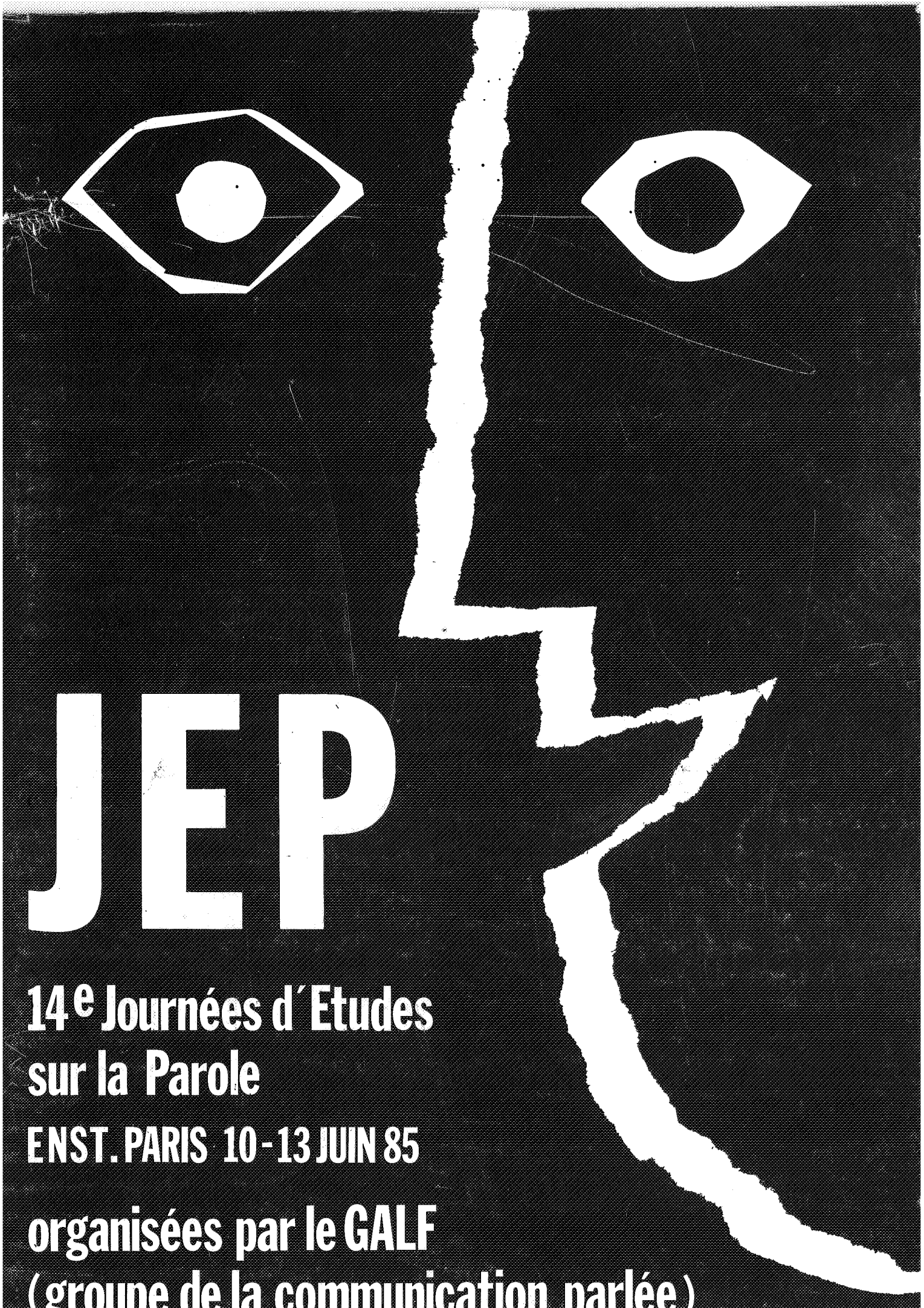
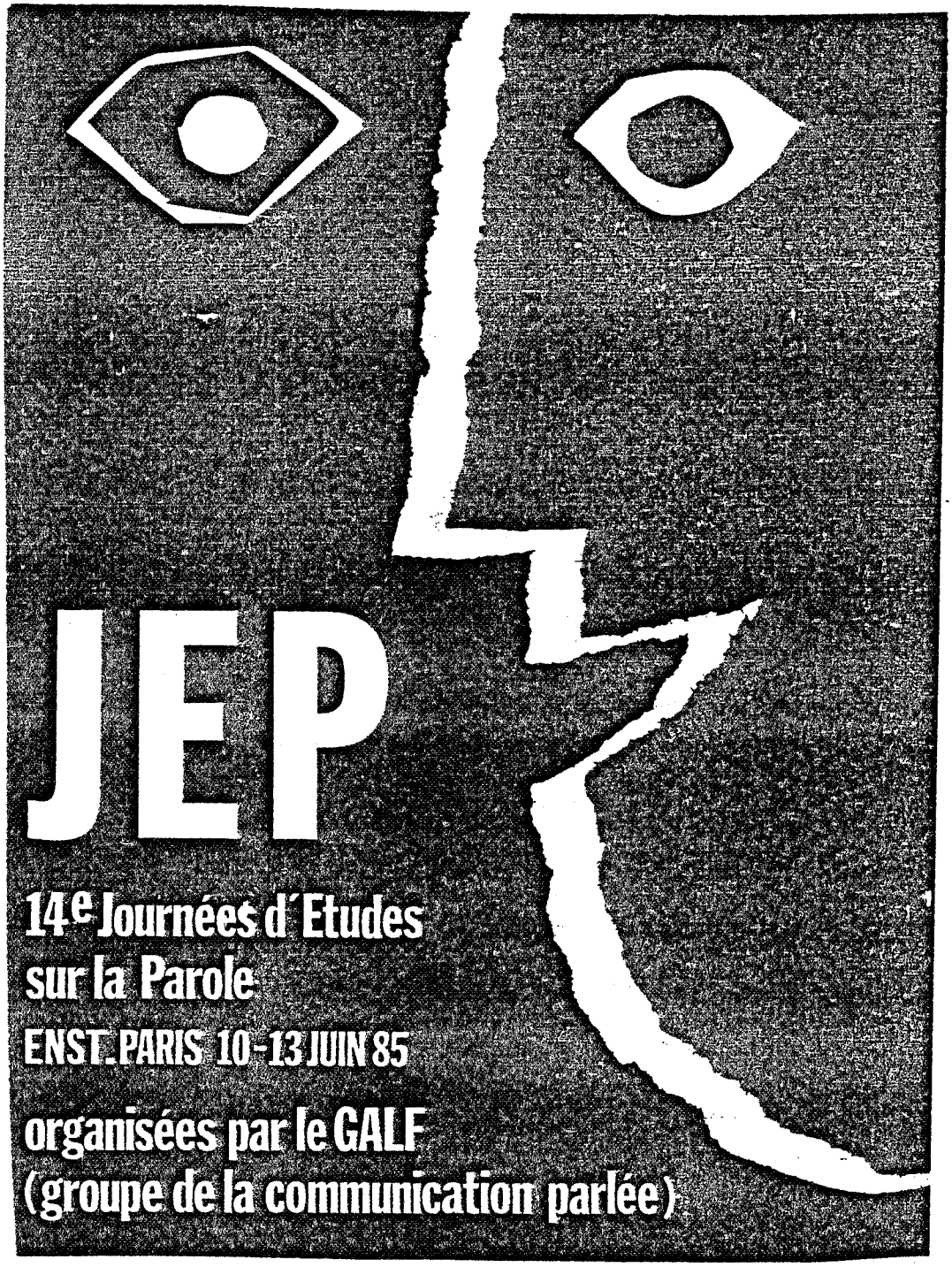


JEP

14e Journées d'Etudes sur la Parole

ENST. PARIS 10-13 JUIN 85

organisées par le GALF
(groupe de la communication parlée)



Ces Journées ont eu lieu à l'Ecole Nationale Supérieure des Télécommunications, 46 Rue Barrault, Paris 13ème.

Elles ont été organisées, sur délégation du Bureau du Groupe de la Communication Parlée du G.A.L.F., par J.P. TUBACH, L. MICLET et G. CHOLLET.

Les Actes de ces Journées ont été édités par J.P. TUBACH
Graphisme: M. GODAY

****
* TABLE DES MATIERES *
****

****
* PERCEPTION *
****

UNE APPROCHE PAYSAGISTE DU DECODAGE D'UNE LANGUE

E. Lhote

Laboratoire de Phonétique
30, rue Mégevand, 25030 Besançon Cedex

RESUME :

To approach a language as a landscape, it is to integrate all the units (sounds, syllabes, words, phrases) as a whole which is more than the putting together of speech units identified in isolation. The decoding process constructs then the meaning not from the temporal sequence of the sonorous events but the orientation system every one acquires in his native language.

Every language to its own listening. To learn a new language means to change his expectation horizon.

Il est maintenant bien connu que les différents éléments sonores d'un message n'exercent pas tous la même fonction sur le comportement perceptif de l'auditeur...

Suivant qu'il connaît ou qu'il ne connaît pas la langue qu'il écoute, l'auditeur accomplit une tâche perceptive différente : dans une langue inconnue, il reconnaît des sons vocaliques, consonantiques, des bruits, des mélodies, des silences; il ne peut pas passer de ce niveau de perception de faits acoustiques au niveau supérieur de perception proprement linguistique qui suppose une mise en relation des faits physiques et des éléments d'un système linguistique déterminé lui permettant d'établir le lien signifiant-signifié, c'est-à-dire d'accéder à la COMPREHENSION.

Après avoir postulé l'existence d'une hiérarchie dans les indices de perception de l'intonation, nous avons observé le comportement d'auditeurs exposés à des mélodies synthétiques et confrontés à une tâche de perception mélodique proprement dite puis à une tâche de perception linguistique (1 - 2)

Nous avons dégagé deux types d'indices participant à la différenciation et à la reconnaissance de quatre patrons intonatifs français, qui ne diffèrent entre eux que par la mélodie (dans la phrase "tu viens demain") ; il n'est naturellement pas surprenant que ces indices s'expriment en termes de hauteur et de durée de notes... puisque ces éléments avaient été choisis au départ. Ce qui est plus intéressant, c'est la *fonction* différenciée de ces deux indices : l'un (l'écart tonal entre la note initiale et la note finale) joue le rôle d'*intégrateur*, dans la mesure où la suite des éléments constitutifs de l'énoncé ne peut être comprise et intégrée à l'expérience antérieure de l'auditeur que dans la partie finale de l'énoncé ; le second (la deuxième note, correspondant au verbe "viens" de "tu viens demain") exerce une fonction de *prédiction*, car l'auditeur peut, dès la deuxième note, prévoir toute - ou une partie de - la signification véhiculée par la suite de l'énoncé.

Ces observations nous ont conduite à mieux comprendre l'originalité du comportement perceptif et nous avons proposé la notion de DECLENCHEMENT qui serait une caractéristique importante de la perception de parole, les éléments déclencheurs n'étant à proprement parler ni des propriétés physiques du signal, ni des propriétés du système récepteur, mais le fruit d'un *comportement linguistique acquis* dans une langue donnée ; le déclenchement serait dans ce cas la preuve que fonctionnent, simultanément et en interaction constante, le niveau acoustique et le niveau de décodage linguistique.

Cette hypothèse sur le comportement linguistique acquiert une nouvelle vitalité si on la compare aux théories de Noizet et de Lederer : Noizet (3) qui s'intéresse à la compréhension de phrases, montre que ce qui caractérise la perception du langage, c'est "la mise en parallèle de plusieurs unités de traitement, correspondant à diverses possibilités de segmentation du stimulus". Lederer (4) qui analyse le fonctionnement de la traduction simultanée, parle d'une *"soudaine synthèse"* entre les informations nouvelles apportées par le discours et les informations antérieurement stockées".

Ce cheminement de la production à la perception de la parole et d'une langue nous a conduite à trouver un modèle qui intègre les passages de niveau à niveau, d'une langue à une autre et qui prenne en compte la trilogie PRODUCTION - PERCEP-

TION - COMPREHENSION.

L'approche paysagiste.

Ce modèle est pour le moment un modèle *paysagiste*, qui, en alliant les modes de deux sens, la vision et l'ouïe, féconde les représentations mentales et enrichit le sonore de la rémanence du visuel. Le concept de paysage sonore intègre l'ensemble des suites sonores que peut capter un individu en *une* seule représentation mentale, c'est-à-dire fixe dans le temps la successivité des stimulations sonores et des images perceptuelles que celles-ci ont fait naître sous forme d'un PAYSAGE qui change avec le moment et avec l'état intérieur de son observateur ; les bruits environnants, la parole, la musique engendrent des paysages variés.

Le compositeur canadien R.M. Shafer (5) ne met dans "son" paysage sonore que l'univers musical et l'environnement sonore urbain ou rural.

Nous proposons quant à nous de considérer que tous les produits oraux d'une Culture (6) participent à l'environnement sonore dans lequel baigne tout individu depuis sa naissance... au même titre que les lieux et le milieu où il vit, qui sont reconnus habituellement comme les constituants premiers de son organisation spatiale et culturelle. De ce point de vue on peut dire que la langue maternelle constitue un élément important du paysage sonore dans lequel se développe le petit d'homme ; chaque langue va alors se caractériser par des paysages sonores qui lui sont propres et chaque auditeur va avoir un comportement d'écoute qui est lié aux paysages qui lui sont familiers, ceux de sa langue maternelle.

Elément du paysage où il évolue et grandit, l'enfant apprend à parler un peu comme il apprend à marcher, c'est-à-dire sans avoir besoin d'une connaissance explicite des éléments de son environnement. Il explore sa langue et sa culture comme il explore sa maison : il développe des *stratégies d'orientation* et des mécanismes de repérage qui lui permettent de devenir un acteur, à l'aise dans son paysage.

Pour l'auditeur, le problème majeur de la compréhension orale est celui du *découpage* qui doit toujours permettre de passer de l'écoute forcément linéaire de la parole à la reconnaissance de formes signifiantes qui, elle, ne procède pas de façon linéaire. Dans la plupart des cas, c'est le RYTHME qui permet (ou qui facilite) le *transfert* entre la suite de sons d'une phrase comme

(avɛklə tɑ̃kilfɛ vole tydwavbjɛ̃puse)

et la suite signifiante

"avec le temps qu'il fait, vos laitues doivent bien pousser".

Les principales caractéristiques d'un modèle paysagiste sont les suivantes :

- il n'est pas de paysage sans observateur
- la parole ne se résume pas plus à une suite de sons qu'une forêt à une somme d'arbres
- l'observateur *fait le paysage* en étant lui-même un élément du paysage.
- entendre (ou voir) un paysage, c'est *intégrer* dans un ensemble unique des éléments temporellement (ou spatialement) éloignés
- les mots-clés de l'approche sont ORIENTATION et POINTS DE REPERE.
  Dans le cas du sonore, s'orienter c'est se trouver des points de repère qui permettent de compenser la fuite du son, et donc de garder la trace un peu plus longtemps à sa disposition.
- le paysage sonore s'évalue par rapport à l'ATTENTE (de l'auditeur).

Dans le cas d'une langue nouvelle

L'approche paysagiste commence toujours par une démarche allant du connu vers l'inconnu : tout observateur regardant un paysage nouveau sait reconnaître une forêt, une ville, une rivière, une falaise, une colline ; ce qu'il ne sait pas, c'est associer ces éléments du paysage à des expériences antérieures. Ce qui le distingue de l'observateur initié qui connaît bien les lieux, c'est son incapacité à créer des réseaux d'images à mémoriser ; en d'autres termes cela veut dire qu'il ne peut pas reconnaître autre chose que des éléments isolés du paysage.

Dans le cas d'une langue nouvelle, l'apprenant, observateur néophyte, identifie des voyelles, des consonnes, des occlusives, des nasales ; il perçoit des syllabes, il entend des mélodies, des groupes rythmiques. Son écoute est celle d'un homme qui entend, qui perçoit mais qui ne peut comprendre... et ne peut donc produire lui-même de nouveaux paysages sonores que l'on puisse intégrer dans l'environnement, parce qu'il n'a pas encore acquis le comportement linguistique qui lui permettrait d'associer des formes sonores et les significations auxquelles celles-ci participent. L'apprentissage de l'*écoute* va donc avoir pour objectif de développer des stratégies d'orientation et des mécanismes de repérage permettant de *déclencher* rapidement des associations entre d'une part des formes sonores nouvelles et la situation de communication dans laquelle ces formes sont employées, et d'autre part cet état de bien-être, recherché par l'auditeur, que seule la compréhension peut engendrer.

On peut alors dire que l'apprenant - observateur construit le système linguistique de la langue qu'il apprend en organisant lui-même son système de *repérage*. Le paysage nouveau s'intègre progressivement aux paysages "antérieurs", et l'apprentissage va développer chez l'apprenant un nouvel *horizon d'attente*.

Cette attente qui caractérise l'auditeur est *double* dans le cas de l'apprentissage d'une nouvelle langue : il faut en effet distinguer celle de l'auditeur natif qui procède par rapport à la langue-cible et celle de l'auditeur "étranger" qui a tendance à perturber le paysage nouveau par le filtrage de celui de sa langue maternelle...

L'approche paysagiste que nous utilisons actuellement dans le laboratoire et qui concerne une douzaine de chercheurs vise à dégager les *points-clés* des paysages propres à une langue, du point de vue du locuteur et de l'auditeur, en associant toujours deux points de vue ; par exemple

- le paysage français "vu par" les sinophones

et le paysage chinois "vu par" les francophones
- les productions françaises des hispanophones et la perception-compréhension de ces productions par les francophones

et dans des applications à la pathologie :
- la parole d'enfants sourds et le degré d'incompréhension des auditeurs...

Comme on le voit, l'approche associe à tout moment la production à la perception et à la compréhension.

En conclusion,

les objectifs de l'approche sont finalisés et explicites : il s'agit de l'apprentissage et de la correction.

Ce que cherche à faire progresser la méthode d'investigation : l'appréciation de la hiérarchie dans les indices, les traits, et plus spécialement les traits rythmiques et prosodiques, dans la circulation du sens entre individus de la même langue et de langues différentes.

Il s'agit donc d'une certaine conception de la reconnaissance de parole ou plus exactement d'une contribution à la reconnaissance de la parole *continue* en situation.

Références.

(1) E. Lhote, "Quelques problèmes posés par l'élaboration de règles prédictives de l'intonation". I.P.S. Miami 1977 ; in *"Current Issues in Linguistic Theory"*, Vol. 9, part I, 1979, 309-319. Ed. H. and P. Hollien, Amsterdam/ John Benjamins B.V.

(2) E. Lhote, *"La parole et la voix :* analyse et synthèse de faits de langue au niveau du larynx"; thèse de doctorat d'état, Strasbourg, mars 1980. Ed. Buske, Hamburg, chapitre 5, pp. 354-523., 1982.

(3) G. Noizet, *"De la perception à la compréhension du langage*, P.U.F., pp. 119-180.

(4) M. Lederer, "La compréhension des textes et des discours vue par la traductologie", dans *"Comprendre le langage"*, Didier, pp. 67, 1980.

(5) R.M. Schafer, *"Le paysage sonore ;* toute l'histoire de notre environnement sonore à travers les âges". Ed. J.C. Lattès, 1979.

(6) E. Lhote, "Les composantes du paysage sonore d'une langue", Colloque C.N.R.S. "Perspectives de l'oralité : KALEVALA 150 ; SUULINEN PERINNE JA NYKYAIKA", mars 1985.

SPECTRES PSYCHOACOUSTIQUES DES CONSONNES VELAIRES EN FRANCAIS

Christel SORIN

Centre National d'Etudes des Télécommunications
22301 LANNION, FRANCE

RESUME

On présente ici les premiers résultats d'une série d'expériences psychoacoustiques visant à étudier la représentation auditive interne des portions du signal de parole essentiels pour les distinctions phonétiques. Ces résultats portent sur la représentation auditive du début de la détente de l'occlusive "K" dans les contextes "AKA" et "UKU". Les "spectres auditifs", obtenus par masquage simultané, ont une forme globale similaire au spectre acoustique du début de la détente mais reproduisent également certaines fluctuations très rapides du spectre DFT. Le degré de détail observé sur ces spectres psychoacoustiques semble très comparable à celui des spectres "physiologiques" basés sur des mesures de synchronisation des réponses sur les fibres auditives unitaires.

1. INTRODUCTION

Le but de cette étude est d'explorer la résolution fréquentielle du système auditif mise en oeuvre lors de l'écoute de sons de parole non-stationnaires.

Les spectres "auditifs" de voyelles stationnaires mesurés par la méthode du seuil de pulsation révèlent un degré relativement grand de résolution spectrale (HOUTGAST 1974). Cette finesse d'analyse ne peut pas être obtenue par un simple filtrage par "tiers d'octave" ou bande critique, filtrage qui est souvent utilisé comme simulation de l'effet des premiers étages du traitement auditif. Il a été suggéré que ce degré élevé de résolution spectrale était dû à l'intervention du processus de suppression à deux tons. Plusieurs études tendent à démontrer que les effets de suppression ne peuvent être observés que si l'on utilise des procédures de masquage non-simultané (HOUTGAST 1974, MOORE 1980 par exemple) conduisant ainsi à l'hypothèse qu'en cas de masquage simultané, tant le son masquant que le son test sont soumis à la suppression, annulant ainsi l'effet de celle-ci. De plus, à notre connaissance, très peu d'études sur les représentations auditives internes ont été faites en utilisant des sons fortement variables dans le temps. La parole étant essentiellement au signal non stationnaire, nous visions à connaître mieux le degré de résolution spectrale pouvant être observé dans de tels cas.

Un autre but de cette étude était de comparer les spectres acoustiques, "auditifs" et "physiologiques" (c'est-à-dire relevés à partir de mesures d'activité des fibres unitaires du nerf auditif) en des points du signal de parole supposés convoyer des informations essentielles pour l'identification phonétique. C'est pourquoi nous avons choisi d'étudier le bruit de détente de consonnes occlusives non voisées, bruit qui semble contenir des informations pertinentes pour l'identification du lieu d'articulation de ces consonnes (BLUMSTEIN and STEVENS 1980).

Nous avons donc mesuré le spectre "auditif" de la partie initiale de la détente de la consonne "K", dans 2 contextes intervocaliques, pour être aussi proche que possible du traitement "naturel" de la parole continue. Ce choix excluait la possibilité d'utiliser une technique de masquage pro-actif (forward masking) et la nature non-stationnaire du stimulus étudié interdisait l'usage d'une procédure de seuil de pulsation. C'est donc la méthode de masquage simultané qui a été utilisée, exclusivement, dans cette expérience.

I - PROCEDURE

1. Les sons masquants :

Deux segments de parole, AKA et UKU, ont été utilisés comme sons masquants. Ils étaient extraits de parole naturelle (en l'occurence, de courtes phrases de type "c'est KUKU çà" enregistrées dans un studio acoustiquement bien isolé) et stockés numériquement sur un PDP 11/24 avec une fréquence d'échantillonnage de 16 KHz. Le début et la fin de ces deux signaux étaient pondérés par une demi-courbe de Gauss avec un temps de montée/descente global de 50 ou 30 ms respectivement pour AKA et UKU (fig. 1 et 2). Ces sons masquants étaient présentés monauralement à un niveau maximum de 77 dB SPL (RMS, constante de temps rapide).

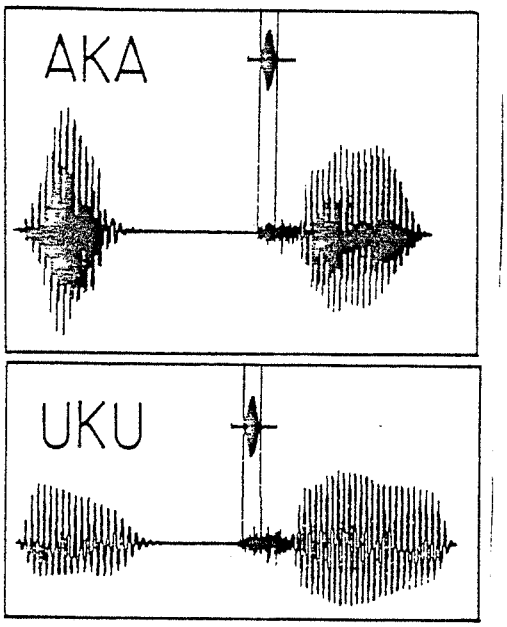


Fig. 1 et 2 : Forme d'onde des deux sons masquants et position du son test.

2. Les sons test :

Le son test était un son pur de durée 16 ms et ayant pour enveloppe une fenêtre de Hamming. Sa largeur de bande était inférieure à 150 Hz à - 10 dB. Ce son test était situé juste au début de la détente du "K" comme cela est indiqué sur les figures 1 et 2. Les fréquences du son test variaient, pour le masqueur AKA, de 800 à 2600 Hz par pas de 50 Hz et pour le masqueur UKU, de 800 à 2400 Hz également par pas de 50 Hz.

3. Méthode :

Les seuils de masquage ont été mesurés en utilisant la méthode d'ajustement adaptatif (LEVITT 1970). Les signaux masquants seuls (c'est-à-dire AKA et UKU) et masquants + son test étaient présentés par paire avec un silence interstimulus de 500 ms. Un silence de 3.5 sec. suivait la réponse du sujet. Le niveau initial du son test était fixé, pour chaque fréquence, de manière à ce que le son test soit clairement audible. Les pas d'accroissement ou de diminution du niveau du son test, suivant la réponse du sujet étaient de 8, 4, 3, 3, 2, 2, 1, 1 dB avant les 1er, 2ème, 3ème... 8ème renversements. Après le 8ème renversement, une première évaluation du seuil était faite sur les 8 renversements et également sur les 4 derniers renversements. Si la différence entre ces deux estimations était inférieure à 1.5 dB, on prenait comme valeur du seuil la moyenne des 4 derniers renversements. Si non, le test continuait pour 4 ou 8 renversements supplémentaires et le seuil relevé était la moyenne des niveaux sur les 8 derniers renversements.

Deux sujets ont été testés, ayant toutes deux une grande pratique dans des tâches similaires. Ces deux sujets avaient des seuils absolus normaux sur toutes les fréquences audiométriques. Les signaux étaient présentés monauralement via un écouteur TDH 39. Les sons masquants et test étaient additionnés de manière numérique et les modifications du niveau du son test en fonction de la réponse du sujet étaient effectuées par programme pendant l'intervalle de silence suivant la réponse.

II - RESULTATS

Pour chaque sujet, les seuils donnés ici correspondent à la moyenne d'au moins 4 ajustements pour une même condition. La moyenne des écart-types était de 0.9 dB pour le sujet CS et de 1.4 dB pour le sujet MB.

Les contours de masquage obtenus pour les masqueurs AKA et UKU sont présentés sur les figures 3 et 4. Chaque courbe correspond aux résultats d'un seul sujet.

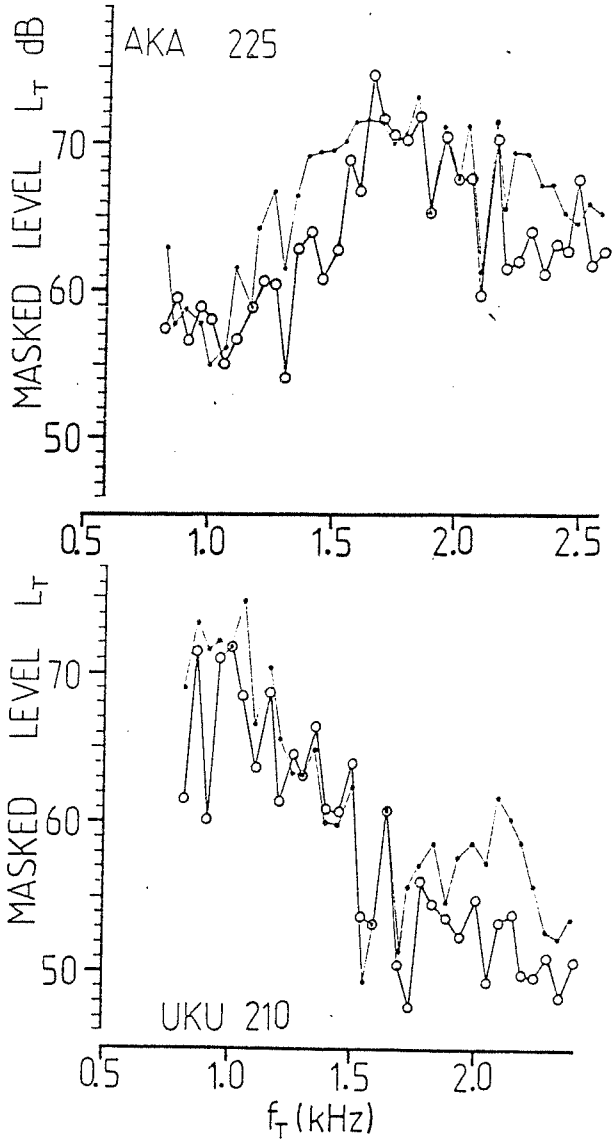


Fig. 3 et 4 : Courbes de masquage obtenues pour les 2 sujets et les 2 sons masquants AKA et UKU (o——o sujet CS, •——• sujet MB)

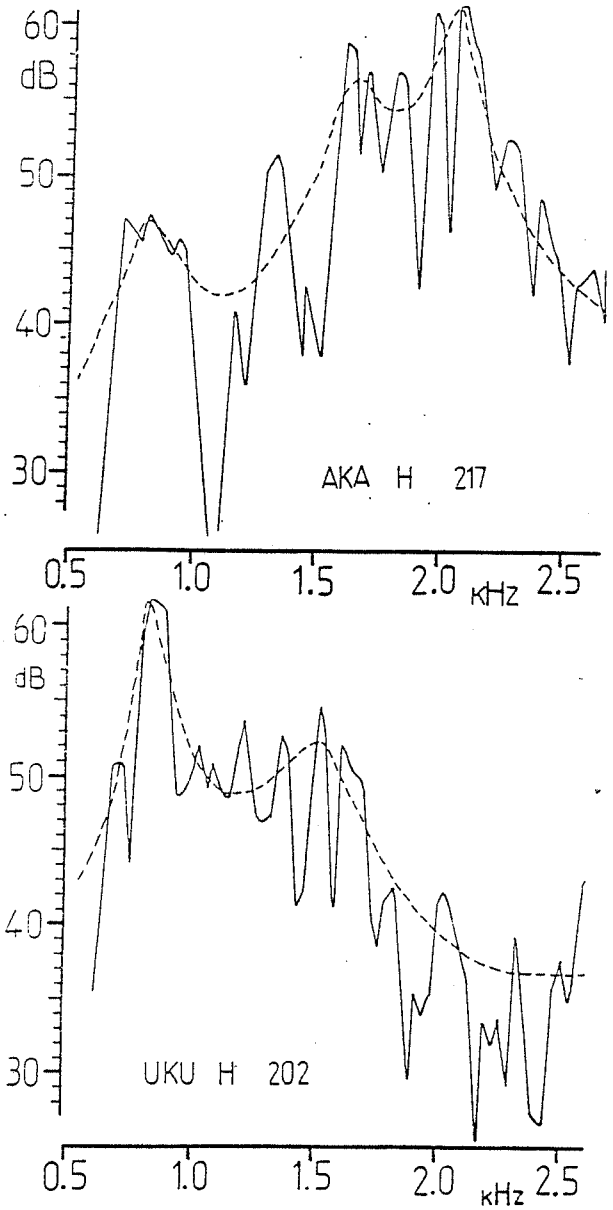


Fig. 5 et 6 : Spectres DFT et LPC (- - -) des 16 premières ms de la détente du "K" dans les 2 contextes AKA et UKU (fenêtre de Hamming)

Les contours de masquage peuvent être comparés aux spectres DFT et LPC de la portion du bruit de détente du "K" sur laquelle le son test était superposé. Ces spectres sont présentés sur les figures 5 et 6 (fenêtre d'analyse de Hamming, de 16 ms, à l'emplacement exact du son test).

On remarque que les formes globales des spectres acoustiques et "auditifs" sont similaires : pour le masquer AKA, une large zone de maxima apparaît dans le "spectre auditif" entre 1.5 et 2.2 KHz, ainsi que dans le spectre acoustique. Pour le masqueur UKU, le maximum du "spectre auditif" est situé aux alentours de 1.0 KHz et est suivi par une pente décroissante d'environ 20 dB/1 KHz, ce qui correspond bien à la forme globale du spectre DFT du début du bruit de détente dans le contexte UKU.

Mais les contours de masquage montrent aussi des fluctuations spectrales rapides : des pics et vallées sont clairement visibles, séparés de moins de 150 Hz, en particulier autour de 2 KHz pour le masqueur AKA et de 1.5 KHz pour le masqueur UKU. Ces fluctuations spectrales rapides dans les "spectres auditifs" semblent être bien corrélés avec les détails fins des spectres DFT correspondants.

L'aspect le plus intéressant de ces résultats est que deux pics du spectre DFT séparés en fréquence de 150 Hz seulement aux alentours de 2 KHz, avec une profondeur de creux de quelques 15 dB, semblent être clairement séparés dans le "spectre auditif". Un degré aussi élevé de résolution fréquentielle semble n'avoir jamais été observé dans cette zone de fréquence (autour de 2 KHz), en particulier en utilisant la méthode de masquage simultané. Les contours de masquage relevés ici reflètent les spectres DFT avec une précision bien plus importante que celle fournie par une analyse utilisant des filtres déduits des formes d'excitation auditive proposés, par ex., par MOORE (1983). Avec de tels filtres, les 2 pics situés, pour AKA, à 1.9 et 2.15 KHz ne seraient pas détectés.

Le degré de détail observé sur ces contours de masquage est également supérieur à celui relevé sur des contours de masquage pour des sons complexes harmoniques, même en utilisant les techniques de masquage non simultané. A titre d'exemple la figure 7 donne le contour de masquage obtenu par HOUTGAST (1974) pour un son complexe contenant les 10 premières harmoniques d'un fondamental à 250 Hz, en utilisant la méthode du seuil de pulsation. On remarque que les creux entre les harmoniques ne sont quasiment plus détectés au-delà de 1.5 KHz bien que la profondeur des "creux" acoustiques correspondants soit bien supérieure à celle observée sur les spectres du "K" et que les harmoniques soient plus écartées que les pics des spectres du "K" de notre expérience.

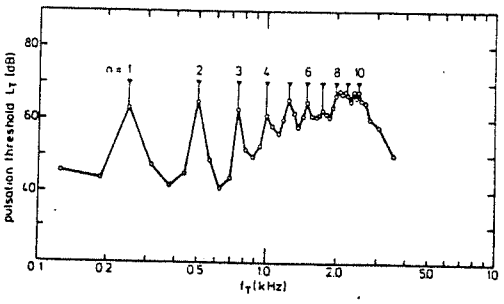


Fig. 7 : Seuils de pulsation pour un son masquant formé des 10 premiers harmoniques d'un fondamental de 250 Hz (d'après HOUTGAST 1974)

III - CONCLUSION

L'utilisation de la procédure de masquage simultanée pour l'étude de la représentation interne des sons a été souvent considérée comme conduisant à une sous-estimation de la résolution fréquentielle de l'ouie : seules les techniques de masquage simultané (proactif ou seuil de pulsation) sont censées rendre compte de l'augmentation de la résolution fréquentielle due au rôle de la suppression à deux tons. Pourtant en utilisant la procédure de masquage simultané et des sons masquants non-stationnaires, nous avons observé ici un degré de résolution fréquentielle apparemment bien supérieur à celui observé pour des sons masquants stationnaires, testés avec des procédures de masquage non simultané. Cela signifie-t-il que la procédure de masquage simultané permet, dans certaines conditions, de prendre en compte l'effet de la suppression et/ou que le degré de résolution fréquentielle observable est augmenté quand on utilise des masqueurs non stationnaires ? De fait FASTL et BECHLY (1983) ont suggéré que l'effet de suppression peut être mesuré même avec le masquage simultané : l'ampleur de la suppression observable augmente quand la durée du son test diminue et un son masquant pulsé "produit" plus de suppression qu'un son masquant continu. Toutefois les auteurs concluent que les effets de suppression obtenus en masquage simultané sont considérablement inférieurs à ceux obtenus en masquage non simultané.

Un mécanisme autre que la suppression est-il responsable du degré élevé de résolution fréquentielle observé ici ? Ou bien doit-on penser que l'effet de la suppression est augmenté par la caractéristique de forte nonstationnarité du signal masquant ? Malheureusement très peu d'expériences psychoacoustiques semblent avoir été menées en utilisant des sons masquants variant fortement dans le temps. La reproduction de l'expérience décrite ici, mais en utilisant comme son masquant un bruit stationnaire ayant le même spectre acoustique fin que celui des 16 ms initiale du bruit de détente du K serait d'un grand intérêt pour répondre à ces deux questions.

La comparaison des contours de masquage présentés ici avec certaines données physiologiques est également très utile. Les ressemblances les plus frappantes apparaissent quand on compare nos données avec les spectres ALSR, définis par YOUNG et SACHS (1979) et basés sur des mesures de synchronisation des réponses des fibres auditives unitaires. La figure 8 présente les spectres DFT et ALSR pour un segment de 20 ms d'un stimulus /da/ observés par SACHS, YOUNG et MILLER (1982). Les fluctuations spectrales rapides du stimulus sont reproduites dans le spectre ALSR avec une précision similaire à celle observée dans nos contours de masquage psychoacoustiques.

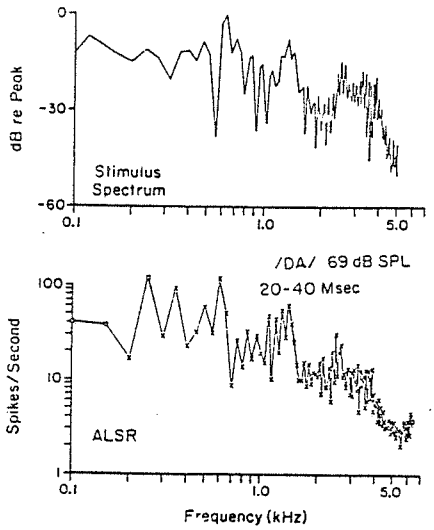


Fig. 8 : Spectres DFT et "Physiologiques" (ALSR) des 20 à 40 ms initiales d'un /da/ (d'après SACHS et al 1982)

Nous concluerons provisoirement de ces observations que les études sur les représentations auditives internes des sons de parole nécessitent encore de nombreux travaux avant de pouvoir proposer aux "traiteurs de parole" (spécialistes de synthèse et de reconnaissance...) l'outil d'analyse du signal qui simulerait avec précision le traitement auditif périphérique auquel le signal de parole est soumis lors de sa compréhension par l'être humain.

REFERENCES

S.E. BLUMSTEIN and K.M. STEVENS (1980) : "Perceptual invariance and onset spectra for stop consonants in various vowel environments", J. Acoust. Soc. Am. 67, 648-662.

H. FAST and M. BECHLY (1983) : "Suppression in simultaneous masking" J. Acoust. Soc. Am. 74(3) 754-757.

T. HOUTGAST (1974) : "Lateral suppression in Hearing" Doctoral thesis, Free University of Amsterdam.

H. LEVITT (1970) : "Transformed Up-Down Methods in Psychoacoustics" J. Acoust. Soc. Am., 49, 467-477.

B.C.J. MOORE (1980) : "Mechanism and frequency distribution of two-tone suppression in forward masking" J. Acoust. Soc. Am. 68, 814-824.

B.C.J. MOORE (1983) : "Suggested formulae for calculating auditory-filter bandwiths and excitation patterns" J. Acoust. Soc. Am. 74(3), 750-753.

E.D. YOUNG and M.B. SACHS (1979) : "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers" J. Acoust. Soc. Am. 66, 1381-1403.

M.B. SACHS, E.D. YOUNG, M.L. MILLER (1982) : "Encoding of speech features in the auditory nerve" in "The Representation of Speech in the Peripheral Auditory System" ed. by R. Carlson and B. Granström, 115-130.

APPLICATION DU TEST DE RIMES A DES PATIENTS SOURDS TOTAUX

J. Génin

Centre National d'Etudes
des Télécommunications
38243 Meylan

ABSTRACT.

Several laboratories in France and abroad propose sensory aids for totally deaf people. The operation of such devices makes use of auditive perceptions evoked by elecrical stimulation of a set of electrodes implanted by surgery close to the patient's cochlea or auditive nerve. These devices are very different from one another and very little is known among how useful they could be for the patients.

This work is related with a restricted version of the diagnostic rhyme test use for performance comparison.

I. INTRODUCTION.

Différents laboratoires étudient et mettent en oeuvre des systèmes de prothèse cochléaire. Ces matériels, qui s'adressent toujours à des patients sourds totaux, utilisent les sensations auditives provoquées par l'excitation électrique d'électrodes implantées par voie chirurgicale au voisinage de la cochlée ou du nerf auditif. Ils diffèrent cependant notablement dans leurs modalités de mise en oeuvre : nombre et nature des électrodes implantées, site d'implantation, nature des signaux électriques appliqués, nature des méthodes de traitement de signal mises en oeuvre... A ce jour, il existe encore peu d'évaluations comparatives des performances que les patients peuvent espèrer de ces thérapeutiques.

C'est pourquoi, en France, les pouvoirs publics ont mis en place un protocole d'évaluation de différents systèmes étudiés par les équipes françaises. Nous avons participé à l'adaptation et à la mise en oeuvre d'un test de rimes (Peckels, 1973) dans le cadre de cette opération.

II. VOCABULAIRE DU TEST.

Le vocabulaire utilisé pour le test a été établi par B. Delgutte, Dépt. RCP, CNET Lannion. Le souci d'aboutir aux séquences de test les plus courtes a fait retenir un système réduit composé de 24 oppositions de paires et 8 triplets pour l'étude des traits phonétiques des consonnes, 26 oppositions de paires pour les voyelles.

Les oppositions phonétiques étudiées pour les consonnes sont :

sourd/voisé	"tard/dard"	8 paires
interrompu/continu	"tâche/sache"	8 paires
compact/diffus	"rate/latte"	2 paires
grave/aigu	"miche/niche"	2 paires
vocalique/non vocal.	"joue/roue"	2 paires
nasal/non nasal	"nous/loup"	2 paires
lieu d'art.	"fa/sa/chat"	8 triplets

Pour les voyelles :

nasal/non nasal	"pente/patte"	6 paires
grave/aigu	"bouse/buse"	8 paires
ouvert/ferme	"beurre/bure"	12 paires

On a essayé d'étendre au maximum la couverture du système phonétique avec ce vocabulaire, mais le souci de limiter la durée des séquences de tests interdit une couverture exhaustive.

III. MISE EN OEUVRE PRATIQUE.

La mise en oeuvre du test se fait de la façon habituelle au moyen de feuilles lues par le locuteur et de feuilles de réponse fournies au patient (figure 1 et 2). Un programme informatique est utilisé pour établir ces feuilles dans lesquelles l'ordre de présentation des paires ou triplets, l'ordre dans lequel les mots sont présentés et la position de la bonne réponse sont tirés au hasard. La seule contrainte est qu'au cours d'une séquence de test, chaque paire soit testée quatre fois (chaque mot prononcé deux fois) et chaque triplet trois fois (chaque mot prononcé une fois). Le programme informatique peut, bien sûr, proposer des tirages différents afin qu'un même patient puisse subir plusieurs fois le même test sans avoir la possibilité d'apprendre les réponses par coeur.

Après saisie des réponses que le patient a cochées sur les feuilles qui lui ont été confiées, un second programme en effectue l'analyse et établit automatiquement les résultats statistiques que l'on peut en tirer.

Ce sont tout d'abord les résultats primaires du test : nombre de bonnes réponses et de confusions pour chaque paire ou triplet proposé. Ces résultats sont également produits par trait et de façon globale pour l'ensemble de la séquence de test.

FEUILLE POUR LE LOCUTEUR

	CURE
VOUTE	
	SEULE
SELLE	
	FAUTE
SAUTE	
BORD	
	GALE
	POTE
BELLE	
	BURE
COLLE	
	PENTE

jeu # 1, alea : 0, page # 2

Figure 1. Exemple de feuille lue par le locuteur.

Les systèmes de prothèse cochléaire qui nous intéressent ici procurent au patient une audition de médiocre qualité et certainement très différente de l'audition naturelle dont il lui reste peut-être le souvenir. Une cause évidente de cela est que les sections de la cochlée les plus accessibles à l'opération chirurgicale sont le premier et le deuxième tour, source de la perception des timbres les plus aigus, souvent au delà des fréquences habituelles de la parole ... et que le patient a perdue le plus tôt. C'est d'ailleurs pour cela que, au delà de l'opération chirurgicale et de la remise de l'appareil de prothèse complet une longue rééducation est nécessaire.

Il arrive ainsi qu'un patient donne pour certaines oppositions ou certains traits phonétiques des résultats plus souvent faux que s'il répondait au hasard. Ceci signifie qu'il distingue mais ne reconnaît pas, défaut dont on peut espérer voir la disparition au cours de la rééducation. Nous cherchons alors un résultat plus pertinent avec le taux de transmission d'information (Miller, 1955).

FEUILLE POUR L' AUDITEUR

COTE	CONTE ✓
BEURRE ✓	BURE
CURE	COEUR ✓
BELLE	BALLE ✓
NATTE	NANTES ✓
DUNE	DINE ✓
BEURRE ✓	BORD
JOUTE ✓	JUTE
GILLES	GEL ✓
SEULE ✓	SOLE
JUTE	GITE ✓
SEL ✓	CIL
SELLE	SALLE ✓

jeu # 4, alea : 9787, page # 7

Figure 2. Exemple de feuille de réponse.

Si le patient répond le mot M_j lorsqu'on a prononcé le mot M_i (i,j = 1 ou 2, 1,2 ou 3 pour les triplets, j = 0 si le patient ne répond pas) et si on relève les fréquences P_{ij} (M_i prononcé, M_j répondu), P_i (fréquence de la prononciation du mot M_i, en fait toujours 1/2 ou 1/3) et P_j (fréquence des réponses j), le taux de transmission d'information s'écrit :

$$T = \frac{\sum_{ij} P_{ij}.\mathrm{Log}(P_{ij}/(P_i.P_j))}{\sum_{i} P_i.\mathrm{Log}(P_i)}$$

On vérifie immédiatement que si la réponse est purement aléatoire, $P_{ij} = P_i.P_j$ entraîne T = 0 alors que si la séparation est parfaite $P_{ij} = P'_j$ entraîne T = 1.

IV. PREMIERS RESULTATS.

A ce jour, l'opération d'évaluation comparative lancée par les pouvoirs publics n'a pas encore fourni suffisamment de résultats. Nous ne disposons que de résultats partiels obtenus auprès de quelques patients ayant subi l'opération

	Traits des consonnes							Traits des voyelles		
	voisé	interr.	compact	grave	vocalique	nasal	lieu	nasal	grave	fermé
patient										
Ro...	2%	30%	31%	14%	55%	66%	16%	29%	11%	32%
An...	0	1%	0	31%	5%	0	2%	1%	0	1%
Fo...	8%	3%	31%	14%	5%	5%	4%	2%	0	1%
De...	8%	14%	0	19%	100%	14%	4%	11%	8%	22%
Ba...	21%	11%	14%	19%	31%	55%	10%	46%	5%	1%
Ra...	7%	2%	20%	10%	10%	9%	1%	29%	3%	27%

Tableau 1. Résultats (taux de transmission d'information).

chirurgicale à Grenoble (CHU de Grenoble, S^{ce} du Pr. Charachon) ou à Paris (CHU St Antoine, S^{ce} du Pr. Chouard). Tous ces patients ont reçu le même type de prothèse cochléaire multicanal à 12 électrodes CHORIMAC (Fardeau, 1979).

Le tableau 1 présente les résultats que nous avons obtenus pour ce taux de transmission d'information avec chaque patient pour chaque trait phonétique étudié.

Les 4 premières lignes de ce tableau sont relatives à des patients qui ont subi l'opération chirurgicale récemment (octobre 1984) ; les 2 derniers patients ont été opérés il y a environ 3 ans. Les résultats relativement mauvais de An... et Ra... s'expliquent par un nombre d'électrodes fonctionnelles faible (mauvais état résiduel du nerf auditif ou rupture des connexions ?...).

La faible dimension statistique des échantillons interdit de rechercher des conclusions précises pour le cas de chaque patient. Cependant l'examen général des résultats fait ressortir les points suivants :

- les traits phonétiques des consonnes "passent mieux" que ceux des voyelles. Ceci est surtout vrai pour les traits compact, grave, vocalique et nasal qui s'appuient sue la présence d'indices spectraux alors que les autres traits s'appuient sur leur variation dans le temps.

- les traits nasal et ouvert des voyelles "passent" mieux. On peut y voir une meilleure efficacité des indices énergétiques par rapport aux indices spectraux, plus importants pour le trait grave.

Ces premiers résultats sont, bien sûr, insuffisants. Il sera intéressant, au moyen de ce test de rimes, de comparer différents dispositifs de prothèse cochléaire, tant à électrodes multiples qu'à électrodes uniques.

REFERENCES.

M. Fardeau & P. Orange, "Les implants cochléaires : l'appareillage", Les Cah. d'O.R.L. 14-6, pp. 606-616, 1979.

G. A. Miller & P. E. Nicely, "An Analysis of Perceptual Confusions among some English Consonants", JASA 27, 1955.

J. P. Peckels & M. Rossi, "Le test de diagnostic par paires minimales", Revue d'Acoust. 27 pp. 245-262, 1973.

RECONNAISSANCE DE MOTS ARTIFICIELS A TRAVERS LA PROTHESE COCHLEAIRE CHORIMAC

R.Mouhssine*, C.Berger-Vachon*, J.Genin**

* Laboratoire GBM, Université de Lyon I
43 bld du 11 Novembre 1918, 69622 Villeurbanne
** C.N.E.T. ZIRST Meylan 38240 Meylan

Abstract

In this paper, the authors study the performances obtained in the recognition of a 129 word list constructed with artificial words embedding phonetical oppositions :

* by two totally deaf patients implanted with the french cochlear prosthesis Chorimac
* by an automatic procedure dealing with the spectrograms and using an euclidian metric and the nearest neighbour for archetype.

The results show that :

* the number of functional channels does not seam to be dramaticaly important
* frequency shifts in the implanted electrodes should improve the performances at the acoustic stage
* results obtained with the automatic procedure are higher than those given by the patients.

I-INTRODUCTION

Jusqu'à ces dernières années les surdités liées à une défaillance de l'oreille interne restaient inaccessibles à la thérapeutique. Mais depuis une décennie un espoir est apparu pour les malades touchés par une dégénérescence de l'organe de Corti et qui conservent un nerf cochléaire encore fonctionnel, puisque les premières tentatives pour exciter directement les dendrites distales du nerf auditif ont donné des résultats intéressants [1].

La prothèse permettant de réaliser cette excitation est construite en France par les laboratoires Bertin à Aix-en-Provence (prothèse Chorimac) [2]. Elle a déjà été implantée à une centaine d'exemplaires par les équipes des professeurs Chouard (Paris), Charachon (Grenoble) et Morgon (Lyon).

Différentes expériences psycho-acoustiques ont permis de chiffrer les performances des implantés mais on ne s'est pratiquement pas intéressé aux spectrogrammes du signal qui étaient distribués à l'oreille interne, ni aux modifications qu'on pouvait lui faire subir pour améliorer la qualité de l'information qu'il transporte.

Dans ce travail, nous nous proposons de comparer les performances de reconnaissance obtenues par un traitement automatique du signal à celles qui sont données par les implantés quand on leur propose une liste de mots artificiels. Ces mots artificiels sont construits à partir d'oppositions phonétiques simples [5] qui sont les suivantes :

- 3 voyelles (/a/, /i/, /o/)
- 2 consonnes voisées (/d/, /z/) et 2 non-voisées (/t/, /s/)
- 2 fricatives (/s/, /z/) et 2 plosives (/d/, /t/).

Leur association a conduit à un vocabulaire de 12 mots formés de syllabes doublées pour une meilleure compréhension (Tableau 1) :

1.DADA (PVA) ; 2.DODO (PVO) ; 3.DIDI (PVI) ;
4.TATA (PVA) ; 5.TOTO (PVO) ; 6.TITI (PVI) ;
7.SASSA(FVA) ; 8.SOSSO(FVO) ; 9.SISSI(FVI) ;
10.ZAZA (FVA) ;11.ZOZO (FVO) ;12.ZIZI (FVI) ;

Tableau 1 - Mots du vocabulaire

Ces mots ont été répétés de façon aléatoire pour constituer une liste de 129 mots contenant entre 10 et 12 occurences de chacun des mots du vocabulaire.

La liste a été lue plusieurs fois à 2 patients de 54 et 55 ans implantés par l'école de médecine lyonnaise en janvier 1983. Le 1er patient (D G) avait 7 électrodes de fonctionnelles et le second (P N) en avait 12.

Il est intéressant de comparer les performances obtenues par les 2 patients aux résultats de la reconnaissance automatique utilisant un nombre équivalent de canaux (7 ou 12).

II-METHODES

2.1 Traitement des tableaux de reconnaissance

En comparant les réponses des patients aux mots prononcés, nous avons établi un tableau 12 x 12 indiquant les bonnes reconnaissances et les confusions observées lorsque la liste a été proposée aux patients et lorsqu'elle a été traitée par des procédures de reconnaissance automatique.

La forme du tableau 12 x 12 est indiquée sur la figure 1 (exemple du premier patient, premier test et absence de lecture labiale).

S \ R	1	2	3	4	5	6	7	8	9	10	11	12
1			4	1		1			1	1		4
2									3		2	5
3			3		1	1			1			4
4			3				1		4	2		
5			2		1	4		2	3			
6			4	1	1	1	1		2		1	
7	2		1	1		1	2		1	1		1
8	3		1	3		1	1		1	1		
9	3		1			1	2		1		2	
10	1			1		4			3	3		
11			3					1	2			4
12		1		1		1		1	2	4		1

Figure 1 - Tableau de reconnaissance du vocabulaire (S = stimulus, R = réponse).

* * *

La trace de cette matrice indique le nombre de mots qui ont été correctement reconnus. En regroupant les lignes et les colonnes de ce tableau [3] on obtient des sous-tableaux indiquant les reconnaissances et les confusions des :

* voyelles
* consonnes
* traits voisé et non-voisé
* traits plosif et fricatif.

A chacun de ces tableaux, on associe un nombre de bonnes reconnaissances. Le test sera ensuite globalement caractérisé par la moyenne de toutes ces reconnaissances.

2.2 Reconnaissance automatique des mots

a) Représentation d'un spectrogramme

Les spectrogrammes fournis par la prothèse Chorimac indiquent l'évolution de la sensation auditive (logarithme de l'énergie) dans 12 bandes fréquentielles avec un échantillonnage toutes les 3 millisecondes [4].

Pour chaque échantillon temporel, nous avons construit la somme de sensations élémentaires associées à chaque filtre. Compte tenu de l'aspect très bruité de cette somme, les spectrogrammes ont été lissés et une fenêtre de filtrage de 24 échantillons a été retenue.

La structure des mots qui sont constitués par une alternance de 2 voyelles et de 2 consonnes dans un environnement bruité fait apparaître assez nettement 2 pics énergétiques. Une procédure automatique de "segmentation", basée sur des critères différentiels a permis de localiser avec certitude le maximum-maximarum du premier pic et le minimum-minimarum situé entre les 2 pics. Le maximum correspond à la première voyelle et le minimum à la deuxième consonne, ce qui est suffisant pour reconstituer le mot. L'image simplifiée du spectrogramme est alors constituée par 2 vecteurs dans un espace à 12 dimensions, soit un vecteur dans un espace à 24 dimensions.

b) Méthode de reconnaissance

La méthode de reconnaissance que nous avons choisie est basée sur les considérations suivantes :

* un spectrogramme X à classifier est comparé, à l'aide du calcul d'une distance, à chacune des 12 classes de mots possibles.
* la distance euclidienne est utilisée.
* l'archétype caractérisant une classe est le représentant de cette classe le plus proche de X (méthode du plus proche voisin). Pour la définition des archétypes, X a été, bien entendu, exclu du corpus des mots prononcés.

III-RESULTATS

3.1 Performances des implantés

Six tests ont été effectués avec le premier patient et quatre avec le second qui a dû être désimplanté neuf mois après l'intervention pour des raisons d'infection sur la prothèse.

Les tests ont eu lieu deux semaines, six semaines, quatre mois, huit mois, seize mois et vingt mois respectivement après l'implantation.

Ils ont conduit aux pourcentages de bonnes reconnaissances indiqués sur le tableau 2.

patient / Taux de reconnais.	D G		P N	
	MINI	MAXI	MINI	MAXI
mots	9,3	26,3	4,7	30,2
voy.	34,9	58,1	24,8	64,3
cons.	27,1	51,1	29,5	42,6
v/nv	58,9	64,3	50,4	58,9
p/f	48,0	74,7	53,5	65,9
Moy.Gle	35,7	55,3	32,6	52,4

Tableau 2 - Pourcentages de bonnes reconnaissances obtenus par les implantés.

* * *

L'évolution des performances des implantés au cours du temps, pour la reconnaissance des mots, est représentée sur la figure 2 (résultats donnés en pourcentages).

Ce tableau indique que le premier et le sixième test pour le premier implanté, et les troisième et quatrième tests pour le second implanté donnent les valeurs extrêmales.

Le quatrième test a été effectué après que les patients aient été privés de leurs machines retournées aux établissements Bertin pour révision. La durée de l'immobilisation de la prothèse a été de un mois environ et dans les deux cas, une chute des performances très nette a été observée.

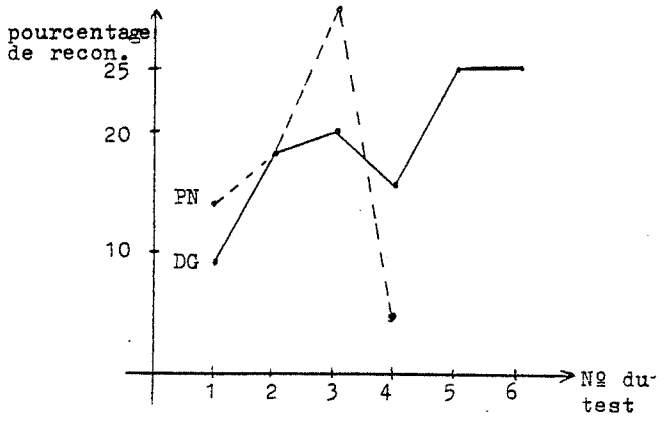


Figure 1 - Evolution des pourcentages de bonne reconnaissance pour les tests effectués avec les implantés.

* * *

3.2 Reconnaissance automatique

La reconnaissance automatique a été effectuée dans les conditions suivantes :

* utilisation des 12 canaux
* utilisation des 7 canaux dont dispose le premier patient, DG, (canaux 1/7/8/9/10/11/12) et détermination du meilleur heptet (canaux 3/4/5/7/8/10/12) et du plus mauvais heptet (canaux 1/2/6/9/10/11/12) parmi les 792 combinaisons.

Les canaux sont numérotés en partant des basses fréquences. Le tableau 3 résume les résultats (donnés en pourcents).

	Meilleur heptet	Heptet DG	Plus mauvais heptet	Douze canaux
Mot	51,2	36,4	27,9	48,1
Voy.	94,6	68,2	69,0	95,3
Cons.	51,2	48,1	38,0	52,0
V/NV	71,3	63,6	60,0	67,4
P/F	67,4	67,4	58,9	72,1
Moy.Gle	67,1	56,7	50,7	67,0

Tableau 3 - Résultats de la reconnaissance automatique.

* * *

L'heptet correspondant au patient DG est classé en 638e position. On remarque que, mis à part le canal nº 1, touté l'information qu'il transmet est située vers les hautes fréquences. Ce mauvais balayage du spectre sonore est, a priori, un handicap pour la reconnaissance.

IV-DISCUSSION

Ce travail montre que les performances obtenues par la machine sont supérieures à celles obtenues par les implantés. En ce qui concerne le premier patient, l'utilisation de l'heptet correspondant à la configuration fréquentielle de l'oreille du patient conduit à une moyenne de 56,7 % pour la reconnaissance automatique, tandis que son meilleur score est de 55,3 %. Avec 7 canaux, on a pu obtenir une moyenne de 67,1 % ce qui permet de penser qu'une transposition fréquentielle pourrait améliorer les performances de reconnaissance de notre vocabulaire par ce patient.

La meilleure moyenne du second patient est 52,4 % et la procédure automatique de reconnaissance conduit à 67,0 %.

Il est aussi remarquable de constater que les performances obtenues avec 7 et 12 canaux sont très comparables tant du point de vue des implantés que de celui des procédures automatiques, dans notre étude.

On a constaté aussi que le second patient a montré des performances initiales supérieures au premier patient mais les résultats se sont ensuite égalisés. La richesse fréquentielle du spectre n'est donc pas le seul élément à considérer. Le profil psychosomatique des sujets ainsi que la qualité et la constance de la rééducation ont probablement aussi un rôle important à jouer.

V-CONCLUSIONS

L'étude que nous avons réalisée a mis en évidence les éléments suivants :

* La richesse fréquentielle du spectre n'est pas un élément particulièrement déterminant pour la reconnaissance du vocabulaire que nous avons étudié; la détermination du nombre optimal de canaux est un problème ouvert.
* Les procédures automatiques de reconnaissance ont conduit à des résultats nettement supérieurs à ceux qui ont été obtenus avec les implantés ayant participé à notre étude.

* On peut envisager avec profit des transpositions fréquentielles du spectre chez les patients n'ayant pas un nerf auditif fonctionnel dans toute la gamme audible.

Tous les résultats méritent d'être précisés dans le futur, surtout maintenant que le développement de la micro-informatique permet d'envisager des traitements élaborés du spectre sonore.

BIBLIOGRAPHIE

[1] C.H. Chouard et Coll., "Résultats cliniques de l'implant cochléaire multiélectrodes", Ann.Otolaryngol., Paris, Vol. 98, pp 593-612, 1981.

[2] M. Fardeau, P. Orange, "L'appareillage", Cahiers d'ORL, Vol. 14-6, pp 609-616, 1979.

[3] A. Morgon, C. Berger-Vachon et Coll., "Cochlear implants. Experience of the Lyon Team" Second Int. Symp. on cochlear implants, Paris, 22-24 sept. 1983, Proc. pp 195-203.

[4] R. Mouhssine, "Analysis of the spectral information produced by the implanted cochlear prosthesis Chorimac", Fourth Spanish-Portuguese-Moroccan workshop on signal processing and its applications, Marrakech, 24-26 sept. 1984, Proc. part B.4.1.

[5] J.P. Peckels, M. Rossi, "Le test de diagnostic par paires minimales", Deuxièmes J.E.P., Aix en Provence, 1-2 avril 1971, Proc. part B-c.

****
* PRODUCTION *
****

LA COPRODUCTION DANS LES GROUPES D'OCCLUSIVES

Alain Marchal

Institut de Phonétique - UA 261 - CNRS
29, Ave. R. Schuman - 13621 Aix-en-Provence

Coarticulation has often been viewed as an accommodative process which serves to smooth out differences between adjacent sounds. Using electropalatography, we look at the spatio-temporal organization of tongue contacts for sequences of stops. Our articulatory and acoustic data indicate that the coarticulatory pattern would be better understood if one aknowledges the existence of a true coproduction process. This principle accounts in a much simpler and natural way for a greater variety of phonetic realizations.

Les théories traditionnelles de la coarticulation considèrent que les phonèmes s'ajustent aux bornes pour permettre le passage harmonieux d'un phonème à un autre. Celles-ci tiennent la coarticulation comme un phénomène phonologique purement abstrait qui peut consister dans le transfert partiel ou total de traits segmentaux. Les résultats de la présente étude montrent que la coextension temporelle entre les "segments" et les traits est un artifice théorique. Celui-ci est démenti par l'analyse des palatogrammes dynamiques et du signal acoustique correspondant des groupes $V_1C_1C_2V_2$. Nous montrons à propos de l'enchaînement de groupes d'occlusives qu'une théorie de la coproduction rend plus "naturellement" compte de l'organisation des gestes articulatoires.

Les différents et innombrables travaux sur la coarticulation ont été élaborés et peuvent s'interpréter dans le cadre de deux types de théories générales : les théories segmentales /1,2/ et les théories de la coproduction /3,4/.

Il est admis que les phonèmes s'influencent réciproquement dans la parole. La question demeure de savoir si les modifications phonétiques observées relèvent d'un mécanisme d'ajustement contextuel ou plutôt d'un mécanisme de coproduction. Dans le premier cas l'unité minimale de programmation est le segment, Kent et Minifie/5/ "Coarticulation is a conceptualization of speech behavior that implies discrete and invariant units serving as input to the system of motor control, and an eventual obscuration of the boundaries between units at the articulatory or acoustic levels". La chaîne parlée résulte d'une concaténation de matrices de traits phonétiques segmentaux ajustés aux bornes pour éviter les effets acoustiques perturbateurs. Dans la théorie de la coproduction, l'unité minimale est au moins la syllabe et on considère qu'il existe au niveau de la production un continuum articulatoire vocalique /6,7/ et éventuellement un continuum consonantique /8,9/. Les faits de coarticulation résultent de l'intéraction de deux systèmes neuromusculaires spécifiques aux voyelles et aux consonnes /10/ dans la réalisation du plan phonétique /11/. Il ne faut donc pas s'étonner de ne pas trouver des unités temporelles discrètes dans le continuum acoustique /4/ "Scientists have not discovered those properties in the acoustic signal, but the reason they have not may be that they have looked for evidence of the wrong kind. They have looked for temporal discreteness when they should have looked for qualitative separetness among temporally overlapping units".

Dans une théorie de type segmental, l'enchaînement de deux occlusives dans une séquence $V_1C_1C_2V_2$ devrait se traduire par l'exécution séquentielle des instructions suivantes : fermeture du tractus pour C_1, tenue de l'occlusion, relâchement articulatoire de C_1 suivi de la fermeture du canal aérifère pour C_2, la tenue et enfin le relâchement pour C_2.

De nombreuses observations en anglais et en français tendent à indiquer que l'organisation spatio-temporelle des événements articulatoires et acoustiques dans les groupes d'occlusives ne répond pas aux prédictions de ce modèle /12,13/. Ainsi, Catford /14/ observe un chevauchement des articulations "In heterorganic close transitions there is no articulatory continuity, but there is articulatory overlap. That is to say, the articulatory stricture for the second consonant is formed before the stricture for the first is released. English examples are [kp] in backpart, [tp] in cut-purse, [tk] in atkins, [zf] in these fairs, and so on. In all such cases, the second articulation is formed well before the

first is released".

Pour le français les avis divergent : Grammont/15/ et Rousselot/16/ considèrent qu'il y a toujours deux articulations successives tandis que Rochette relève plusieurs cas où le relâchement de C_1 n'intervient qu'après la mise en place du barrage occlusif de C_2.

Nous pouvons faire remarquer que les travaux consacrés à cette question n'ont jamais abordé que l'étude d'un seul aspect du problème à la fois, et qu'aucune recherche ne fait le lien entre les modifications de la forme de la cavité buccale et les éventuelles modifications de signal acoustique qui l'accompagnent.

Dans cette étude, nous avons examiné l'organisation spatio-temporelle des appuis linguo-palatins dans les groupes d'occlusives hétérorganiques en français et la structure acoustique détaillée du signal de parole correspondant.

La rencontre de deux consonnes de même mode ou de mode différent peut se produire dans le même mot avant ou après chute de "ə" Ex : tactique [aktik] ; Bagdad [agdad] ; acqueduc [akdy] ou le plus souvent à la rencontre de deux mots différents. La frontière syntaxique peut être plus ou moins forte. Daniloff et Hammarberg(1975) distinguent par exemple 7 types de frontière : a)à l'intérieur d'un mot.intrasyllabique, b)à l'intérieur d'un mot.extrasyllabique, c)dans un mot composé, d)à la frontière de mots dans un syntagme, e)à la frontière de syntagme-intraclause, f)frontière de syntagme-extraclause, g)à la frontière de phrase. Ex : un coup de tête [ydt] ; Jacques tire [akt] ; t'as passé le bac ? dis! [akdi] .

Notre corpus était composé de 120 phrases naturelles de 4 à 6 syllabes. Les groupes d'occlusives orales sous l'accent dans les mêmes mots ou des mots différents étaient précédés et suivis des voyelles a, i, u. Nous avons réalisé deux séries d'enregistrements simultanées et synchronisées des données EPG et acoustiques avec deux locuteurs (hommes de 35 et 40 ans) parlant un français exempt de traits régionaux marqués.

Etude articulatoire

L'analyse des appuis de la langue au palais a été réalisée à partir d'informations électropalatographiques.(Marchal,Floutier, 1984).On a utilisé un électropalatographe portable contrôlé par un microprocesseur Rockwell-Aim 65 qui permet l'acquisition de 100 "images" du palais par seconde. Les palais artificiels, fabriqués en acrylique, selon la méthode de la cire perdue, sont munis de 64 électrodes uniformément réparties sur la plaque palatine. Rappelons que dans ce système le locuteur est mis à la masse et que la fermeture du circuit langue-palais induit le passage d'un courant de 5uA.

Analyse acoustique

Nous avons utilisé concurremment les méthodes analogiques traditionnelles: sonographie(Kay Electric 6061 B. 0.8000Hz. bande large=300Hz, bande étroite=45Hz) et les méthodes de traitement de signal. Nous avons en particulier utilisé le logiciel ILS, commercialisé par Signal Technology (version 1V) pour obtenir les oscillogrammes de la parole numérisée; les formants, l'intensité, la fréquence fondamentale et l'évolution du spectre en fonction du temps (commandes : DSP.INA.API.SGM.FTR.SDI) La parole a été échantillonnée à 10KHz. Le contexte était de 100 points par trame. Nous avons enfin mesuré l'énergie moyenne et calculé la répartition de l'énergie à l'aide d'un programme d'analyse du LIMSI/17/.

Résultats

Notre étude exclut délibérement toute interprétation statistique. Nous n'avons pas en effet cherché à savoir quel était le mode le plus fréquent de production de cette séquence de phonèmes. Nous avons voulu au contraire essayer de dégager toute les stratégies possibles. Leur examen a permis de mettre en évidence 4 classes de réalisations différentes de C_1C_2.
1)une seule tenue articulatoire/une seule tenue acoustique - 2)une seule tenue articulatoire/présence d'un bruit faible au relâchement de C_1 - 3)deux tenues articulatoires consécutives/bruit au relâchement de C_1 - 4)deux consonnes indépendantes/explosion entre C_1 et C_2 .
La dernière classe (4)fait apparaître l'absence du phénomène de coarticulation linguale. Il y a une organisation séquentielle et discrète des gestes articulatoires requis dans la production de C_1 et C_2. Ce type de réalisation caractérise des suites C_1C_2 où les deux consonnes sont séparées par une frontière majeure de phrase.

Pour les autres cas (1.2.3)il y a derrière la variabilité superficielle observée le même type d'organisation des appuis linguo-palatins. Le principe de coproduction permet d'en rendre compte.L'enchaînement de [t] à [k]dans "il n'y a pas plus de quat(re) cas" (Fig.1.)illustre bien l'organisation spatio-temporelle des appuis de la langue au palais dans une suite d'occlusives. Il met notamment en évidence :
-l'anticipation de l'articulation de C_2 pendant la tenue de C_1
-la présence d'une phase de double occlusion C_1C_2
-l'apparition d'un "clic" au relâchement de C_1.
Les deux premiers phénomènes sont l'expression de la coproduction de deux consonnes. On peut essayer de l'interpréter ainsi : un groupe de deux occlusives n'est pas la simple concaténation de deux consonnes. La commande (consonantique) provoque une élévation relativement indifférenciée de la langue. Lorsque les appuis latéraux sont établis, la spécificité de C_1 et C_2 intervient mais celle-ci est limitée pour assurer la différenciation acoustique au moment de

l'implosion. Dès que le barrage de C_1 est réalisé, la langue peut se déplacer librement pour mettre en place le barrage de C_2. Cette anticipation peut dans certains cas (1 et 2) provoquer l'apparition d'une phase de double occlusion.

Ces observations invitent à reconsidérer la notion de coarticulation.Celle-ci ne peut être assimilée à un simple ajustement des segments aux bornes. Il convient tout autant de rejeter la conception traditionnelle et naïve selon laquelle le segment peut être représenté par une matrice de traits exclusivement segmentaux et réalisés simultanément. La production de la parole est d'abord un phénomène temporel : les gestes articulatoires ne peuvent être définis de façon statique. Chaque articulateur possède sa dynamique propre. L'interrégulation des activités musculaires est assuré par un ensemble de "structures coordinatives". La coproduction des voyelles et la coproduction des consonnes met en évidence ce mécanisme. Afin de parvenir à un plus grand degré d'adéquation descriptive et d'adéquation explicative, il faut modifier les systèmes actuels de traits afin qu'ils tiennent compte de ces propriétés essentielles et universelles de l'appareil de production.

Références bibliographiques

/1/ R.G. Daniloff & R. Hammarberg, "on defining coarticulation", j.of.phon, 1, pp 239-248, 1973.

/2/ R. Hammarberg, "on redefining coarticulation", j.of phon.,10,pp 123-137,1982

/3/ C. Fowler, Timing control in speech Production, Indiana University Linguistics Club, Bloomington, (1977).

/4/ C. Fowler, "Coarticulation and theories of extrinsic timing", j.of phon., 8, pp 113-133, 1980.

/5/ R. Kent & F.D. Minifie, "Coarticulation in recent speech production models", j. of phon., 5, pp 115-133, 1977.

/6/ V.A. Kozhevnikov & L.A. Chistovich, Speech : articulation and Perception, Transl, US Dept. of Commerce, (1965)

/7/ S. Ohman, "Coarticulation in vcv utterance : spectrographic measurements" j.acoust.soc.Am, 39, pp 151-168, 1966.

/8/ K. Stevens & A.S. House, "Perturbation of vowel articulations by consonantal context : an acoustical study" j.speech & hear.res.,-, pp 111-128, 1963

/9/ A. Marchal, "Coarticulation or coproduction", Revue d'acoustique, 4, pp 255-258, 1983.

/10/J.S. Perkell, Physiology of speech production : results and implications of a quantitative cineradiographic study, MIT Press, Cambridge, (1969)

/11/P. Linell, "The concept of phonological form and the activities of speech production and speech perception", j.of phon., 10, pp 37-72, 1982.

/12/W.J. Hardcastle & P. Roach, "An instrumental investigation of coarticulation in stop sequences, W.in Progress, phonetics lab., Reading, 1,pp 27-44, 1977.

/13/Cl. E. Rochette, Les groupes de consonnes en français, Klincksieck, Paris, (1973).

/14/J. C. Catford, Fundamental problems in phonetics, E.U.P., Edinburgh,(1977).

/15/M. Grammont, Traité de Phonétique, Delagrave, Paris,(1933)

/16/J. Rousselot, Principes de phonétique expérimentale, Welter, Paris (1901).

/17/F. Manceron, Contribution à l'analyse spectro-temporelle du signal de parole considéré comme une unité d'impulsions acoustiques-Thèse de docteur ingénieur, LIMSI, Paris (1982).

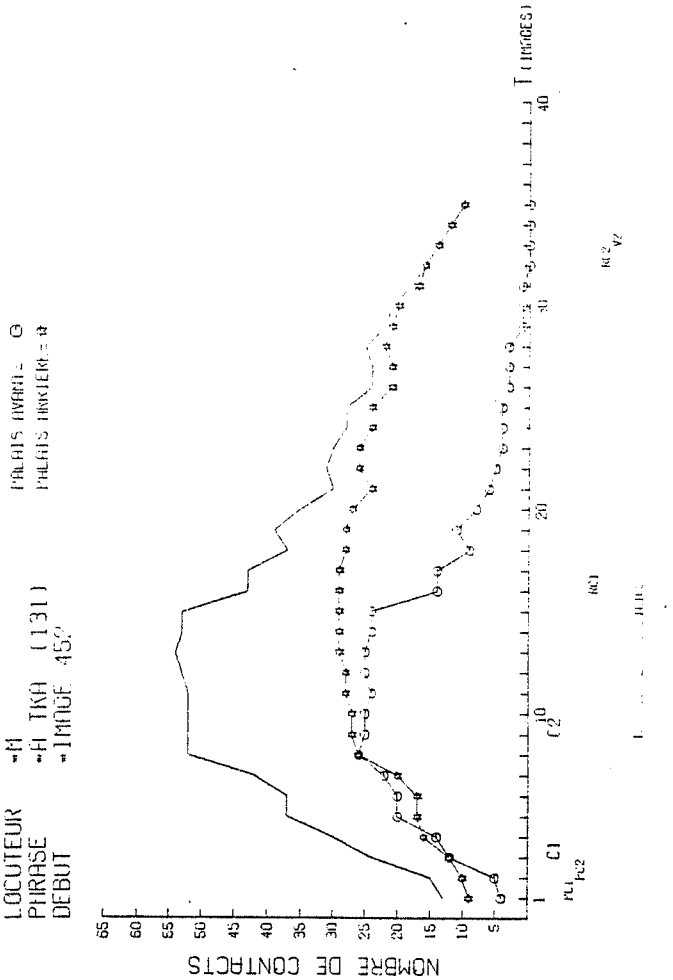
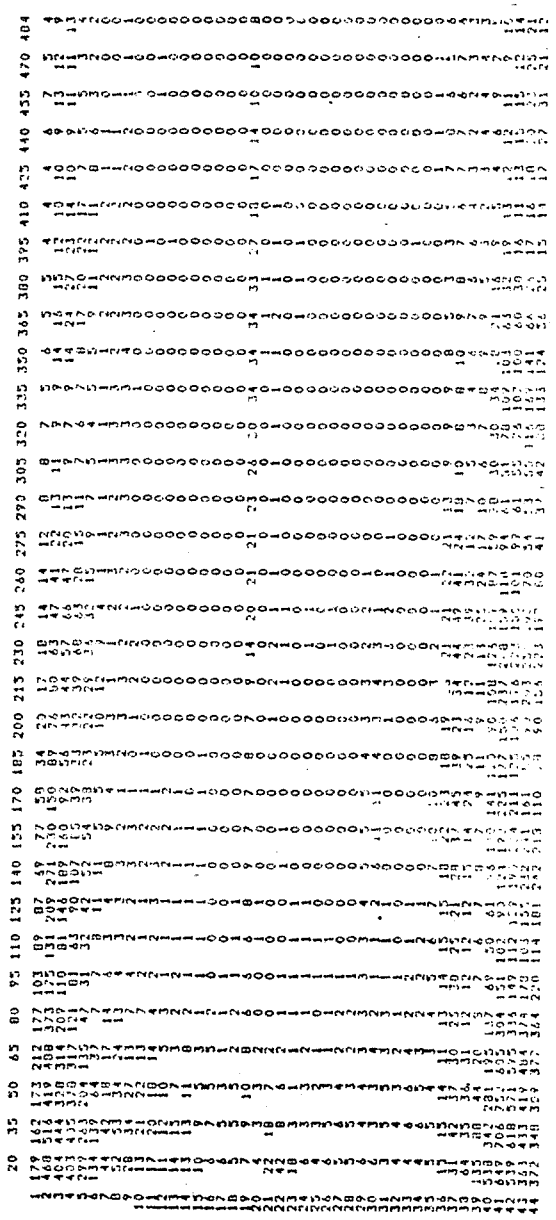


Fig.1.évolution des appuis linguo-palatins dans le groupe tk

PC_1;PC_2 = préparation C_1 ; C_2
C_1 ;C_2 = tenue C_1 ; C_2
C_1 C_2 = double occlusion
RC_1 ;RC_2= relâchement C_1 ; C_2



Tab.1. répartition de l'énergie dans le groupe tk .
clic à la trame 19 correspondant au relâchement de C_1

L'ACTIVITE MUSCULAIRE DE TROIS ORGANES ARTICULATOIRES : ORGANISATION TEMPORELLE

Michèle GENTIL & Thomas GAY

INSERM - U3, 47 bd de l'Hopital, 75651 Paris Cédex 13, France &
The University of Connecticut Health Center, Farmington, Con.06032,USA

ABSTRACT

The present study aims at exploring the muscular coordination of the labial, lingual and mandibular systems. To this end, the activity patterns of three muscles of the lips, three muscles of the tongue and one muscle of the jaw were recorded along with the acoustic signal during consonant-vowel-consonant syllables produced at two different speaking rates by an American-English subject. Results indicated long anticipatory coarticulatory effects for the three articulatory organs and similar muscular patterns across the two speaking rates.

INTRODUCTION

Cette étude a pour but d'examiner la coordination musculaire de trois organes articulatoires : lèvres, langue et mâchoire, considérant à la fois des unités segmentales et syllabiques. Plusieurs organes articulatoires étant pris en compte, la coarticulation est observée au niveau de l'organisation temporelle de l'activité musculaire des divers articulateurs.

Avec un tel objectif, l'activité de trois muscles des lèvres, trois muscles de la langue et un muscle de la mâchoire était enregistrée simultanément avec le signal acoustique. Des syllabes consonne-voyelle-consonne étaient produites par un sujet anglais américain à deux différents débits de parole (conversationnel et rapide). Nous analysions :

- L'organisation temporelle intrasegmentale, c'est-à-dire pour chacun des organes articulatoires et à chaque vitesse de parole les effets coarticulatoires des diverses activités musculaires en relation avec le signal acoustique.
- L' organisation temporelle intrasyllabique, c'est-à-dire considérant tous les organes articulatoires,l'agencement de l'activité musculaire au cours de la réalisation de deux segments phonétiques (consonne-voyelle et voyelle-consonne), cet examen étant effectué à chaque débit de parole.

METHODE

Un adulte mâle anglais américain servait de sujet dans cette expérience.

Le matériau phonétique comportait dix monosyllabes de type consonne-voyelle-consonne, les consonnes étant /p - b - w - t - d - s/ et les voyelles /i - a/. Chacune de ces productions était introduite dans une phrase porteuse "It is ....again" et distribuée au hasard dans plusieurs listes. Elle était répétée vingt fois à deux débits de parole (conversationnel et rapide).

Au moyen d'électrodes à crochets (1), on enregistrait l'activité de sept muscles. Trois de ces muscles étaient associés aux mouvements des lèvres supérieure et inférieure. Ce sont l'orbiculaire supérieur (OOS), le mentalis (MENT) et le dépresseur (DLI). Trois autres muscles étaient en relation avec le mouvement de la langue : génioglosse (GG), longitudinal supérieur (SL) et inférieur (IL). Enfin un muscle, le ventre antérieur du digastrique, intervenait dans l'abaissement de la mâchoire.Ces signaux electromyographiques étaient enregistrés simultanément avec le signal acoustique.

Le traitement des données était effectué sur ordinateur du laboratoire du département d'"Oral Biology". Les potentiels electromyographiques étaient redressés et intégrés. A partir du disque analogique, les données étaient entrées répétition par répétition en utilisant des fenêtres de 2.5 secondes. Pour chacune des répétitions enregistrées, on déterminait un repère (line-up) sur le signal acoustique. Celui-la correspondait à l'entrée de l'explosion acoustique de la première consonne. Enfin de manière à limiter le problème de variabilité entre les diverses répétitions toutes les données étaient moyennées.

RESULTATS

Notre analyse des aspects temporels de l'activité musculaire comportait deux parties, l'une concernant l'organisation temporelle intrasegmentale, l'autre concernant l'organisation temporelle intrasyllabique.

Organisation temporelle intrasegmentale

L'étude de l'organisation intrasegmentale considère séparément chaque segment phonétique de la syllabe consonne-voyelle-consonne. Afin d'examiner les effets coarticulatoires, nous observions les différents modèles d'activité musculaire pour chacun des organes articulatoires et calculions la différence de temps entre l'entrée de l'activité de chacun des muscles et l'entrée du signal acoustique correspondant. Ces calculs étaient faits pour les deux débits de parole respectivement.

Cette approche est illustrée en Figure 1 qui présente pour la production /pip/ (dans les deux vitesses de parole) les entrées des divers muscles correspondant à chacun des trois segments de cette syllabe. Les mesures sont faites depuis la line-up qui sert de référence. Les différents symboles qui sont utilisés sont les suivants: triangle = entrée des muscles de la langue, cercle = entrée des muscles des lèvres, diamant = entrée du signal acoustique. Nous remarquons que l'activité musculaire commence toujours longtemps avant le signal acoustique dans les deux débits de parole c'est-à-dire pour les muscles des lèvres approximativement 95 msec en vitesse conversationnelle et 80 msec en vitesse rapide. En comparant les deux débits de parole, on peut voir que l'entrée de l'activité musculaire est étroitement liée à l'entrée du signal acoustique. Ainsi en vitesse conversationnelle, l'entrée du signal acoustique pour le /p/ initial est plus éloignée de la line-up qu'en vitesse rapide, de même les entrées d'activités musculaires sont également plus éloignées de la line-up en vitesse conversationnelle qu'en vitesse rapide.

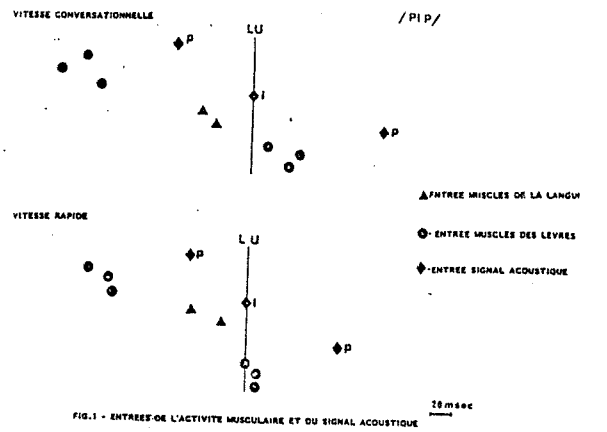


FIG.1 - ENTREES DE L'ACTIVITE MUSCULAIRE ET DU SIGNAL ACOUSTIQUE

Le tableau 1 rassemble les résultats concernant les paires minimales /pip/ et /bip/. Ce tableau présente les différences (en msec) entre les entrées de l'activité des différents muscles et les entrées respectives du signal acoustique pour chaque segment phonétique.Comme précédemment, toutes les entrées sont calculées depuis la line-up. Les signes négatifs ou positifs devant les valeurs signifient que les entrées sont situées respectivement à gauche ou à droite de la line-up. Considérant la première consonne de chaque production, nous notons pour la consonne sourde /p/ des entrées d'activité musculaire antérieures à celles observées pour la consonne voisée /b/. Toutefois, l'entrée en activité du génioglosse est temporellement identique après consonne sourde et consonne voisée quel que soit le débit de parole. Ceci laisse à penser que le VOT est en relation uniquement avec le mouvement de la langue et ne dépend pas de la coordination consonantale.

STIM	VIT.	AC SIG. ENT.	OOS ENT.	OOS DIFF.	INFT ENT.	INFT DIFF.	OLI ENT.	OLI DIFF.	GG ENT.	GG DIFF.	IL ENT.	IL DIFF.
/PI/	CONV	- 66.6	- 162.8	96.2	- 185	118.4	-148	81.4				
	RAP.	- 55.9	- 154.6	98.7	- 135	79.1	-133	77.1			-120	64.1
/I/	CONV	L.U.							- 51.8	51.8	- 37	37
	RAP.	L.U.							- 55	55	- 25.8	25.8
/P/	CONV	+125.8	+ 14.8	111	+ 44.4	81.4	+ 37	88.8				
	RAP.	+ 85	L.U.	66	+ 8.6	77.4	+ 8.6	77.4			+ 60.2	25.8
/B/	CONV	- 31.4	- 185	103.6	- 181.5	99.9	- 148	66.6			125.8	44.4
	RAP.	- 77	- 159	81.6	- 129	52	- 120	43			107	30
/I/	CONV	L.U.							- 52	52	- 29.6	29.6
	RAP.								- 51.6	51.6	8.6	8.6
/P/	CONV	+ 88.8	+ 22.2	66.6	+ 37	51.8	+ 33.3	55.5			+ 70	18.8
	RAP.	+ 73	+ 8.6	64.4	+ 21.5	51.5	+ 12.9	60.1			+ 34.4	38.6

TAB.1 - RELATIONS ENTREES ACTIVITES MUSCULAIRES ET SIGNAL ACOUSTIQUE - /PIP/ & /BIP/

- A GAUCHE DE LA LINE UP
+ A DROITE DE LA LINE UP

Organisation temporelle intrasyllabique sur deux segments phonétiques

Dans une deuxième analyse, nous examinions le chevauchement temporel des activités musculaires relatives à des segments individuels à l'intérieur des diverses syllabes. Nous calculions :

- Le temps entre la fin d'activité des muscles associés au mouvement pour la première consonne et l'entrée en activité des muscles intervenant dans le mouvement pour la voyelle.

- Le temps entre la fin d'activité des muscles liés au mouvement pour la voyelle et l'entrée en activité des muscles intervenant dans le mouvement pour la seconde consonne.

En ce qui concerne la monosyllabe /pip/ en vitesse conversationnelle, ces divers calculs de temps conduisaient à la représentation schématique de la figure 2. L'activité du longitudinal supérieur et la seconde activité du génioglosse correspondent au mouvement pour le /s/ et pour le /g/, rappelons que /pip/ est produit dans la phrase porteuse /itspip›g£n/.

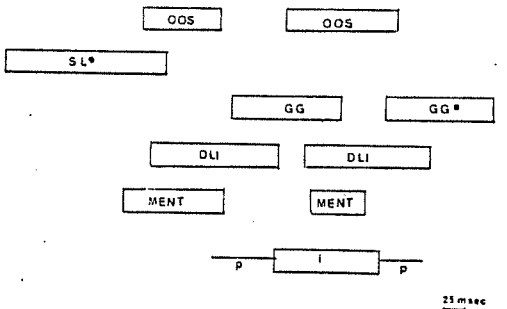


FIG.2 - ACTIVITE INTERMUSCULAIRE POUR /PIP/ DEBIT CONVERSATIONNEL

* Activité musculaire correspondant au mouvement pour /s/ & /g/ (phrase porteuse)

Les chevauchements temporels d'activités musculaires pour cette production dans les deux débits de parole sont indiqués dans le tableau 2 qui présente également les chevauchements d'activités musculaires pour la production /bip/ à des fins de comparaison.Durant la production de /pi/, /bi/ et /ip/, l'orbiculaire supérieur, le mentalis et le dépresseur assurent le déplacement des lèvres pour les consonnes et le génioglosse est actif dans les mouvements de la langue pour la production de la voyelle. Ainsi nous mesurions : (1) Le temps entre l'entrée en activité du génioglosse et la fin d'activité de l'orbiculaire supérieur, le mentalis et dépresseur pour les syllabes /pi/ et /bi/,. (2) Le temps entre l'entrée en activité de l'orbiculaire supérieur, mentalis, dépresseur et la fin d'activité du génioglosse pour les syllabes /ip/.

Les astérisques signalent un interval plutôt qu'un chevauchement. Ce tableau révèle des modèles d'activité intermusculaire similaires dans les deux vitesses de parole : nous trouvons des degrés semblables de chevauchement et interval musculaires. Toutefois, des variations dans la durée des chevauchements sont évidentes en fonction des changements de vitesse de parole. Ces variations sont liées au type de syllabes. Dans les syllabes voyelle-consonne, /ip/, la durée de chevauchement est plus grande en vitesse conversationnelle qu'en vitesse rapide. En ce qui concerne les syllabes consonne-voyelle, /pi/ et /bi/, au contraire nous notons un chevauchement temporel plus court en vitesse conversationnelle comparée à vitesse rapide. Ce dernier effet va dans le sens de celui prédit pour les modèles de production de parole montrant un chevauchement temporel ou coarticulation qui s'accroit lorsque le débit de parole s'accroit (2), (3).

TABLE 2 - ORGANISATION TEMPORELLE INTRASYLLABIQUE

	OO & GG		MENT & GG		DLI & GG	
	CONV.	RAP.	CONV.	RAP.	CONV.	RAP.
/pi/	15.0*	5.2	7.4*	13.8*	59.2	64.0
/ip/	70.0	52.0	40.6	43.4	48	43.4
/bi/	7.5	17.6	10.9*	17.2*	59.4	62.2
/ip/	74.0	47.3	59.2	34.4	62.9	43.0

*Interval et non chevauchement

CONCLUSION

L'analyse des modèles d'activité musculaire correspondant aux mouvements des trois organes articulatoires : lèvres, langue et mâchoire conduisent à deux constatations générales.

- L'activité des muscles commence longtemps avant le signal acoustique correspondant. L'anticipation musculaire est plus longue et plus variable pour les muscles des lèvres en comparaison des muscles de la langue. Ces observations sont valables pour les deux différents débits de parole, quoique la durée d'anticipation au niveau musculaire soit plus grande en débit conversationnel.

- La coordination musculaire des trois systèmes : labial, lingual et mandibulaire présente des modèles d'activité similaires dans les deux débits de parole avec cependant des variations dans la durée de chevauchement des activités musculaires.

REFERENCES

(1) J.V. BASMAJIAN & G.A. STECKO, "A new bipolar indwelling electrode for electromyography", J.Appl.Physiology, Vol.17, p 849, 1962

(2) V.KOZHEVNIKOV & L.CHISTOVICH, "Speech : Articulation and Perception", Moscow-Leningrad (English translations JFRS, Washington DC No JFRS 30453), 1965

(3) T. GAY,"Cinefluorographic and electromyographic studies of articulatory organization" in Dynamic Aspects of speech production ed. by Masayaki Sawashima & Franklin Cooper, Uni of Tokyo press, pp 85-105, 1977

ANALYSE INTER-LOCUTEURS DU COMPORTEMENT DE L'OS HYOIDE EN CHAINE PARLEE

André Bothorel

Institut de Phonétique
Université des Sciences Humaines - STRASBOURG II

In this paper, we present the hyoid bone associated with articulatory gestures. We find that its vertical movements do not vary from subject to subject for the same vowel. Its horizontal movements are related to those of the root of the tongue. We discuss these movements with respect to articulation, not to pitch control mechanism.

INTRODUCTION

Le comportement de l'os hyoïde est un sujet controversé dans la mesure où, selon certains chercheurs, ses mouvements varient en fonction des sujets pour les mêmes voyelles (Söderberg [1], Rossi et Autesserre [2]), alors que pour d'autres -les plus nombreux- sa position varie en fonction des voyelles pour un même sujet (Perkell [3], Menon et Shearer [4], Bothorel [5], entre autres): "With respect to articulation, its position is higher in the vowel /a/ than in the vowel /i/... The extent of the vertical movement was found to be larger in vowel change than in pitch change" (Honda 1983, p. 276) [6].

Notre contribution, fondée sur l'analyse du comportement de l'os hyoïde en chaîne parlée à partir de l'exploitation de radiofilms -quatre sujets français, corpus identique-, tentera de donner quelques indications supplémentaires sur les déplacements verticaux et horizontaux de cet os (élément isolé dans le squelette humain) qui sert de support à la base de la langue et dont les mouvements ne peuvent être qu'en rapport avec la nature des réalisations articulatoires.

PROTOCOLE EXPERIMENTAL

Nous disposons de films radiologiques (50 images/secondes, réalisés en 1982) de quatre sujets français, à savoir deux masculins (L.K., né en 1957 -Bas-Rhin-, et A.D., né en 1957 à Bordeaux où il a vécu jusqu'à 20 ans) et deux féminins (D.F. née en 1957 -Haut-Rhin-, et P.B. née en 1954 à Fribourg-en-Brisgau). Les quatre sujets ont le français comme langue maternelle.

Le corpus est constitué de courtes phrases du type: "ma chemise est roussie"
"six beaux tapis"
"il fait des achats"
"la petite massue"
"donne un petit coup"...
Les voyelles dites accentuées (en dernière syllabe de phrase) sont les suivantes: /i,y,u,e,a/. Elles sont précédées des consonnes /p,t,k,b,d,g,f,s,ʃ,ʒ/. Les syllabes accentuées sont du type CV (cinq exemples seulement sont du type CVC).

Les mesures (en grandeur réelle) ont été prises au point d'intersection indiqué sur la Fig. 1 à partir d'une projection directe des radiofilms, deux ou trois images avant le début de l'abaissement du voile du palais marquant la fin de la phrase. Quelques relevés de la position de l'os hyoïde ont été effectuées pour certaines voyelles inaccentuées, à savoir /u,ɛ,ɑ̃/. Le nombre des occurrences s'élève à une quinzaine environ pour chacune des voyelles accentuées.

RESULTATS

1. Le déplacement vertical de l'os hyoïde en chaîne parlée.

1.1. La position de l'os hyoïde pour chaque voyelle.

Le déplacement vertical de l'os hyoïde

n'est pas erratique: le relevé des mesures effectuées montre qu'il est en position la plus élevée en moyenne pour la voyelle accentuée /a/ par rapport aux voyelles accentuées /i,y,u/, cette dernière étant en position la plus basse pour les quatre sujets. Si les relevés que nous présentons en Fig.2 (sujet masc. A.D.) indiquent bien une position moyenne plus élevée pour /a/ par rapport à /u/, on doit convenir qu'ils ne sont pas très significatifs. Chez les trois autres sujets, les différences sont plus nettes.

En Fig. 2 apparaît également la différence importante qui existe entre les voyelles /u/ accentuées (position basse de l'os hyoïde) et inaccentuées (position plus élevée). Les quatre exemples de /u/ inaccentué sont tirés des phrases suivantes: "voilà des bougies", "un tour de magie", "un fourré touffu" et "il part pour Vichy".

1.2. L'amplitude de la dispersion des mesures relevées pour chaque voyelle.

Le relevé des mouvements de l'os hyoïde étant fait sur deux axes (vertical et horizontal), nous pouvons nous rendre compte de la dispersion des mesures sur les graphiques que nous établissons. Ainsi, en Fig.2, la dispersion est plus grande pour la voyelle /a/ que pour la voyelle /u/ (sujet A.D.). A l'examen de la dispersion des mesures pour les quatre sujets, les tendances suivantes apparaissent: en moyenne la dispersion est moins importante pour les voyelles /a,u/ que pour les voyelles /y,i/.

En considérant la consonne qui précède ou suit la voyelle, nous avons noté qu'elle influençait jusqu'à un certain point la position verticale et horizontale de l'os hyoïde: c'est ainsi que les voyelles au contact de /s/ se trouvent au sommet du "nuage" et le plus éloignées des vertèbres alors que celles qui sont au voisinage de /p,b,g/ se situent dans sa partie basse et le plus proches des vertèbres. Cette répartition est vérifiée pour chaque voyelle et pour chaque sujet. Ajoutons que dans les cas de "Vichy" et d'"ici", la voyelle /i/ se situe dans une position intermédiaire au plan vertical et plus avancée (plus éloignée des vertèbres au plan horizontal.

1.3. Rapports entre le déplacement vertical de l'os hyoïde et du larynx.

Les mesures que nous avons effectuées sur un seul sujet (A.D.) montrent que le déplacement vertical du larynx correspond à celui de l'os hyoïde, ainsi que l'ont vérifié la plupart des chercheurs: les deux structures sont en position la plus haute pour /a,ɑ̃/ et en position basse pour /i,y,u/. En prenant en compte la nature de la consonne avoisinante, nous avons observé que le déplacement du larynx et le mouvement de l'os hyoïde coïncidaient: c'est ainsi que, pour la voyelle /u/, les deux structures sont en position la plus élevée dans les exemples qui comportent /s/, comme "à sous" ou "secousse", et en position la plus basse dans les exemples comme "du bout" "boubous", "du coup","hibou".

Il nous apparaît nettement qu'en ce qui concerne le déplacement vertical du larynx et de l'os hyoïde, il y a "analogie de comportement", "similitude de mouvement", "rapport d'interdépendance" [7].

2. Le mouvement horizontal de l'os hyoïde en chaîne parlée.

Les résultats que nous avons obtenus concernant le déplacement horizontal de l'os hyoïde montrent que ce mouvement est en rapport avec la nature de l'articulation réalisée: pour les voyelles fermées du type /i,y,u/, l'os hyoïde s'éloigne des vertèbres, alors que pour les voyelles du type /a,ɑ̃/ il s'en rapproche. Ceci est en accord avec la position de Rossi et Autesserre [2] qui, réinterprétant les tableaux de Bothorel [5] précisaient: "The dividing line... comes not between front and back vowels but between the two classes (ɑ,ɑ̃) and (i,u). The position of the hyoid bone on the horizontal axis is linked not to that of the back of the tongue but to that of the tongue-root" (pp.236-37). C'est également l'opinion de Honda [6](p.274): "...high vowels are associated with forward position and low vowels with back position of the hyoid bone due to anatomical connections with the tongue-root."

DISCUSSION

Analysé du point de vue de sa fonction dans la réalisation articulatoire des sons de la parole, le comportement de l'os hyoïde n'est pas indifférent, soumis au hasard. On comprendra d'autant mieux son comportement que sera approfondie la nature de ses rapports avec les autres articulateurs auxquels, par des muscles et ligaments divers, il est anatomiquement relié. L'ambiguïté de son statut vient du fait qu'il se situe dans une position intermédiaire entre la langue (et les autres articulateurs) élaborant des mouvements articulatoires et le larynx, siège de la phonation: "its position is affected by both pitch control and articulatory gestures" (Honda [6], p.269).

En considérant le déplacement vertical de l'os hyoïde, nous pouvons dire, à l'examen de nos documents et en accord avec la plupart des chercheurs, qu'il est relativement stable pour une voyelle donnée quel que soit le sujet: il est en position la plus haute pour les voyelles les plus ouvertes du type /a,ɑ̃/ et en position la plus

basse pour les voyelles les plus fermées du type /i,y,u/. Le rapport est ainsi inversé entre la hauteur de la langue dans la cavité buccale pour la réalisation des voyelles et la position de l'os hyoïde.

Que la voyelle soit accentuée ou inaccentuée n'est pas indifférent. Ainsi que nous l'avons observé (voir Fig.2), les voyelles en position inaccentuée ne se superposent pas sur les voyelles accentuées: n'ayant pas le temps d'atteindre laur position-cible, elles sont davantage soumises aux influences de l'entourage consonantique. De ce fait, l'analyse du comportement de l'os hyoïde en chaîne parlée doit tenir compte du caractère accentué ou inaccentué des articulations considérées.

D'autre part, l'examen de la dispersion des mesures révèle un autre aspect qu'il convient également de souligner: la position de l'os hyoïde ne sera pas la même pour une voyelle au contact de /s,ʃ/ que pour la même voyelle au contact de /p,g,b/. Les premières consonnes relèveront sa position, les dernières l'abaisseront. Comme pour les voyelles, l'os hyoïde a un comportement déterminé selon la nature articulatoire de la consonne: il est plus relevé pour /s,ʃ/ et les consonnes non-voisées en général, plus bas pour les consonnes voisées correspondantes. En cas de contiguïté, la coarticulation joue, selon des règles qui peuvent être établies sans trop de difficultés. C'est ainsi que la tendance au relèvement de l'os hyoïde pour /s,ʃ/ est contrebalancée par l'influence abaissante des deux /i/ environnants dans les cas de "ici" ou "Vichy" et sa position pour le /i/ accentué sera alors intermédiaire entre le /i/ relevé de "hachis", "esquisse" et celui abaissé de "gui", "bougies".

En rassemblant les différents résultats obtenus,nous disposons d'un certain nombre d'indications qui sont de nature à nous permettre de privilégier quelques hypothèses d'explication, parmi celles qui ont déjà été proposées. Pour rendre compte du comportement inverse du complexe os hyoïde/larynx par rapport à la position de la langue dans le sens vertical, Honda [6] propose l'argumentation suivante: "Larynx elevation in low vowels is suggested to be due to hyoglossus muscle activity. Contrarily, larynx depression might be caused indirectly by the transformation o- tongue tissue. Contraction of the posterior fibers of the genioglossus raises the tongue dorsum and at the same time pushes the hyoid bone and the tongue base downward, since the insertion point of the posterior fibers of the genioglo-sus is just above the hyoid bone. The volume of the tongue mass being constant, decreases in the horizontal dimension of the tongue result in increase in its vertical dimension, both raising the dorsum of the tongue and lowering its base. This transformation of the tongue seems to be primarily relevant to vertical movement of the larynx in vowel articulation."(p.280).

En ce qui concerne son mouvement horizontal, l'os hyoïde est en position avancée (éloignée des vertèbres) pour les voyelles fermées, de petite aperture, du type /i,y,u/ et reculée pour les voyelles /a,ɑ̃/. Lorsque la racine de la langue est tirée vers l'avant pour la réalisation des voyelles "hautes", l'os hyoïde, qui en dépend, est également entraîné vers l'avant. La même explication convient pour les consonnes du type /s,ʃ,ʒ/ qui l'attirent aussi vers l'avant. Pour les voyelles de type /a,ɑ̃/, la racine de la langue est reculée, plus proche des vertèbres, aussi l'os hyoïde est-il également plus reculé (Fig.2).

CONCLUSION

Les résultats que nous avons présentés montrent que l'os hyoïde, même en chaîne parlée, a un comportement bien défini en fonction des articulations réalisées, explicable en termes de rapports de force au niveau articulatoire (essentiellement en considération des positions de la langue). Il reste qu'un autre aspect (que nous n'avons pas développé) est à considérer pour bien comprendre la fonction de l'os hyoïde: c'est celui du rôle qu'il joue dans le mécanisme de contrôle de la fréquence fondamentale, rôle qui n'a pas encore été défini avec précision.

La prise en compte du comportement de l'os hyoïde dans l'élaboration des modèles articulatoires ne doit pas être négligée dans la mesure où, sans être un articulateur actif, il participe efficacement à la redondance attachée à la réalisation des sons de la parole.

[1] Söderberg, C.-G. 1977, "Displacement of the tongue-hyoid-larynx column" in Provincial and universal linguistic themes, Congratulatory Essays to K.H. Dahlstedt (Otterbjörk,R. et Sjöström,S. éd.), Umeâ University, pp. 137-150.

[2] Rossi, M. et Autesserre, D. 1981, "Movements of the hyoid and the larynx and the intrinsic frequency of vowels", Journal of Phonetics 9/2, pp. 233-249.

[3] Perkell, J.S. 1969, "Physiology of speech production: results and implications of a quantitative cineradiographic study", The MIT Press.

[4] Menon, K.M.N. et Shearer, W.N. 1971, "Hyoid position during repeated syllables", Journal of Speech and Hearing Research 14, pp. 858-864.

[5] Bothorel, A. 1975, "Positions et mouvements de l'os hyoïde dans la chaîne parlée", Travaux de l'Institut de Phonétique de Strasbourg 7, pp. 80-132.

[6] Honda, K. 1983, "Relationship between pitch control and vowel articulation", Haskins Lab. S.R. on Speech Research 73, pp. 269-282.

[7] Bothorel, A., Brock, G. et Maillard-Salin, G. 1980, "Contribution à l'étude des rapports entre les mouvements de l'os hyoïde et le déplacement du larynx", Travaux de l'Institut de Phonétique de Strasbourg 12, pp. 225-269.

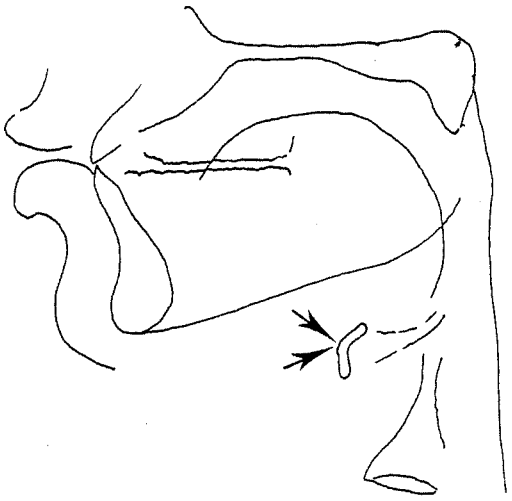


Fig. 1. Voyelle /u/ dans "sous" (sujet A.D.). Les flèches indiquent le point qui a été pris comme base pour le relevé des mesures.

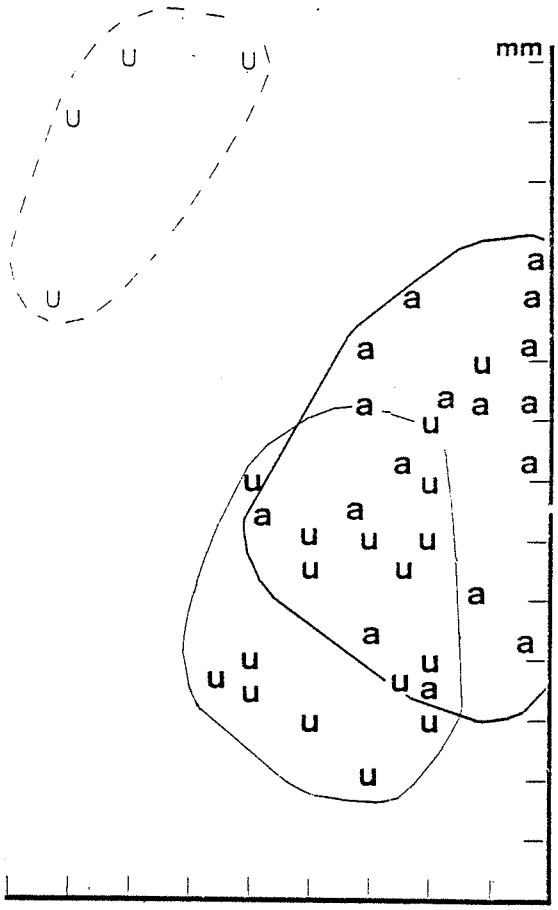


Fig. 2. Relevé des mesures du mouvement de l'os hyoïde pour les voyelles accentuées /a/ et /u/. En haut, mesures pour /u/ en position inaccentuée. (sujet A.D.)

VERS LA DYNAMIQUE: ETUDE DE QUELQUES TRANSITIONS ARTICULATOIRES, |VCV|.

H. Sanchez-Salguero[1], T. Gay[2] & L.J. Boë[1]

[1]Institut de la Communication Parlée - GRENOBLE
Laboratoire Associé au CNRS n° 368
[2]Health Center - FARMINGTON - Connecticut - USA

Summary :

In this article, we shall study some maximal articulatory transitions obtained from radiocinematographics (64 images/-sec). These radiocinematographics, constitute our starting point. Therefore the passage from the sagittal view to area function and the simulation of the vocal tract by an analog model, allowed us to obtain the evolution of formant values for each configuration. We will present, in brief, our experimental method.

A. INTRODUCTION.

Le processus de production de la parole est plus accessible au niveau acoustique qu'articulatoire et depuis longtemps des études ont été menées pour essayer d'induire les contraintes articulatoires à partir du signal acoustique (LEHISTE & PETERSON, 1961; STEVENS & HOUSE, 1955, 56).

De nombreux travaux ont été dirigés pour élaborer des modèles mathématiques du fonctionnement des paramètres acoustiques, caractérisant les fréquences formantiques en fonction d'un petit nombre de paramètres, (STEVENS, HOUSE & PAUL, 1966; LIBERMAN et al., 1954, 67; KELLY & LOCHBAUM, 1962), mais trouver des invariants dans les signaux acoustiques produits pour toutes les conditions possibles de la parole, est un problème extrêmement difficile, au vu des contraintes qui entrent en jeu au niveau neuromusculaire; il nous semble raisonnable de faire l'hypothèse que ces contraintes, qui ne sont pas très nettes sur le signal acoustique, puissent être spécifiées plus clairement à partir de la production.

La modélisation directe est une des techniques les plus employées pour l'étude de ce problème, avec généralement un modèle électrique analogue du conduit vocal, mais il faut faire évoluer l'aire de chaque section en fonction du temps. Le nombre des variables rend le contrôle difficile. Pour diminuer le problème on cherche à représenter l'aire par une spécification articulatoire simple.

On dispose de plusieurs types de modèles, mais nous avons préféré opérer à partir de mesures directes. Dans notre cas, nous disposons, à partir de cinéradiographies, de l'évolution temporelle du conduit vocal lors de la production de certaines réalisations. Il nous était facile de commander un modèle analogue (simulation par ligne électrique), avec les fonctions d'aire ainsi obtenues.

Le but de cette approche est d'essayer de simuler les configurations qui nous permettrons d'étudier les problèmes de certaines transitions, lesquelles ne sont pas simples à induire à partir du niveau acoustique.

Les études perceptives des laboratoires Haskins (COOPER & al., 1952; LIBERMAN & al., 1954; HARRIS & al., 1958; LIBERMAN et al. (1967), LIBERMAN & STUDDERT (1978) ont montré que le lieu d'articulation peut être spécifié uniquement à partir de l'information des transitions formantiques. Ces travaux ont été confirmés récemment par STEVENS & BLUMSTEIN (1978, 79, 80).

Le travail présenté ici concerne des transitions maximales entre les 3 voyelles cardinales |i|, |a|, |u| pour un locuteur anglophone.

B. LES DONNEES.

Grâce à une collaboration étroite avec le département de biologie orale de Farmington, il a été possible d'avoir un certain nombre de configurations articulatoires (36 par son), qui ont été acquises par cinéradiographie (64 images/s) pour les réalisations extrêmes |API|, |APU|, |ATI|, |IPA|, |IPU|, |ITA| de l'anglais.

C. LA METHODOLOGIE.

Un logiciel à été développé à l'Institut de Phonétique pour traiter les configurations articulatoires : la longueur des sections transversales, l'aire par section et l'aire frontale aux lèvres. La figure B.1 montre une configuration initiale et finale d'un des sons étudiés.

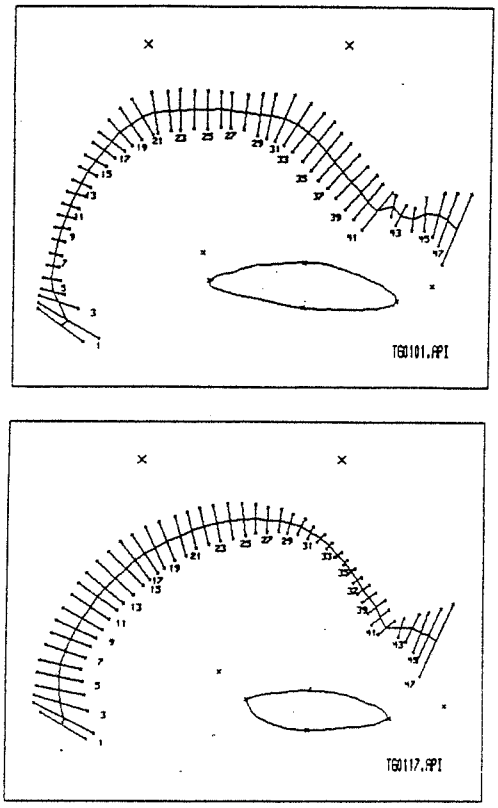


Figure B.1. Coupe sagittale du conduit vocal. Configurations initiale et finale de |API|.

Il est difficile de savoir sur un tracé radiologique avec exactitude où commencent et où finisent les différentes parties du conduit vocal (larynx, pharynx, cavité buccale, région alvéolaire, lèvres), après une analyse détaillée des différentes configurations et après de nombreux essais nous avons divisé le conduit vocal en un nombre de zones égales à celles obtenues dans l'étude du moulage (SANCHEZ H. & BOE L.J, 1984). L'application des résultats nous permet de connaître l'aire par section, comme le montre la figure C.1. (cf.fig.B.1).

(a)

(b)

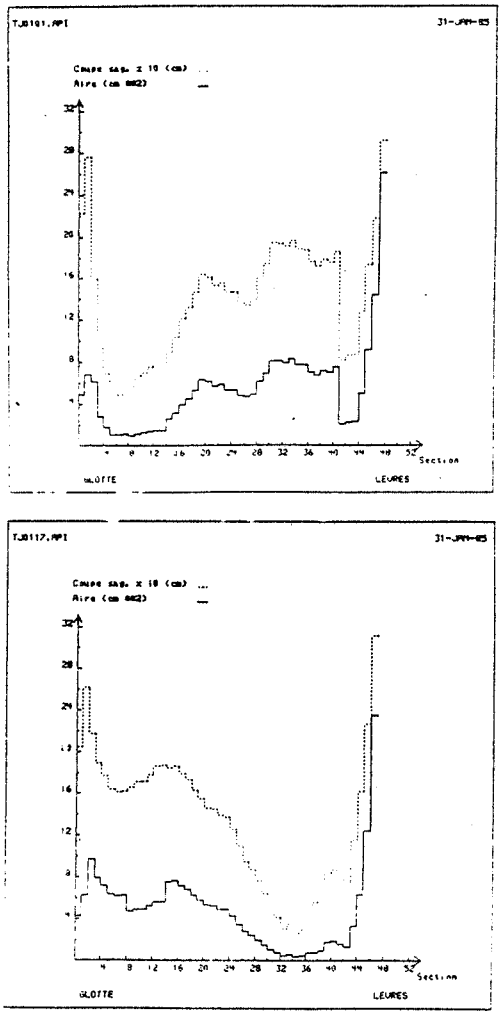


Figure C.1. Deux images de |API|.
a) Section transversale mesurée;
b) aire par section calculée.

Pour la simulation nous avons été obligé de réduire à 20 les 48 à 50 sections initialement mesurées (fig. C.2). Ceci a été réalisé grâce à un programme de lissage de la fonction d'aire.

---

* Passage de la fonction sagittale à la fonction d'aire à partir de coefficients estimés par ailleurs (SANCHEZ H. & BOE L.J., (1984).

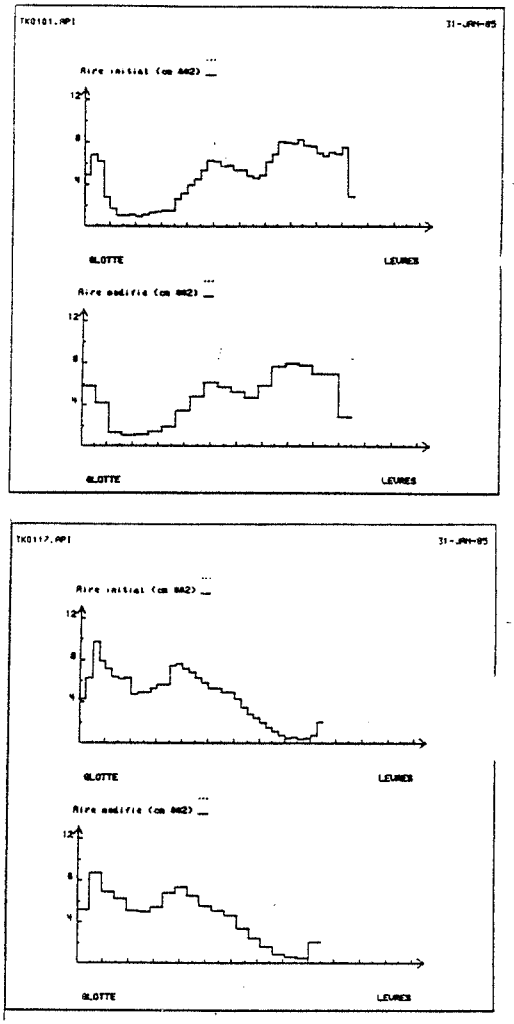


Figure C.2. Lissage de la fonction d'aire :
a) fonction d'aire calculée;
b) fonction d'aire simplifiée.

A partir des fonctions d'aire ainsi calculées nous avons procédé au calcul des formants, bandes passantes, amplitudes pour toutes les configurations dont nous disposons.

Nous avons considéré dans cette étude la simulation de toutes les pertes du conduit vocal.

D. LES RESULTATS.

Le traitement des données radiologiques, nous a permis d'établir la trajectoire des formants pour des transitions maximales. On a pu obtenir, ainsi pour chaque configuration articulatoire un suivi formantique. Aux figures D.1(a,b) sont présentées les valeurs des formants (a) et bandes passantes (b).

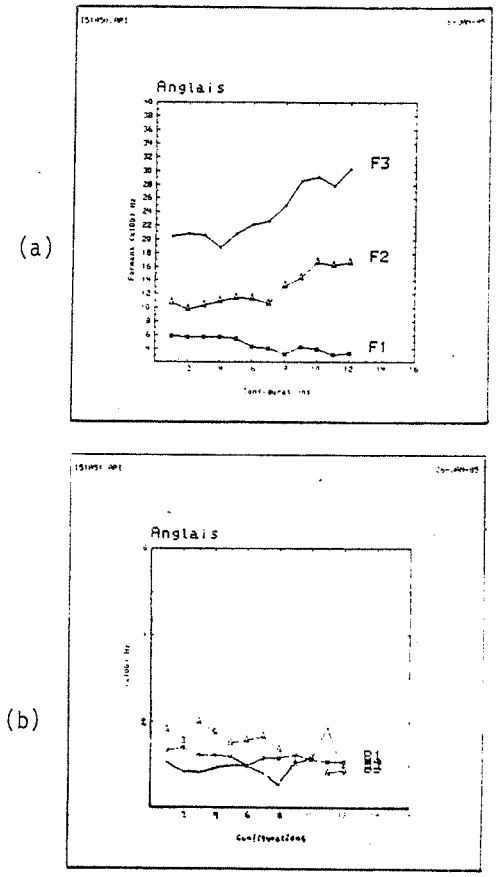


Figure D.1. Valeurs obtenues pour |API| :
a) formants;
b) bandes passantes.

La valeur de chaque point dans la figure D.1 est le résultat d'une configuration articulatoire, c'est-à-dire, la forme du conduit vocal à l'instant t_1, puis à l'instant t_2, etc... On peut vérifier facilement celle-ci. la fonction d'aire n'est pas la même d'un instant à l'autre, ce changement continu du filtre acoustique (conduit vocal) va modifier le signal glottique.

Dans la réalité le problème est encore plus complexe, puisqu'au changement de forme du conduit vocal s'ajoute celui de l'onde de débit, à laquelle peut se superposer une autre source d'excitation générée en un point de constriction ou à la suite d'un occlusion.

L'évolution des deux premiers paramètres a été présentée dans le plan F1/F2 et référencée par rapport aux mesures de PETERSON & BARNEY (1952). (fig. D.2).

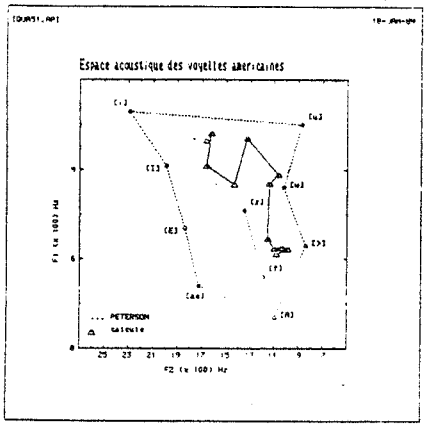


Figure D.2. Représentation spatiale de F1 et F2. Cas de |API|.

Nous pouvons constater que les cibles initiales et finales correspondent aux valeurs de référence. Dans d'autres cas on note que, la cible de la voyelle finale n'est pas atteints, cette différence peut-être due au fait que les tracés au voisinage de la glotte, ne correspondent pas véritablement à la réalité (ils sont très difficiles à interpréter).

On a pu aussi constater que différentes configurations peuvent produire le même résultat (problème de compensation).

Finalement, les transitions formantiques ont été définies comme le segment où l'on détecte un changement dans les valeurs des formants.

E. CONCLUSIONS.

Les résultats présentés sont relativement encourageants; à partir de données radiologiques nous avons pu obtenir par modélisation les trajectoires formantiques pour des transitions |VCV| maximales (avec les bandes passantes et les fonctions de sensibilité associées). Pour ce faire, le passage coupe sagittale fonction d'aire a été amélioré d'un façon notable. Les résultats permettent de déboucher sur de la synthèse par modéle articulatoire qui pourra se commander avec un nombre restreint de paramètres (cf. par exemple le modèle de S. MAEDA, 1978).

Actuellement nous réalisons la synthèse des ces transitions en utilisant un modèle temporel. Nous pourrons ainsi tester l'ensemble de notre demarche sur le plan perceptif.

F. BIBLIOGRAPHIE.

BLUMSTEIN S.E. & STEVENS K.N. (1979)
Acoustic Invariance in Speech Production : Evidence from Mesurements of the Spectral Characteristics of Stop Consonants.
J. Acoust. Soc. Am. 66, 1001-1017.

BLUMSTEIN S.E. & STEVENS K.N. (1980)
Perceptual Invariance and Stop Onset Spectra for Stop Consonants in Different Vowel Environments.
J. Acoust. Soc. Am. 67, 648-662.

COOPER F.S. DELATTRE P.C. LIBERMAN A.M. BORST J.M. & GERSTMAN L.J. (1952)
Some Experiments on the Perception of Synthetic Speech Sounds.
J. Acoust. Soc. Am. 24, 597-606.

HARRIS K.S. HOFFMAN H.S. LIBERMAN A.M. DELATTRE P.C.& COOPER F.S. (1958)
Effect of Third Formant Transitions on the Perception of the Voiced Stop Consonants.
J. Acoust. Soc. Am. 30, 122-126.

KELLY J.L. & LOCHBAUM C. (1962)
Speech Synthesis.
Stockholm Speech Communication Seminar R.I.T.

LEHISTE I. & PETERSON G.E. (1961)
Transitions, Glides and Diphthongs.
J. Acoust. Soc. Am. 33, 268-277.

LIBERMAN A.M. COOPER F.S. SHANKWEILER D.P. & STUDDERT-KENNEDY M. (1967)
Perception of the Speech Code.
Psychol. Rev. 74, 431-461.

LIBERMAN A.M. DELATTRE P. COOPER F.S. & GERTSMAN L.J. (1954)
The role of Consonant-Vowel Transitions in the Perception of the Stop Nasal Consonants.
Psychol. Monogr. 68, 1-13.

LIBERMAN A.M. & STUDDERT-KENNEDY M. (1978)
Phonetic Perception in Handbook of Sensory Physiology : Perception.
Springer-Verlag,New-York, 143-179.

MAEDA S. (1978)
A Statistical Approach in the Construction of an Articulatory Model.
In: Recherches/Acoustique 5, 77-93.

MAEDA S. (1978)
Un Modèle Articulatoire basé sur une Etude Acoustique.
9-ièmes J.E.P. du G.A.L.F., 35-55.

PAUL A.P. HOUSE A.S. & STEVENS K.N. (1964)
Automatic Reduction of Vowel Spectra : An Analysis by synthesis Method and its Evaluation.
J. Acoust. Soc. Am. 36, 303-308.

SANCHEZ H. & BOE L.J. (1984)
De la Coupe Sagittale à la Fonction d'Aire du Conduit Vocal.
Bulletin de L'Inst. de Phonétique, 13, 1-24.

STEVENS K.N. & BLUMSTEIN S.E. (1978)
Invariant Cues for Place of Articulation in Stop Consonants.
J. Acoust. Soc. Am. 64, 1358-1368.

STEVENS K.N. & HOUSE A.S. (1955)
Development of a Quantitative Description of Vowel Articulation.
J. Acoust. Soc. Am. 27, 484-493.

STEVENS K.N. & HOUSE A.S. (1956)
Studies of Formant Transitions using a Vocal Tract Analog.
J. Acoust. Soc. Am. 28, 578-585.

LE RAPPORT ENTRE SYLLABE ET COARTICULATION EN ITALIEN

Massimo Pettorino et Antonella Giannini

Istituto Universitario Orientale, Laboratorio di Fonetica Sperimentale
Napoli, Italia.

In a previous article we examined the extension of the coarticulatory phenomenon in VV and VCV sequences of Italian. We found that the syllabic boundaries don't coincide with those of the traditional segments of consonants and vowels. The aim of this experimental research is to verify those data by the exam of a wider spectrographic material in order to identify the programming units through the observation of the extension of the coarticulatory effect of the lip rounding. Moreover we have examined whether a relation exists between the syllabic boundary and the duration of the consonant.

A model of the syllabic organization of Italian is proposed.

INTRODUCTION

L'orientation des modernes études de phonétique expérimentale tend à considérer le phénomène de la coarticulation comme strictement lié au problème de l'individuation des limites syllabiques. D'un côté la coarticulation, autrefois vue comme l'effet des limitations d'ordre méchanique de l'appareil phonatoire, est à présent considérée comme le reflet de l'organisation en unités de programmation dans la production du parlé; de l'autre côté la syllabe, autrefois vue comme le résultat d'un "chest pulse" (Stetson 1951) [1], est à présent considérée comme syllabe articulatoire, qui est formée d'un groupe d'éléments coproduits au niveau de commande motrice.

Par la suite de ces considérations, étant donné que soit la coarticulation soit la syllabe sont le résultat d'une unique commande motrice, les limites d'extension de l'une ne doivent dépasser les limites d'extension de l'autre.

Il y a eu plusieurs tentatives d'individuer l'unité fondamentale du parlé, grâce à l'observation des effets de coarticulation. Kozhevnikov et Chistovich 1965 [2] suggèrent que le type fondamental de syllabe est formé d'une voyelle et de toutes les consonnes qui la précèdent, tandis que d'autres études expérimentales sont en conflit avec cette hypothèse. Öhman 1966 [3], par example, dans une analyse spectrographique des séquences V_1CV_2 (voyelle-consonne-voyelle), trouve que la première voyelle subît l'effet de coarticulation de la deuxième voyelle à travers la consonne intervocalique.

Dans un précédent article (Pettorino et Giannini 1984) [4] nous avons examiné le rapport existant entre coarticulation et syllabe en italien, mais la recherche était limitée à l'analyse des consonnes brèves et longues en position intervocalique. Les résultats de cette étude indiquaient que :

1) V_1 est influencée par V_2 seulement dans la séquence V_1V_2;
2) V_2 est inflencée par V_1 dans les séquences V_1V_2 et V_1CV_2;
3) il n'y a aucun effet de coarticulation entre les deux voyelles dans la séquence V_1CCV_2 (où CC indique la consonne longue).

Ces données ont révélé que la séquence V_1V_2 est constituée par une seule unité de programmation, V_1CV_2 est constituée par deux unités dont la coupure a lieu pendant la métastase de V_1 et, enfin, V_1CCV_2 est constituée par deux unités dont la coupure a lieu au milieu de la tenue de la consonne longue. Toutes ces données indiquent que, dans les séquences examinées, les limites des unités de programmation ne coïncident jamais avec celles des traditionnelles unités de consonnes et voyelles.

Le but du présent article est d'analyser le type d'organisation syllabique de l'italien par l'examen d'une plus large variété de séquences de consonnes et de voyelles. En outre nous nous proposons de vérifier s'il existe une rélation entre la coupure syllabique et la durée de la consonne.

PROCEDURE

Pour vérifier l'extension de l'effet de coarticulation et, par conséquent, les limites des unités syllabiques, nous avons choisi, entre les différents traits articulatoires, celui qui résulte plus évident sur le traçé du sonagramme, c'est à dire celui de la protrusion des lèvres. En fait, sur le sonagramme, une articulation arrondie présente des formants à des valeurs plus basses que la même articulation non labialisée.

Dans nos séquences V_2 est [u] ou [a], toujours en position accentuée, à fin d'avoir respectivement la présence ou l'absence de la protrusion des lèvres et nous avons déterminé à quel instant, avant la voyelle [u] la labialisation commence.

Pour ce qui est des consonnes intervocaliques, nous avons pris en examen les dentales parce qu'elles ont un articulateur qui n'influence pas l'arrondissement des lèvres. Seulement dans le cas de l'occlusive, au lieu de [t] et [d], on a pris en considérations les vélaires [k] et [g] car elles ont un locus du deuxième formant à 3000 c/s quand elles sont suivies d'une voyelle non-arrondie et à une valeur beaucoups plus basse quand elles sont suivies d'une voyelle arrondie. Cela permet de visualiser d'une manière assez claire sur le sonagramme si l'occlusive est labialisée ou non labialisée.

Les séquences prises en examen sont V_1CV_2, $V_1C_1C_2V_2$, $V_1C_1C_2C_3V_2$. Dans tous les contextes V_1 est [a] et, comme nous avons dit, V_2 est [a] ou [u]. Dans la séquence V_1CV_2 , C est une quelconque consonne du système phonologique de l'italien. Dans la séquence $V_1C_1C_2V_2$ les consonnes sont [l] [r] [n] [s] [k] ou [g] combinées dans tous les possibles contextes phonologiques de l'italien. Dans la séquence $V_1C_1C_2C_3V_2$, C_1 est [l] [r] [n] ou [s], C_2 est [k] ou [g] et C_3 est [l] ou [r] .

De chaque séquence nous avons effectué un sonagramme à bande large et nous avons mesuré les durées de tous les segments consonantiques et vocaliques. Le matériel a été prononcé par deux locuteurs de langue maternelle italienne.

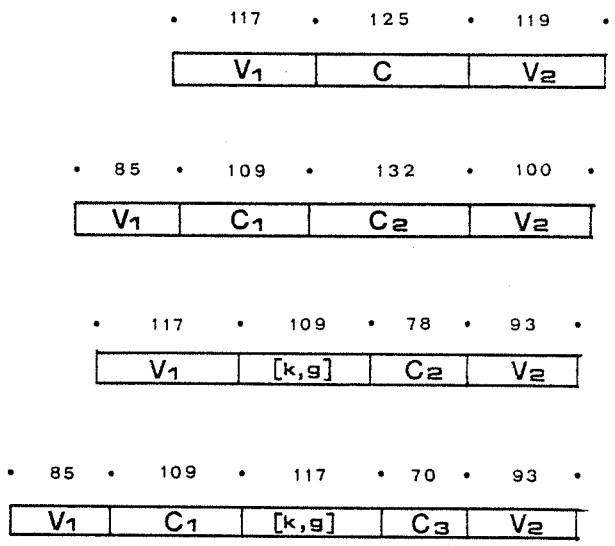
RESULTATS

Pour ce qui concerne les valeurs des formants des voyelles, les traçés ne révèlent aucune variation considérable dans toutes les séquences examinées.Le Tableau I montre les valeurs moyennes des durées des consonnes et voyelles dans les différentes situations contextuelles.

L'analyse spectrographique indique que dans la séquence V_1CV_2 , où V_2 est [u], la consonne est toujours arrondie et la protrusion des lèvres a lieu à l'intérieur de la première voyelle. Ce fait est évident parce que pendant la métastase de V_1 le deuxième formant subît une déviation négative en comparaison de la même séquence se terminant par [a]. Nous avons calculé que la labialisation commence environ 150 ms avant le début de [u] .

Pour ce qui est de la séquence $V_1C_1C_2V_2$, nous devons faire une distinction entre le cas où C_1 est une des dentales prises en considération et le cas où C_1 est une occlusive vélaire.

Tableau I. Valeurs moyennes en millisecondes des durées des voyelles et des consonnes.

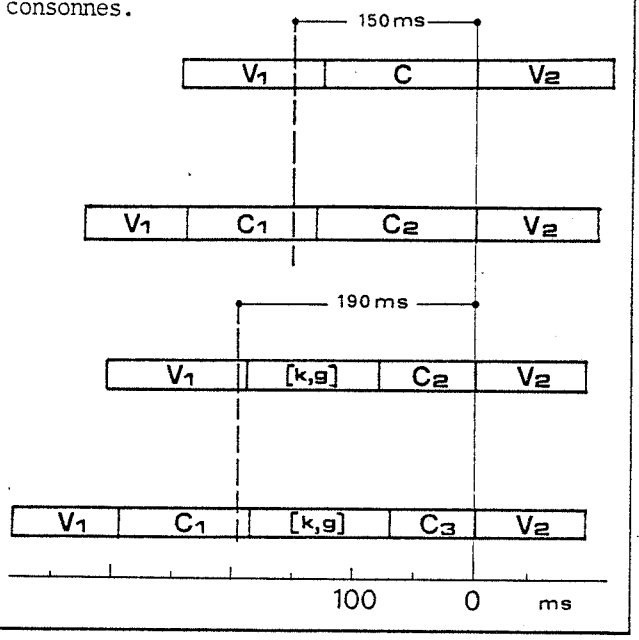


Dans le premier cas seulement C_2 est arrondie et la protrusion des lèvres commence pendant la métastase de C_1 ; en termes de temps, cela a lieu environ 150 ms avant le début de la voyelle [u]. Dans l'autre cas, soit C_1 soit C_2 sont arrondies et la protrusion commence pendant la métastase de V_1 , à environ 190 ms avant le début de la voyelle [u] .

Pour ce qui concerne la séquence $V_1C_1C_2C_3V_2$ où, comme on a déjà dit, C_2 est [k] ou [g], C_2 et C_3 sont arrondies et l'arrondissement des lèvres commence pendant la métastase de C_1, environ 190 ms avant le début de la voyelle [u].

Toutes ces données sont résumées dans le Tableau II.

Tableau II. Position de la coupure syllabique dans les différentes séquences de voyelles et consonnes.



DISCUSSION

Le but de cette recherche expérimentale est, comme nous avons dit, celui d'examiner l'organisation syllabique de l'italien et, pour faire ça, nous avons pris en considération un particulier trait articulatoire, la protrusion des lèvres. Les données que nous avons obtenues nous indiquent que :

1) toutes les séquences examinées sont produites par deux distinctes commandes motrices de programmation et par conséquent il y a dans tous les cas deux syllabes;
2) la consonne qui précède immédiatement la V_2 fait toujours partie de la deuxième unité syllabique;
3) dans la séquence $V_1C_1C_2V_2$ la coupure syllabique a lieu à la fin de la première consonne, sauf le cas où C_1 est une occlusive. En ce cas-là la coupure syllabique tombe à la fin de la première voyelle. Dans la séquence $V_1C_1C_2V_2$ nous avons donc deux types possibles d'organisation syllabique, V_1C_1 - C_2V_2 et V_1-$C_1C_2V_2$;
4) dans la séquence $V_1C_1C_2C_3V_2$, où C_2 est toujours une occlusive, la coupure syllabique a lieu à la fin de la première consonne, en donnant lieu à deux syllabes du type V_1C_1-$C_2C_3V_2$.

Une remarque à faire à ce point est que la syllabe articulatoire que nous avons examinée, qui se base sur le concept d'ensemble d'éléments coproduits au niveau de commande motrice, coïncide avec la syllabe traditionnelle de l'italien, sauf le cas de [s] suivi de consonne. En fait, en ce cas, la tradition groupe [s] avec la ou les consonnes qui suivent (par exemple "ciascuno" est divisé [tʃa] - [sku] - [no]) tandis que, comme nous avons vu, il y a toujours une coupure syllabique entre [s] et la consonne qui suit ([tʃas] -[ku]-[no]).

Nos données ne concordent pas avec celles obtenues par Daniloff-Moll 1968 [5] et Benguerel-Cowan 1974 [6] qui, suivant des méthodes de recherche différente de l'analyse spectrographique, ont trouvé que l'arrondissement, dans une séquence $V_1C_1..C_nV_2$ débute pendant la première consonne d'un groupe de quatre (Daniloff-Moll) jusqu'à six (Benguerel-Cowan) consonnes. Pourtant, il faut souligner que ces articles se réfèrent l'un à l'anglais américain et l'autre au français. Il serait intéressant d'étudier l'organisation syllabique de ces deux langues et celle de l'italien en suivant la même méthode de recherche.

Un autre article très intéressant sur la coarticulation de la protrusion des lèvres est celui de Bell Berti-Harris 1979 [7], qui se base sur l'analyse électro-myographique de séquences VCV et VCCV de l'anglais américain. La thèse de cet article est que "the onset of lip rounding is not synchronized to activity related to phonetic segments in the utterance other than the vowel itself" (p. 1269), c'est à dire que la commande motrice pour la protrusion des lèvres pour [u] commence un temps déterminé avant le début de la voyelle, indépendemment du nombre de consonnes intervocaliques.

Nos données, au contraire, semblent indiquer que, lorsque la labialisation intéresse seulement une consonne, le temps d'anticipation de ce phénomène est environ de 150 ms, lorsque la labialisation intéresse deux consonnes le temps d'anticipation est environ de 190 ms. Il y aurait donc deux classes temporelles pour la labialisation, selon que les deux consonnes appartiennent à une ou à deux syllabes. Ces données suggèrent des ultérieures considérations.

Comme il résulte du Tableau I, dans la séquence $V_1C_1C_2V_2$ la durée des consonnes, indépendemment de la labialisation, est d'environ 240 ms lorsque les deux consonnes sont séparées par une coupure syllabique, tandis qu'elle est réduite à environ 187 ms lorsque les deux consonnes font partie de la même syllabe. Cela est dû à la compression temporelle des éléments consonantiques à l'intérieur d'une syllabe: plus sont les consonnes dans une syllabe, plus leur durées sont réduites. Cela veut dire aussi qu'une déterminée combinaison de consonnes doit avoir une durée différente selon qu'elle présente ou non à son intérieur une coupure syllabique.

Etant donné que dans les séquences examinées dans cette étude on ne peut pas varier la position de la coupure syllabique sans altérer au moins l'ordre de C_1 et C_2 (par example /arguti/ et /agrumi/), on a pris en considération la séquence [sk] soit en position initiale soit en position intervocalique (par example /scuro/ et /ciascuno/).

Les sonagrammes indiquent que [sk] en position initiale, est entièrement labialisé et donc appartient à une même syllabe. Pour ce qui est de la durée, [sk] en position initiale a une durée d'environ 160 ms , tandis que en position intervocalique a une durée d'environ 220 ms.

Cette différence de durée confirme le modèle d'organisation syllabique que nous avons proposé sur la base de l'examen du phénomène de coarticulation par rapport à la protrusion des lèvres.

Les résultats de cette étude, obtenues par l'analyse spectrographique, suggèrent l'opportunité de vérifier s'il existe un rapport entre coupure syllabique et dégré d'aperture de la consonne puisque, comme nous avons vu, le déplacement d'une occlusive de C_1 à C_2 détermine un déplacement de la coupure syllabique.

En outre, il serait intéressant de vérifier par la synthèse la validité, sur le plan de la perception, de ce modèle d'organisation syllabique de l'italien.

REFERENCES

1 R.H. Stetson, "Motor Phonetics: a Study of Speech Movements in Action",Amsterdam, North Holland, 1951.

2 R.V. Kozhevnikov et L. Chistovich, "Speech: Articulation and Perception" (translated from Russian), Joint Publications Research Service, Rep.30, 543, Washington D.C., 1965.

3 S.E.G. Öhman, "Coarticulation in VCV Utterances: Spectrographic Measurements", Journal of Acoustical Society of America, Vol.39, pp.151-168, 1966.

4 M. Pettorino et A. Giannini, "Some Aspects of Coarticulation in Italian: a Spectrographic Analysis of VV, VCV and VCCV Utterances", Wiener Linguistische Gazette, Suppl.3, Wien, 1984.

5 R. Daniloff et K. Moll, "Coarticulation of Lip Rounding", Journal of Speech and Hearing Research, Vol.11, n°4, pp.707-721, December, 1968.

6 A.P. Benguerel et H.A. Cowan, "Coarticulation of Upper Lip Protrusion in French", Phonetica, Vol. 30, n° 1, pp.41-55, 1974.

7 F. Bell-Berti et K.S. Harris, "Anticipatory Coarticulation: Some Implications from a Study of Lip Rounding", Journal of Acoustical Society of America, Vol.65, n° 5, pp.1268-1270, May, 1979.

EFFETS TRANSSYLLABIQUES DE COARTICULATION VOYELLE-VOYELLE EN ITALIEN PARLE'

Mario Vayra

Scuola Normale Superiore
Piazza dei Cavalieri 7, 56100 Pisa

The experiment shows the occurrence of coarticulatory influences among stressed and unstressed vowels in spoken Italian. The outcomes seem to suggest - for this subject - a directional asymmetry in the effects of coarticulation among non contiguous vowels, with stressed to unstressed carryover effects tending to exceed the anticipatory ones. Interestingly, these asymmetries seem coherent with those exhibited by English [6], [7], a language traditionally opposed to Italian, by linguists, as far as the phenomenon of reduced vowels is concerned.

Le problème. Cette expérience a été conçue dans le but d'obtenir, pour la langue italienne, un test préliminaire de cette hypothèse générale - exprimée plusieurs fois par Fowler ([6] , [7] p. ex.):

1. La production des voyelles est "continue"; et, plus spécifiquement,
2. Les voyelles inaccentuées sont produites ainsi qu'un "détour" de la trajectoire de la langue, "en marche" vers la voyelle accentuée suivante (cfr. 8 pour une similaire et précoce interprétation dynamique des "targets" vocaliques).

En termes traditionnels, cette expérience vise à découvrir l'existence d'éventuels indices d'une influence coarticulatoire portée par la voyelle accentuée sur la voyelle inaccentuée (précédente ou suivante). Dont l'hypothèse que, si les voyelles sont articulées d'une façon "continue", la présence d'une consonne intervocalique n'arrive pas à neutraliser les effets voyelle-voyelle d'irradiation timbrique; cela, surtout, lorsque la langue n'est pas l'articulateur primaire de la consonne (c'est le cas, ici, de /b/).

Et encore, l'individuation dans l'italien - langue classifiée dans la tradition des études comme "syllable timed" ([1] , [2], p. ex.) - d'interactions articulatoires entre segments non contigus, pourra nous offrir quelques informations ultérieures (ou quelques doutes) sur l'organisation temporelle de sa production. Ainsi, si l'on détermine que les frontières syllabiques ne dérangent pas nécessairement les interactions acoustiques ou articulatoires entre voyelles accentuées on aura alors un indice assez fort de l'existence - même en italien - d'une périodicité articulatoire fonctionnellement significative, dans le domaine inter-accentuel. En effet, selon l'interprétation des faits coarticulatoires en termes de réelle "coproduction" premièrement suggérée par Öhman , [9] , les "compensations temporelles" entre segments de la chaîne phonique pourraient indiquer que les consonnes et les voyelles sont produites "de concert", et donc les "marges" des voyelles sont progressivement masqués par les consonnes. Mais, si pour ce qui concerne l'anglais - langue traditionellement classifiée comme "stress timed" (voir [3] p. ex.) paraît raisonnable l'hypothèse que consonnes et groupes de consonnes sont produits sur un fond d'articulation continue de la voyelle accentuée [10] , [7] , qu'est-ce qu'il se passera pour l'italien? Car, dans une langue "stress timed" c'est l'intervalle interaccentuel qui pourrait constituer le domaine phonétique des "compensations" temporelles que la voyelle accentuée révèle en fonction des voyelles inac-

centuées précédentes ou suivantes [6], [7]; tandis que dans le cas de la langue italienne c'est la frontière syllabique qui, au contraire, devrait jouer un rôle primaire au niveau articulatoire et phonétique: en conditionnant ainsi les stratégies articulatoires et les procès phonétiques, ou en émergeant des cycles articulatoires à dimension syllabique.

Admettre la presence d'interactions articulatoires transsyllabiques dans une langue telle que l'italien nous obligerait a re-conceptualiser la notion traditionelle de "timing" comme contrôle temporel sur "unités de production" à dimension syllabique ou inter-accentuelle (voir [1] pour une position dubitative similairesur la notion de "contrôle" des durées au niveau soit syllabique que inter-accentuel).

Mais encore: s'il n'y a pas de doute que la coarticulation intervocalique n'est pas du tout équivalente à la "réduction" vocalique, il est aussi vrai que, pour l'anglais, on a fait dériver la réduction de la coarticulation, et non pas le contraire [6]. Et nous savons très bien que la réduction a été consacrée comme une des "marques" phonétiques primaires dans le langues "stress timed" (dans cette tradition et sur l'italien voir aussi [12] [13]).

Deux types différents d'issues expérimentales semblent suggérer un ré-examen des positions traditionelles au sujet de la réduction dans le vocalisme inaccentué de la langue italienne. Il s'agit d'une part des données recuellies par [11] sur la présence de faibles et inconsistentes "compressions temporelles" au niveau intra-syllabique ; de l'autre des résultats obtenus par [4] sur l'abaissement de F1 dans la voyelle /a/ lorsqu'elle est inaccentuée. On peut observer, à ce propos, que la voyelle /a/, dont on étudiait les variations formantiques, était insérée - dans l'étude de [4] - dans des séquences phonétiques trisyllabiques où, tandis que la structure rythmique changeait systematiquement, la couleur timbrique restait constante (/'papapa:pa'papa:papa'pa/, p.ex.).

Selon les Auteurs "only F1 exhibits systematic variations related to stress", while F2 and F3 vary according to the consonantal context..." [4]:126. Il sera donc très interessant de vérifier si des changements sont possibles même dans F2 de la voyelle inaccentuée - ou dans son degré d'antérieurité/postérieurité - lorsque nous en modifions le contexte vocalique accentué.

Ces résultats ont motivé l'expérience qu'on va décrire:la prédiction est que la coarticulation intervocalique, où elle agit, produira des glissements timbriques sur les voyelles inaccentuées inclues dans l'interval accent-accent: c'est à dire, elle produira des variations fréquentielles soit dans F1 que dans F2 des voyelles inaccentuées en fonction du timbre de celles accentuées, précédentes et suivantes.

Mais ce sont des effets coarticulatoires de cette sorte qui on fait parler, pour l'anglais, de cycle articulatoire continu accent-accent [6]. Ainsi, la découverte d'effets coarticulatoires similaires dans l'italien, pourrait indiquer que les distinctions typologiques traditionnelles en termes de "stress/syllable timing", méritent des vérifications ultérieures, sur le plan théorique aussi que expérimental.

L'expérience.

Protocole expérimental. Le matériau qui a servi de base à la présente recherche comprend un corpus de trois listes de trisyllabes de la forme /V1bV2bV3/ (p.ex.: /ibabi/). La voyelle médiane est toujours /a/, tandis que celles initiales et finales sont /i/,/a/,/u/; ces voyelles ont été systématiquement alternées dans la position initiale et finale, jusqu'à obtenir neuf énoncés pour chaque liste. Chaque liste diffère des autres quant à la structure rythmique du trisyllabe (accent initial, médian, final). Un locuteur a lu les trisyllabes qui ont été présentés sur des fiches cartonnées en ordre randomisé. Dans chaque fiche le trisyllabe était précedé par un mot-cible "réel" (/'amano/, p.ex.) dont le locuteur devait répliquer le modèle rythmique en produisant le trisyllabe (/'ibabi/, p.ex.). Il a été enregistré à la chambre sourde du Laboratorio di Linguistica de la Scuola Normale Superiore, au moment où sa production était devenue fluente. Les listes ont été lues trois fois ainsi qu'on a obtenu, au total, 81 "pseudo-mots".

Les mesures. L'analyse des fréquences formantiques a été effectuée, chez le Laboratorio di Linguistica de la Scuola Norma-

le Superiore, par un calculateur PDP 11/34 à l'aide d'un programme qui utilise des méthodes de prédiction linéaire autorégressive (I.L.S.). L'analyse est realisée à partir des coefficients d'autocorrélation. La fenêtre d'analyse est de 20 ms.. 14 coefficients de prédiction sont calculés. Après analyse (FFT), l'amplitude, la fréquence fondamentale et les valeurs des pics spectraux sont présentées sur un écran de visualisation. Un "pseudo-spectrogramme" (de séquences) du signal analysé pouvait aussi être représenté.

Le sujet. Un locuteur italophone de Florence, chercheur universitaire chez la Scuola Normale Superiore, non marqué quant aux traits géo-linguistiques de sa production.

Résultats et discussion.

Les mesures consignées aux Tableaux I et II sont relatives aux voyelles médianes /a/, accentuées (a) et inaccentuées (b), des trisyllabes utilisés dans l'expérience. Ces valeurs, qui ont été relevées dans le point médian de la zone stable de la voyelle, correspondent aux fréquences centrales des bandes relatives à F1 (Tableau I)et F2 (Tableau II).

Le Tableau I présente les valeurs moyennes de F1 pour la voyelle médiane /a/ dans le contexte de /i/, /a/, /u/, initiales (première colonne) et finales (seconde colonne). Chaque moyenne dans la première colonne comprend les trois possibles contextes vocaliques finaux. Ainsi, la valeur de 902.05 Hz pour /i/ initial, dans la première colonne du Tableau I, correspond à la moyenne de F1 pour /a/ médian en /'ibabi, 'ibaba, 'ibabu/. La même procédure dans la seconde colonne (ici les moyennes relatives à /a/ médian comprennent les trois contextes initiaux (p.ex.:/iba'bi, aba'bi, uba'bi/).

Le Tableau II montre les valeurs correspondantes pour F2. Dans ces Tableaux on peut valuer les effets coarticulatoires en comparant F1 et F2 relatifs à /a/ en différents contextes. Ainsi, quant à F1, un effet coarticulatoire de /i/ et /u/ sur /a/ en devrait abaisser le formant. Toutefois, dans le cas de /a/ médian accentué il n'y a, peut-être, de significatif qu'un effet de /u/ initial. En ce qui concerne le même effet sur la médiane inaccentuée on découvre une tendance plus forte à l'abaissement de F1, dans le cas, surtout de /u/ final (601.38 Hz), et de /u/ initial (659.36 Hz).

Les Tableaux I (a,b) paraissent confirmer ce qu'on connaissait déjà de la langue italienne: en effet, quant a F1, ils révèlent des effets d'interaction articulatoire transsyllabique très faibles. Selon les issues de [4] ,c'est l'accent, ici, et non pas la coarticulation., qui semble jouer le rôle primaire dans les variations de F1: Il est interessant d'observer l'analogie de ces données avec celles présentées par [6]. Il s'agit d'un étude similaire et précédente sur l'Anglais, partiellement répliqué ici en accord avec l'Autrice. Même en Anglais "stress of medial vowel had a striking effect, with the stressed medial vowel having a substantially higher first formant than the unstressed vowel"(p.131).

Mais il est, peut être, bien plus intéressant de noter que l'analogie entre les deux langues poursuit quant aux effets coarticulatoires sur F2. A cet égard, un effet coarticulatoire de la voyelle antérieure /i/ sur /a/ médian (ou /ə/) en élèverait la valeur du relatif F2: tandis qu'une influence par la voyelle postérieure /u/ en provoquerait l'abaissement. Les Tableaux II (a,b) indiquent que dans l'italien comme dans l'Anglais les effets progressifs sont majeurs que ceux anticipatifs: et encore, que les influences sur la voyelle accentuée par la voyelle inaccentuée sont moins fortes que celles contraires [6] . L'effet de l'accent sur F2, qui dans l'Anglais élève le formant, apparaît inexistant ou inconsistent en italien. La Figure 2 représente séparément les effets progressifs et anticipatifs sur les voyelles /a/ moyennes, accentuées et inaccentuées. Ainsi la voyelle médiane /a/ inaccentuée présente dans les trisyllabes qui ont un /i/ initial un F2 moyen de 1586.35 Hz; lorsque la voyelle initiale est /a/ ou /u/ la moyenne est respectivement de 1379.35 Hz et de 1317.23 Hz.

Dans la fig. 1 on a nommé "variations coarticulatoires" les differences entre ces valeurs moyennes. Elles arrivent, dans le cas de la coarticulation progressive sur la voyelle inaccentuée jusqu'aux 238.06 Hz.

Les asymétries directionnelles entre les effets coarticulatoires anticipatifs et progressifs (fig.1) paraissent cohérentes avec celles qu'on a montré dans l'Anglais [6] .

En conclusion, les présentes données révèlent, pour l'italien, des phénomènes significatifs de glissement timbrique dans la voyelle inaccentuée en fonction du contexte transsyllabique. La question est que des phénomènes semblables ne semblent pas trouver une place aisée à l'interieur des traditionnelles typologies linguistiques en termes de "syllable/stress timing".

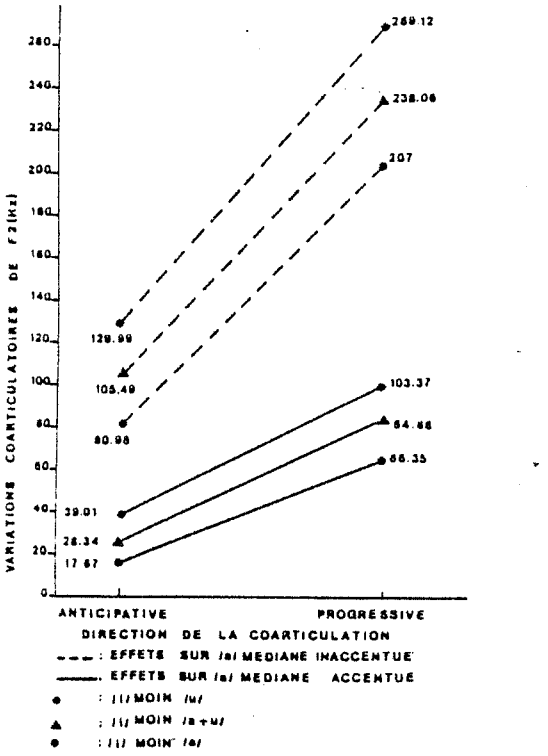


FIGURE 1. Effets coarticulatoires anticipatifs et progressifs sur /a/ médian accentué (ligne continue)et inaccentué (ligne hachée). La degré des effets a été calculé en sustrayant la valeur moyenne de F2 pour /a/ dans le contexte des voyelles bordantes /a/,/u/ (et moyenne de /a+u/) de celle pour /a/ en contexte de /i/.

/a/ voyelle médiane

(a)	accentuée		(b)	inaccentuée	
	INITIALE	FINALE		INITIALE	FINALE
/i/	920.05	909.06		689.03	646.34
/a/	915.38	909.36		729.02	671.36
/u/	869.69	892.70		659.36	601.38

TABLEAU.I. Valeurs moyennes de F1(Hz) pour les /a/ centraux accentués (a) et inaccentués (b) dans le contexte de /i/,/a/,/u/ initiaux (première colonne) et finaux (deuxième colonne). Voir le text pour une explication.

/a/ voyelle médiane

(a)	accentuée		(b)	inaccentuée	
	INITIALE	FINALE		INITIALE	FINALE
/i/	1372.04	1334.36		1586.35	1383.01
/a/	1305.69	1316.69		1379.35	1302.03
/u/	1268.67	1295.35		1317.23	1253.02

TABLEAU II. Valeurs moyennes de F2(Hz) pour les /a/ centraux accentués (a) et inaccentués (b), dans le contexte de /i/,/a/,/u/ initiaux (première colonne) et finaux (deuxième colonne). Voir le text pour une explication.

REFERENCES

[1] P.M. Bertinetto, Strutture prosodiche dell'italiano, Firenze, 1981.

[2] P.M. Bertinetto, "Ancora sull'italiano come lingua ad isocronia sillabica", en Scritti linguistici in onore di G.B. Pellegrini, Pisa,pp 1073-1082,1983.

[3] R.M. Dauer, "Stress-timing and syllable timing reanalyzed", J. of Phon.,11, pp 51-62,1983.

[4] E. Farnetani et S. Kori, "Lexical stress in spoken sentences:A study on duration and vowel formant pattern", Quaderni Centro Studi Ricerche Fonetica, Padova, pp 106-133, 1982.

[5] F. Ferrero,E. Magno Caldognetto,K. Vagges et C Lavagnoli, "Some acoustic and perceptual characteristics of the italian vowels", J. of It. Ling. 3, 1, pp 87-96, 1978.

[6] C.A. Fowler,"Production and perception of coarticulation among stressed and unstressed vowels", J. of Speech and Hear. Res., 46,pp 127-139, 1981.

[7] C.C. Fowler, "Converging sources of evidence on spoken and perceived rhythms of speech: cyclic production of vowels in sequences of monosyllabic stress feet", J. of Exp. Psych.:Gen.,112,pp 386-412, 1983.

[8] J. Martin, "Rhythmic (hierarchical) vs. serial structure in speech and other behavior", Psych. Rev., 79, 6, pp 487-509, 1972.

[9] S. Öhman, "Coarticulation in VCV utterances", J.A.S.A., 39, pp 151-168, 1966.

[10] B. Tuller,J.A.S. Kelso, K. Harris, "Further evidence for the role of relative timing in speech: A reply to Barry", Haskins Lab.SR-74/75, pp 187-195, 1983.

[11] M. Vayra, C. Avesani, C. Fowler, "Patterns of temporal compression in spoken italian",Proc. 10th Int. Cong. Phon. Sc., Dordrecht, pp 541-546, 1983.

[12] I. Vogel, La sillaba come unità fonologica, Bologna, 1982.

[13] G. Marotta, Aspetti della struttura ritmico-temporale in italiano, Pisa, 1984.

UNE SOURCE D'EXCITATION COHERENTE DANS LES OCCLUSIVES

Shinji Maeda

CNET, Lannion 22301

RESUME

L'écoulement de l'air durant le relâchement des occlusives a été calculé avec l'aide d'un modèle simple du conduit vocal. Il est apparu que le conduit vocal peut être excité par une source cohérente et par une source de bruit. La source cohérente est créée par deux mécanismes aérodynamiques différents; une modulation de l'écoulement de l'air dûe à un excés de la pression d'air dans les cavités situées derrière la constriction supraglottique et par un écoulement d'air dirigé vers l'intérieur résultant de l'expansion rapide de l'aire de la constriction. Le niveau sonore des signaux dus à une source cohérente de cette sorte est relativement élevé, et il peut être porteur d'information phonétique dans certains cas.

INTRODUCTION

Tout son de parole est généré soit à partir de la modulation de l'écoulement de l'air à travers une constriction dont l'aire varie rapidement (une source cohérente), soit à partir de l'instabilité de l'écoulement lui même (une source de bruit turbulente). L'intensité d'un tel bruit peut être calculée en fonction du débit de l'écoulement et par la taille de la constriction [1], ou par le nombre de Reynolds [2]. Les deux sources de sons peuvent exciter simultanement le conduit vocal. L'écoulement d'air est, en conséquence, un ingrédient essentiel dans la génération du son, et en particulier, pour les occlusives et les fricatives. Stevens [3] a proposé un cadre théorique pour calculer l'écoulement dans les cas stationnaires. Dans le présent article, nous étendons les études de Stevens aux cas dynamiques pour examiner comment l'écoulement de l'air peut être modulé lorsque la constriction supraglottique s'agrandit rapidement. Nous espèrons que la connaissance acquise par cette étude aidera à mieux comprendre les caractéristiques principales des signaux lors du relâchement des occlusives dans la parole naturelle, et de synthétiser de facon précise ces signaux avec l'aide, par exemple, d'un synthétiseur du conduit vocal [4].

UN MODELE DU CONDUIT VOCAL POUR LE CALCUL DE L'ECOULEMENT

Durant la production des occlusives, l'écoulement de l'air est déterminé principalement par la pression d'air subglottique P_s et la configuration des constrictions glottiques et supraglottiques. Dans cette analyse, P_s et la glotte sont fixes et seule la constriction supraglottique varie avec le temps. Dans ce cas, le conduit vocal peut être representé par le reseau sur la Figure 1.

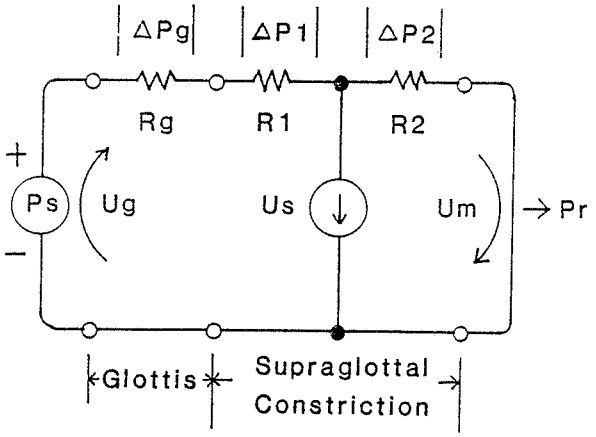


Fig 1 Un modèle du conduit vocal pour calculer l'écoulement de l'air, U_m, pendant le relâchement des occlusives.

Nous supposons que les deux constrictions peuvent être représentées par un tube uniforme qui est défini par son aire de section, A, et sa longueur, L. La chute de pression à travers chacune des deux résistances cinétiques R_1 et R_2, correspond à la chute de pression de Bernouilli à l'entrée de chaque constriction, qui est définie par

$$P = k(\rho U^2/2A^2),$$

ou "ρ" est la densite de l'air. "U" est le débit volumique de l'écoulement de l'air à l'entrée, et "k" est appelé "un coefficient de perte d'entrée".

Sa valeur varie de 0.5 a 2.0, en fonction de la forme de la constriction et de la magnitude de U. Nous assumons k=1. L'autre résistance cinétique, R_2, à la sortie de la constriction est importante seulement lorsque le débit volumique de sortie, U_m, est négatif, c'est-à-dire lorsque l'air s'écoule dans la constriction venant de l'extérieur.

La source parallèle courante, U_s, rend compte de l'écoulement d'air venant de l'intérieur (de l'air aspiré) lorsque la constriction s'agrandit apres le relâchement. La source du l'écoulement U_s est égale à la dérivée temporelle du volume du tube de constriction. Comme nous avons assumé que sa forme est représentée par un tube uniforme,

$$U_s = d(L_c A_c)/dt. \qquad (1)$$

Pour que le U_s calculé soit significatif, l'expansion de la constriction doit être specifiée de facon réaliste en terme de longueur, L_c et d'aire, A_c.

En tenant compte des données présenteés par Fujimura [5], l'expansion de l'aire, au moins dans le cas des occlusives labiales, semble être décrite de facon raisonnable par la fonction exponentielle,

$$A_c = A_o(1 - e^{-t/T_c}), \qquad (2)$$

ou A_o est l'aire cible en fin d'expansion et T_c est une constante de temps indiquant la vitesse d'expansion. La valeur de T_c, 40 ms, est utilisée comme valeur standard dans cette étude. Comme il n'existe pas de données publiées, disponibles, pour le calcul de la variation de longueur, nous supposons la fonction adhoc suivante pour le calcul de L_c,

$$L_c=(L_o-L_i)(1-e^{-t/T_c})+L_i, \qquad (3)$$

ou L_o est une longueur cible (fixée a 1 cm), et L_i est une longueur initiale correspondant à la longueur de l'occusion juste avant le relâchement. L'inclusion de L_i était nécessaire pour expliquer les signaux observés lors du relâchement des occlusives dans de la parole naturelle, que nous décrirons plus tard.

Il peut être utile de démontrer brievement quelques unes des caractéristiques de base de l'écoulement de sortie, U_m dans notre modèle. Par la résolution du circuit représenté à la Figure 1, U_m peut être dérivé de la facon suivante,

$$U_m=-U_s/2 - cP_s/2U_s \quad \text{for } U_m<0, \qquad (4a)$$

$$=-U_s - \sqrt{cP_s} \quad \text{for } U_m>0, \qquad (4b)$$

where

$$c = 2/\rho\,(1/A_c^2+1/A_g^2). \qquad (5)$$

Il est intéressant de noter que la valeur initiale de "c", c'est-à-dire la valeur au moment même du relâchement, est égal à zero, puisque $A_c=0$ au moment t=0. Cela signifie que la valeur initiale de U_m dépend seulement de celle de U_s. Des équations 1, 2 et 3, on déduit

$$U_s|_{t=0}=L_i(A_o/T_c). \qquad (6)$$

L'amplitude de U_s au moment t=0 est alors proportionelle à la vitesse initiale de l'expansion de l'aire, c'est-à-dire, A_o/T_c et a L_i. Si L_i n'est pas égale à zero, U_s à t=0 est alors différente de zéro et positive. En conséquence, U_m montrera une évolution instantannée dans le sens négatif au début du relachement. Comme la pression d'aire radiée est approximativement proportionelle à la dérivée de U_m, un son d'impulsion négative doit être produit à cet instant.

De plus, à partir des équations 4b et 5, U_m est gouverné par la constriction supraglottique seulement quand A_c est inférieur a A_g. Après que A_c excède A_g, l'écoulement est essentiellement gouverné par A_g et il devient alors constant. De l'équation 4b, l'amplitude d'un tel ecoulement "saturé" est proportionnelle à A_g et à la racine carrée de P_s, pression de l'aire subglottique. Le temps de montée de U_m, en conséquence, dépend non seulement de la vitesse initiale du relâchement, A_o/T_c mais aussi de A_g. La source cohérente, P_r, devrait indiquer, juste après l'occurence d'une impulsion négative, une "queue" qui consiste en une montée rapide suivie d'une descente graduelle.

RESULTATS

Démontrons maintenant quelques uns des résultats des calculs utilisant des valeurs représentatives des paramètres du conduit vocal. Les effets aérodynamiques de ces paramètres, tels que A_g, L_i et T_c, durant le relâchement des occlusives, sont évalués en terme de U_m, un nombre relatif de Reynolds (Re, qui est proportionnel a U_m^2/A_c) , et P_r. Le nombre de Reynolds a été utilisé comme une indication grossière de l'intensité du bruit de turbulence. Dans ce calcul, des résistances et des inductances additionnelles en séries ont été incluses pour prendre en compte les pertes dues à la viscuosité et l'inertie de l'air lorsque la constriction supraglottique forme une tube étroit et relativement long au début du relachement.

Les effets de L_i sont démontrés dans la Figure 2, ou les U_m, Re et P_r calculés, du haut jusqu'en bas, sont indiqués en fonction du temps pour trois valeurs differentes de L_i. Les autres paramètres, P_s, A_g, A_o, et T_c, ont été fixés selon leur valeurs standards indiquées à la Figure 2.

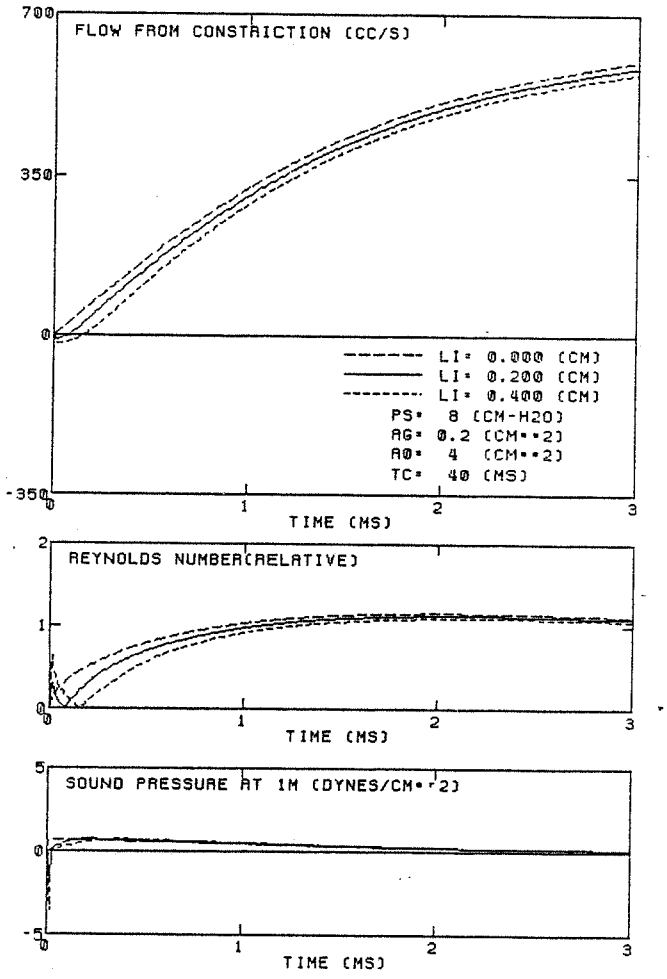


Fig 2 L'effet de la longueur initiale de la constriction, Li, sur U_m (en haut), Re (milieu), et P_r (en bas).

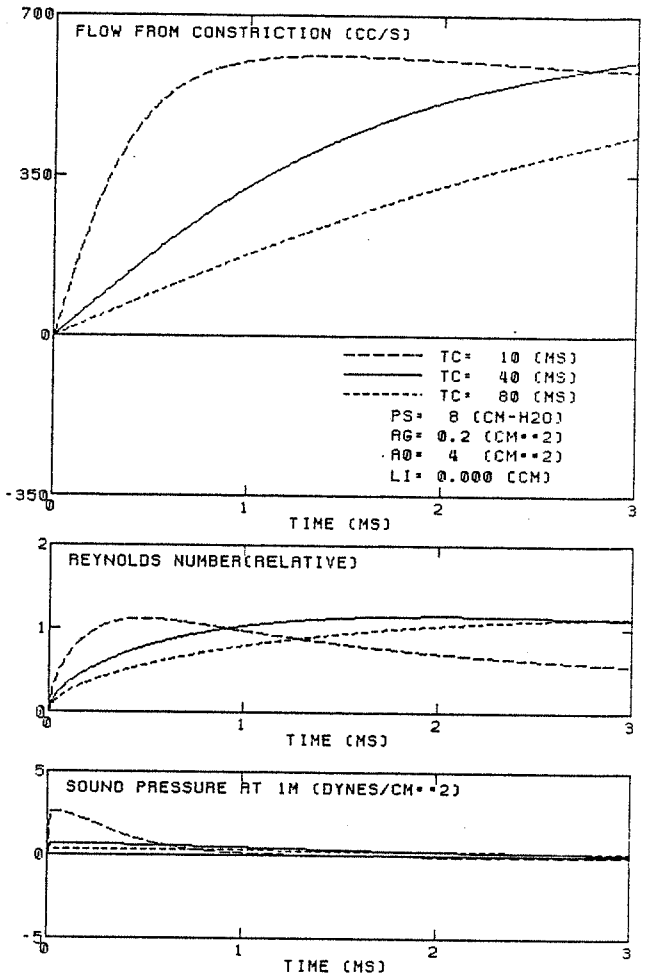


Fig 3 L'effet de la constante de temps du relâchement, T_c.

Due a une augmentation dans l'intensité de U_s, les courbes de U_m sont dérivées légèrement vers la droite, lorsque L_i augmente du Ø.Øcm a Ø.4cm. On peut noter que les deux courbes U_m, pour L_i=Ø.2cm et Ø.4 cm, augmentent à partir de valeurs négatives, comme attendu. Chacune des réponses correspondantes P_r montre l'impulsion négative au début avec un niveau sonore relativement élevé, qui est suivi par la "queue". La pente de la montée initiale de la "queue" varie en fonction de L_i. Ceci est dû à l'effet des inductances en série, qui est proportionnel a L_i. Pour un L_i petit, l'amplitude de l'impulsion au debut est petite, et la montée de la "queue" est abrupte. Au contraire, pour un grand L_i, l'impulsion de début est forte, et la montée est moins abrupte. L'effet principal de L_i se produit en conséquence, grosso modo, sur la pente spectrale de la source cohérente.

Les effets de T_c sont montrés sur la Figure 3. Ici nous utilisons le temps constant, T_c, comme un paramètre pour le contrôle de la rapidité de l'expansion de la constriction, A_o/T_c. Comme attendu, une rapidité d'expansion plus rapide résulte en une montée plus abrupte de U_m, comme il est indiqué en haut de la figure. En conséquence, T_c et A_o ont un effet sur les deux sources, la source cohérente et celle de bruit. On peut noter que la montée du temps de U_m pour T_c=1Ø msec est de 1 msec environ. La courbe Re correspondante, montrée au milieu, indique un pic a Ø,5 msec après le relâchement. Ceci est un temps beaucoup plus court que celui auquel on pourrait s'attendre avec un temps contant d'expansion de la constriction, qui est de 1Ø msec dans ce cas particulier. En ajustant A_g de facon appropriée, il devrait être possible de produire une montée abrupte au début de bruit du un burst, même si le relâchement de l'occlusion est relativement graduel.

CONCLUSIONS

Il devient donc évident que le conduit vocal peut être excité par deux sources différentes de sons, une source cohérente et une source de bruit. L'excitation cohérente devrait se produire au tout début du relachement. Alors que l'intensité de la source de bruit augmente de facon un peu moins abrupte que celle de la source cohérente, son intensite elle peut être soutenue pendant une période relativement longue. Ceci est probablement ce qui se produit pour une occlusive sourde ou sonore. Pendant le relâchement d'une nasale, /m/ ou /n/, le conduit vocal devrait être seulement excité par la source cohérente, puisqu'il n'y a pas formation de pression intraorale pendant le relâchement.

La figure 4 illustre le spectogramme et la portion de signal correspondant à la sequence /mi/. On peut remarquer sur le spectogramme que le relachement du /m/ est concommitent avec une "barre de relachement", indique par une flêche. La forme du signal indique clairement l'existence d'une excitation suplémentaire qui est superposée sur le son voisé (excité par les vibrations des cordes vocales). Dans cette représentation du signal, la polarité est renversée, c'est-à-dire qu'une pression acoustique positive correspond à la partie inférieure. La direction de l'excitation suplémentaire semble être négative, suggérant qu'au moment du relâchement, le conduit vocal a été excite seulement par une impulsion négative au début dûe à la source coherente.

Nous faisons maintenant des expériences de synthèse des occlusives en exploitant les résultats décrits dans cet article. On espère que cette étude permettra d'apporter quelques lumières sur la relevance perceptive de ces sources d'excitation.

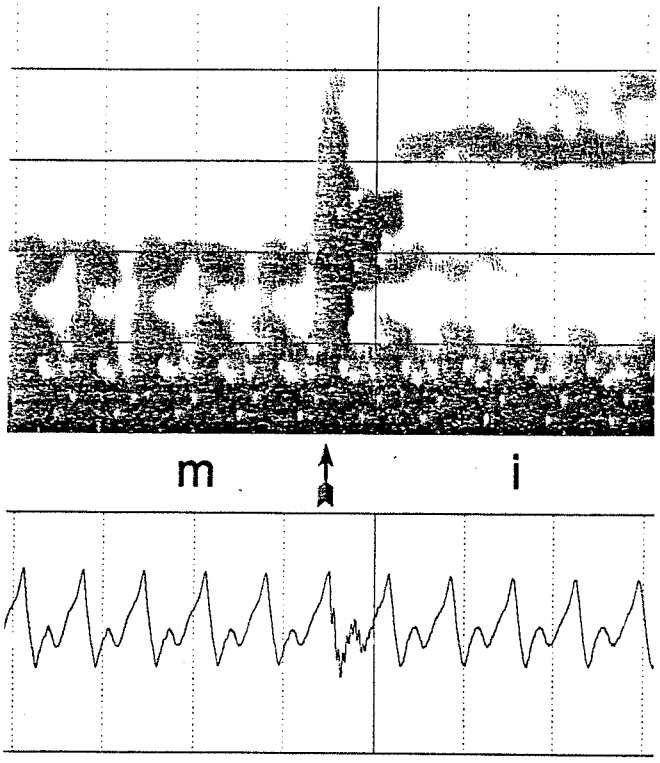


Fig 4 Spectogramme (en haut) et onde du signal (en bas) corresponding à la sequence /mi/. La flêche indique le moment du relâchement de /m/.

REMERCIEMENTS

Cette étude a été réalisée à MIT, avec le support partiel du contrat NINCDS(NS-04332).

BIBIOGRAPHIE

[1] Shadle,C.H., The Acoustics of Fricative Consonants, PhD thesis, MIT, 1985.

[2] Flanagan,J.L. and Ishizaka,K., "Automatic Generation of Voiceless Excitation in a Vocal Cord-Vocal Tract Speech Synthesizer," IEEE/ASSP-24, 2, 163-170, 1976.

[3] Stevens, K.N., "Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations," J. Aco. Soc. Am., 50, 4(Part2), 1180-1192, 1971.

[4] Maeda, S., "A Digital Simulation Method of the Vocal-Tract System," Speech Communication, 1, 199-229, 1982.

[5] Fujimura,O., "Labial Stop and Nasal Consonants: a Motion Picture Study and its Acoustic Implications," J. of Speech Hearing Research, 4, 3, 233-247, 1961.

ETUDE ELECTROGLOTTOGRAPHIQUE DE LA DYNAMIQUE D'ACQUISITION DU TRAIT VOISEMENT DES OCCLUSIVES "SONORES" EN FRANÇAIS

Jean-Pierre Goudaillier

Laboratoire de Phonétique, UER de Linguistique, Université R. Descartes / Paris; 104, Quai Clichy, 92110 CLICHY

In this electroglottographic study (EGG), 2046 'voiced' stops / b /, / d / & / g / are analysed, as they have been pronounced by thirty-one 6-10 year old children of the North of France (Vervins, Aisne). Thanks to a classification in terms of laryngeal vibrations times (L.V.T.), it is possible a) to determine how devoicing differences can be established between boys and girls at this age, b) to determine that devoicing is less important (in percentages) by children of the higher school levels (CM1 & CM2) than by children of the lower school levels (CP & CE1).
Henceforth, this study contributes to a better knowledge of the dynamics of acquisition of the feature 'voicing' by young French children.

1. INTRODUCTION

Le trait voisement des consonnes occlusives intéressent les chercheurs depuis de nombreuses années, plus particulièrement l'indice V.O.T. [1] [12][13][17]. En ce qui concerne le français, le délai d'établissement du voisement (D.E.V.), entre autres indices, a été utilisé par l'équipe du Laboratoire de Phonétique de Bruxelles, sans que ne soit pour autant oublié le rôle important de la durée d'occlusion et d'autres indices pour la perception [3][15][16][17] (cf. aussi [2] pour le rôle du D.E.V. dans la perception des occlusives et des constrictives par des déficients auditifs). Pour l'anglais et l'espagnol (de même pour la variante portoricaine), des études (cf., entre autres, [4][14]) nous renseignent au sujet de l'acquisition du trait voisement par de jeunes enfants.

La présente recherche fait suite à une série d'études ([6][7][8][9][10]), que j'ai menées ces dernières années à propos de la langue orale de l'enfant francophone a) dans le cadre de l'I.N.R.P. jusqu'en 1980 [11], b) dans le cadre du C.I.L.O.(1) depuis 1980. Ces études s'inscrivent depuis 1983 dans le cadre d'un projet intégré franco-québécois de recherche intitulé "systèmes phonologiques d'enfants scolarisés francophones français et québécois âgés de 6-10 ans"(2).

Concernant le trait voisement, mes recherches ont permis d'établir que les phénomènes de dévoisement des occlusives phonologiquement "sonores", chez des enfants en début de scolarité (CP et CE1) dans le Nord de la France, relèvent de régionalismes et ne sont pas imputables à un quelconque retard d'acquisition [6][7]. Par ailleurs, j'ai pu établir que plus le taux de dévoisement des occlusives "sonores" est important (en pourcentages) chez un enfant, plus celui-ci opère une distinction de force articulatoire (différence, entre autres, de durée des articulations) entre /ptk/ et /bdg/ [8][9].

Dans cette étude, je procède à un classement des occlusives "sonores" en termes de *moments d'apparition des vibrations laryngiennes* (*M.A.V.L.*) [5; p. 142-148] en mettant en valeur la différenciation sexuelle, qui intervient, ainsi que la dynamique d'acquisition du trait voisement : augmentation (en pourcentages) du nombre d'articulations entièrement voisées pour les enfants de CM1 et CM2.

2. POPULATION D'ENQUÊTE ET METHODE INSTRUMENTALE

En mai 1980, les réponses de 18 enfants de CP et CE1 de l'Ecole Ceccaldi de Vervins (Aisne) (pour le questionnaire de l'enquête, cf. [11], p. 113-124) ont été enregistrées (microphone + électroglottographe). La méthodologie instrumentale est celle qui a été présentée dans mes publications antérieures (cf., entre autres, [8], p. 267). En mai 1983, 12 autres enfants de CP et CE1 de cette même école ont été enregistrés, ainsi que 12 enfants de CM1 et CM2. L'ensemble des enquêtes a fourni 5388 occlusives "sourdes" et "sonores" analysables, dont 2046 occlusives "sonores" (voir le Tableau 1 de la page suivante pour plus de détails).

Les mesures ont été effectuées à partir de tracés oscillographiques (microphone) et électroglottographiques (EGG) de chacune de ces occlusives (cf. Figures 1 à 4); pour certaines séquences, la courbe d'énergie (intensité globale du signal) a aussi été utilisée pour procéder à la segmentation, à l'identification des divers segments et à leurs mesures.

(1) Centre Interuniversitaire de recherches sur la Langue Orale de l'enfant et de l'adolescent.

(2) Responsables du projet : 1983 et 1984 : J.-P. Goudaillier; 1985 : Anne-Marie Houdebine.

Nr. d'enfants	Niveau scolaire	Sexe	Date d'enquête	Nr. total d'occlusives sonores	Nr. total d'occlusives
4	CP	F	5/80	415	1310[1]
7	CP	M	5/80		
4	CE1	F	5/80	388	998[2]
3	CE1	M	5/80		
3	CP	F	5/83	244	585
3	CP	M	5/83		
3	CE1	F	5/83	320	815
3	CE1	M	5/83		
3	CM1	F	5/83	325	812
3	CM1	M	5/83		
3	CM2	F	5/83	354	868
3	CM2	M	5/83		
42[3]				2046	5388

Tableau 1 : Population d'enquête et occlusives analysées (par niveau scolaire)

3. MOMENTS D'APPARITION DES VIBRATIONS LARYNGIENNES (M.A.V.L.)

Si l'on observe les séquences [bat$^{(h)}$o], [d$^{(h)}\phi$], [$\underset{\leftrightarrow}{d}{}^{(h)}\phi$] et [$\underset{o}{bat}{}^{(h)}$o] ci-contre (Figures 1 à 4), on constate quatre types de réalisations différentes à l'initiale de chacune d'entre elles. Le [b] de [bat$^{(h)}$o] est entièrement voisé (durée : 120 ms). L'explosion elle-même est voisée (aucune interruption du voisement lors du passage de la consonne à la voyelle). Les initiales de [d$^{(h)}\phi$] et [$\underset{\leftrightarrow}{d}{}^{(h)}\phi$] sont différentes sur ce dernier point : leurs explosions ne sont pas voisées; de même en ce qui concerne les bruits de friction suivant ces explosions (V.O.T. positifs : +10 ms et +15 ms respectivement); l'occlusion de [d$^{(h)}\phi$] est entièrement voisée (durée : 125 ms), celle de [$\underset{\leftrightarrow}{d}{}^{(h)}\phi$] ne l'est que partiellement (segment voisé : 70 ms; segment non voisé : 45 ms; durée de l'occlusion : 115 ms). L'occlusion de [$\underset{o}{bat}{}^{(h)}$o] est entièrement dévoisée (V.O.T. positif : +5 ms). En termes de *moments d'apparition des vibrations laryngiennes* (M.A.V.L.), [bat$^{(h)}$o], [d$^{(h)}\phi$], [$\underset{\leftrightarrow}{d}{}^{(h)}\phi$] et [$\underset{o}{bat}{}^{(h)}$o] sont respectivement de types 1, 2, 3 et 4.

4. DYNAMIQUE D'ACQUISITION DU TRAIT DE VOISEMENT

Les 2046 occlusives analysées ont pu être classées en fonction des quatre types de M.A.V.L. définis en 3. ci-dessus. Les résultats sont indiqués dans le Tableau 2 ci-contre, ainsi qu'aux Schémas 1, 2 et 3, et permettent un certain nombre de constatations relatives à la dynamique d'acquisition du trait voisement.

- une première constatation s'impose : quel que soit le niveau scolaire des enfants, les filles désonorisent plus leurs articulations que ne le font les garçons. Par exemple, en CP et CE1, 30% des réalisations de / b / sont dévoisées (M.A.V.L. 2, 3 et 4) chez les filles, 8% chez les garçons; en CM1 et CM2, ces pourcentages sont respectivement de 20% et 2,5% (Schéma 1; cf. Tableau 2 pour d'autres exemples).
- deuxième constatation : les pourcentages d'articulations dévoisées décroissent pour les niveaux scolaires les plus élevés (CM1 et CM2), sans pour autant atteindre des valeurs négligeables. Pour / b / (cf. ci-dessus), / d / et / g / (cf. Tableau 2 et Schémas 2 et 3), le dévoisement est relevé dans 8%, 38,5% et 49,5% des cas chez les garçons de CP et CE1, dans 2,5%, 30% et 34,5% des cas chez ceux de CM1 et CM2 (à noter les pourcentages importants pour / d / et / g /)[4].
- autre fait important : les réalisations à M.A.V.L. de type 4 (dévoisement complet) n'existent pratiquement plus en CM1 et CM2 : elles ne représentent plus que 3% au total, ce qui est négligeable.
- on constate, de manière corollaire, une augmentation en pourcentages des articulations à M.A.V.L. de type 2 : pour / d / (cf. Tableau 2 et Schéma 2), par exemple, les réalisations à occlusion sonore et à explosion + friction sourdes ne constituent que 18,5% et 13,5% pour les filles et les garçons de CP et CE1 respectivement, alors que ces pourcentages sont de 42% et 25% en CM1 et CM2. La même tendance est constatée pour / b / et / g /(cf. Tableau 2 et Schémas 1 et 3).

5. CONCLUSION

La présente étude montre l'utilité d'opérer un classement en termes de moments d'apparition des vibrations laryngiennes (M.A.V.L.) des réalisations voisées, dévoisées et non voisées des consonnes occlusives "sonores". Elle permet par ailleurs de mettre au jour un certain nombre de faits relatifs à la dynamique d'acquisition du trait de voisement, tels que ceux-ci peuvent être relevées dans une population d'enfants scolarisés francophones français du Nord de la France : un dévoisement plus important chez les filles que chez les garçons, ce quel que soit leur niveau scolaire (du CP au CM2); une diminution du nombre d'articulations dévoisées chez les enfants de CM1 et CM2, ce par rapport à ce qui peut être relevé en CP et CE1; ceci s'accompagne du fait que les articulations à M.A.V.L. de type 4 (réalisations entièrement dévoisées) dispa-

(1) cf. [6].
(2) cf. [8][9].
(3) dont 11 enfants de CM1 et CM2 déjà enregistrés en 1980 (en CP et CE1).

(4) Les présentes données confirment ce que j'avais déjà pu constater par ailleurs : les réalisations de / b / présentent les pourcentages de dévoisement les moins élevés et celles de / g /les pourcentages les plus élevés ([8]; p. 268); cf. à ce sujet Tableau 2.

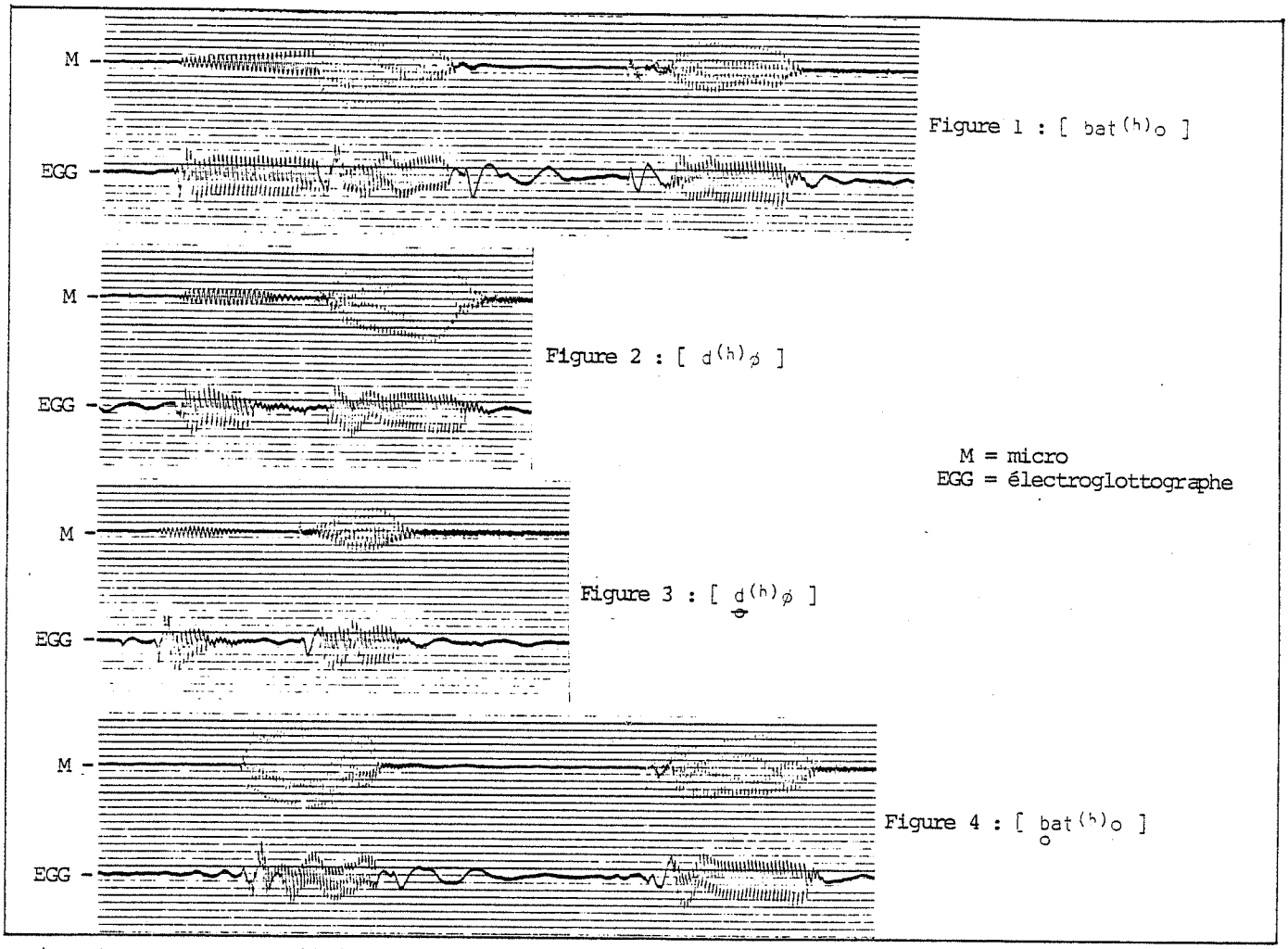


Figure 1 : [bat$^{(h)}$o]

Figure 2 : [d$^{(h)}\phi$]

M = micro
EGG = électroglottographe

Figure 3 : [$\underset{o}{d}^{(h)}\phi$]

Figure 4 : [$\underset{o}{bat}^{(h)}$o]

raissent en presque totalité en CM1 et CM2, alors que les articulations à M.A.V.L. de type 2 (explosion + friction non voisées) augmentent, quant à elles, de manière conséquente en pourcentages. Dans le cadre du projet intégré franco-québécois indiqué en 1., le prolongement nécessaire à une telle étude est en cours de réalisation : il s'agit de la comparaison de ces données avec celles d'autres régions en France et avec celles relatives au Québec.

REFERENCES BIBLIOGRAPHIQUES

[1] Christian Abry & Louis-Jean Boë, "Le trait de voisement, ses indices et leurs propriétés. Implication pour les détections automatique et perceptive d'une analyse phonologique au-delà du trait", *Recherches sur la prosodie du français*, 1979,p. 57-93.
[2] S. Barth, R. Ben Fadhel & G. Majo, "Le D.E.V. dans la parole des déficients auditifs - Comparaison avec les entendants", *11èmes J.E.P.*, 1980, p. 65-75.

		/ b /				/ d /				/ g /			
		1	2	3	4	1	2	3	4	1	2	3	4
CP + CE1 (1980 + 1983) 30 enfants (1367 occl.)	♀ (14 enfants)	70,0%	3,5%	16,5%	10,0%	41,5%	18,5%	21,5%	18,5%	37,0%	16,0%	27,5%	19,5%
	♂ (16 enfants)	92,0%	2,0%	1,0%	5,0%	61,5%	13,5%	10,5%	14,5%	50,5%	24,0%	11,5%	14,0%
CM1 + CM2 (1983) 12 enfants (679 occl.)	♀ (6 enfants)	80,0%	14,0%	4,0%	2,0%	42,0%	42,0%	15,0%	1,0%	40,0%	48,0%	12,0%	
	♂ (6 enfants)	97,5%	2,0%	0,5%		70,0%	25,0%	5,0%		65,5%	27,5%	7,0%	

Tableau 2 : Pourcentages des différents types (1, 2, 3 et 4) de M.A.V.L. des occlusives / b /, / d / et / g /, toutes positions confondues.

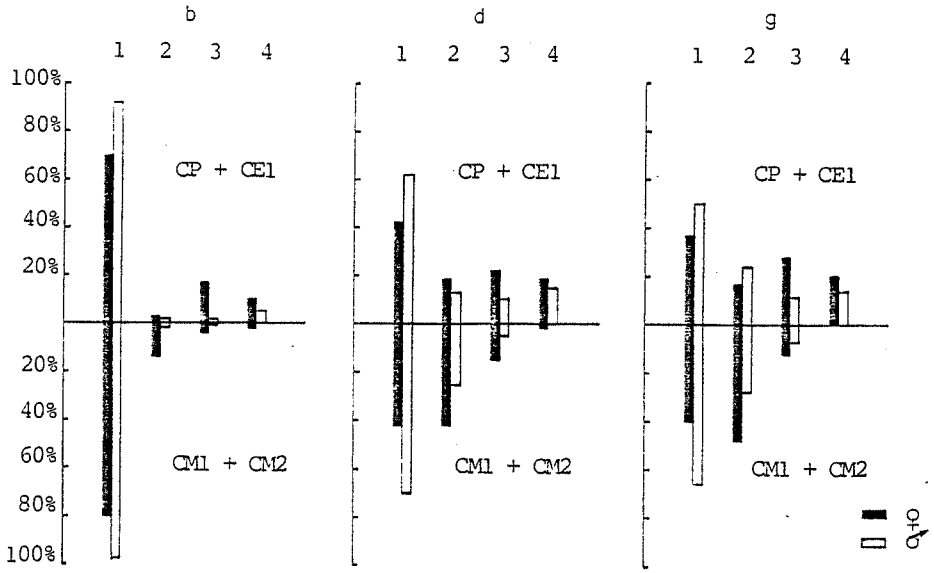


Schéma 1 Schéma 2 Schéma 3

Comparaison des pourcentages des divers types de M.A.V.L. pour les occlusives "sonores" / b / (schéma 1), / d / (schéma 2) et / g / (schéma 3) en fonction du niveau scolaire et du sexe des enfants.

[3] Pierre Bejster, "Le rôle de la durée de l'occlusion dans la perception du voisement chez l'enfant", *R.A. Labo Phonétique*, Bruxelles, 10/1, 1976, p.23-38.

[4] Rebecca E. Eilers, D. Kimbrough Oller & Carmen R. Benito-Garcia, "The acquisition of voicing contrasts in Spanish and English learning infants and children : a longitudinal study", *Journal of Child Language*, 11, 1984, p. 313-336.

[5] Jean-Pierre Goudaillier, *Phonologie fonctionnelle et phonétique expérimentale - Exemples empruntés au luxembourgeois*, Hamburg, Buske Verlag, 1981 (Hamburger Phonetische Beiträge, 36), XII + 476 pages.

[6] J.-P. Goudaillier, "Exemple de traitement de l' opposition de "sonorité" par des enfants de Cours-Préparatoire - Utilisation de la méthode électroglottographique", *12èmes J.E.P.*, 1981, p. 377-391.

[7] J.-P. Goudaillier, "Voicing contrast by eleven 6-7 year old children of a school in the North of France", *14th Annual Meeting of the Societas Linguistica Europaea*, Copenhague, 1981 (non publié).

[8] J.-P. Goudaillier, "Diverses possibilités de matérialisation du trait de voisement : étude électroglottographiques des occlusives d'enfants âgés de 7-8 ans d'un Cours-Elémentaire 1ère Année du Nord de la France", *11ème I.C.A.*, Paris, 1983, p. 267-270.

[9] J.-P. Goudaillier, "Voicing contrast in French : the case of seven 7-8 year old children of the North of France (an electroglottographic study)", *10th Intern. Congress of Phon. Sciences*, Utrecht,1983,p.627.

[10] J.-P. Goudaillier, "Etude électroglottographique du voisement. Le cas des occlusives d'enfants scolarisés du Nord de la France", *Journée d'Etudes 'Phonétique instrumentale et Linguistique'* (Laboratoire de Phonétique, Université R. Descartes, Paris, Mai 1982), p. 121-124 (Etudes de Phonologie, phonétique et linguistique descriptive du français, 1, Hamburg, Buske Verlag, 1984).

[11] Anne-Marie Houdebine, *Aspects de la langue orale des enfants à l'entrée au Cours-Préparatoire* (ouvrage collectif sous la direction de...), I.N.R.P./Paris, C.R.D.P./Besançon, 1983, 359 pages.

[12] Patricia E. Keating, "Phonetic and phonological representation of stop consonant voicing", *Language*, 60/2, 1984, p. 286-319.

[13] Leigh Lisker & Arthur S. Abramson, "A cross-language study of voicing in initial stops : acoustical measurements", *Word*, 20, 1964, p. 384-422.

[14] Marlys A. Macken & David Barton, "The acquisition of the voicing contrast in English : a study of voice onset time in word-initial stop consonants", *Journal of Child Language*, 7, 1979, p. 41-74.

[15] Willy Serniclaes, "Fenêtre de prélèvement temporel des indices d'occlusives", *13èmes J.E.P.*, 1984, p. 69-78.

[16] Willy Serniclaes & Pierre Bejster, "Différences interlinguistiques dans le traitement perceptif des indices de voisement", *Rapport d'Activités de l'Institut de Phonétique de l'Université libre de Bruxelles*, 10/1, 1978, p. 83-94.

[17] Max Wajskop, "Indices temporels des occlusives intervocaliques en français", *Rapport d'Activités de l'Institut de Phonétique de l'Université libre de Bruxelles*, 12/2, 1978, p. 71-98.

CONTRIBUTION A LA CARACTERISATION DE PATHOLOGIES AUDIO-PHONATOIRES PAR LE SPECTRE VOCAL MOYEN

Bernard Harmegnies

Département de Phonétique et de Psycho-acoustique.
Université de l'Etat
avenue du Champ de Mars, 7000 Mons - Belgique.

The voices of six subjects with severe hearing impairment have been studied by means of the long term average spectrum. The effect of the hearing aids on the voice quality as well as the difference between pathological and normal voices have been studied thanks to the interspectral correlation coefficient. Such a technique appears to be of use for purpose of objectivating the judgments of quality of the deafs' voices.

1. INTRODUCTION

A l'occasion de recherches récentes [1,2], nous avons étudié comment le spectre moyen à long terme (SMLT) de la voix subit l'influence de diverses sources de variations. Les répercussions des changements de texte, de langue et d'appareil phonateur ont ainsi été étudiées, de même que la variation naturelle du SMLT en l'absence de sources de variations contrôlées. Toutes ces sources de variations ont comme caractéristique commune le fait d'entretenir des rapports directs avec la production du matériel vocal. Dans la présente étude, nous nous centrerons au contraire sur l'effet de sources de variations indirectement liées aux mécanismes de la phonation. Dans ce but, nous avons choisi d'étudier ici les répercussions vocales de perturbations de la boucle de régulation audio-phonatoire. Un public d'élection s'offrait à nous : celui des déficients auditifs. Chez ces sujets, en effet, non seulement le trouble auditif lui-même mais aussi le recours aux aides prothétiques constituent d'importantes modifications de l'appareil récepteur, dont il semble logique de considérer qu'elles peuvent influer sur l'émission vocale. Par ailleurs, ce projet nous semblait se justifier d'autant plus que la majorité des recherches en matière de voix de sourds négligent l'étude de la qualité vocale, se centrant préférentiellement sur les aspects rythmiques [3,4] ou intonatifs [5,6] de l'élocution du déficient auditif.

2. EXPERIMENTATION

2.1. Echantillonnage

Les sujets étudiés ici sont issus de l'échantillon occasionnel des sujets traités à l'IES de Mons-Ghlin. Ils ont été sélectionnés au nombre de six afin de constituer un groupe qui fût le plus homogène possible. Tous sont atteints de surdité congénitale profonde de perception et étaient âgés de 16 à 19 ans au moment de l'expérience. Le tableau 1 fournit à leur propos des données complémentaires. L'intelligibilité habituelle des sujets y est ainsi grossièrement évaluée dans une échelle à trois divisions, à partir d'une appréciation subjective fournie par l'équipe éducative ayant ces individus en charge. Le déficit moyen y est, quant à lui, calculé selon la formule dite de FOURNIER, qui consiste à moyenner, pour une oreille déterminée, ses pertes à 500, 1000 et 2000 Hz mesurées par audiométrie tonale. Ajoutons que tous les sujets présentent un déficit auditif croissant légèrement avec la fréquence. Tous sont appareillés au moyen de prothèses amplifiant préférentiellement les basses fréquences (surtout en dessous de 2 KHz).

		Déficit auditif		
Sujets	Age	Oreille gauche	Oreille droite	Intelligibilité
BOR	17	98 dB	92 dB	-
BRA	19	95 dB	97 dB	+
DEL	17	98 dB	102 dB	+/-
HAR	18	113 dB	118 dB	+/-
LUX	19	103 dB	112 dB	-
MAR	16	103 dB	102 dB	-

TABLEAU 1 : caracteristiques des sujets de l'échantillon.

Enfin, tous les sujets ont pour langue maternelle le français; ils maitrisent bien la lecture et aucun ne présente de trouble associé.

2.2. Dispositif expérimental

Chaque sujet a été invité à lire un texte continu d'une durée approximative de 15 secondes dont la composition phonémique reflète celle du français. Les corpus ainsi produits sont identiques à ceux recueillis lors d'expériences antérieures impliquant 10 sujets normaux [1,2]. Dans le cadre de la présente expérience, les six sujets produisirent chacun le corpus dans 2 conditions : une fois dépourvus de leurs prothèses auditives (condition 1) et une fois munis de celles-ci (condition 2). Les données furent recueillies dans le même environnement (cabine insonorisée) et au moyen de la même chaîne d'enregistrement (magnétophone Nagra IV S et microphone Neuman KM 84) que celles exploitées dans les expériences antérieures impliquant des sujets normaux.

2.3. Analyse acoustique

Chaque production du texte a fait l'objet d'une analyse spectrale par la méthode du spectre moyen à long terme (SMLT), au moyen de l'analyseur digital BRUEL KJOER 2033. Chaque SMLT ainsi calculé constitue la moyenne de 256 spectres instantanés. Ceux-ci sont établis via un algorithme de FFT produisant des spectres définis sur 400 fréquences. Dans le cas qui nous occupe, la gamme d'analyse retenue s'étend jusqu'à 5000 Hz. Pour chaque spectre considéré, nous disposons donc d'une mesure d'énergie tous les 12.5 Hz sur toute la gamme d'analyse.

2.4. Traitement statistique

C'est le coefficient de corrélation interspectrale qui a été retenu en vue de la comparaison des spectres. Défini et testé antérieurement [1,2], il devait en effet nous permettre de confronter les variations spectrales observées lors de la présente expérimentation à celles constatées chez des sujets normaux.

Rappelons que le coefficient de corrélation interspectrale est un indice de similarité spectrale variant entre 0 et 1 en valeur absolue et dont la valeur est d'autant plus élevée que les spectres se ressemblent. Soit $r_{ss'}$, le coefficient de corrélation entre les spectres s et s' on l'obtient par :

$$r_{ss'} = \frac{\sum s_i s_i' - \frac{1}{K}\left[\sum s_i \sum s_i'\right]}{\sqrt{\left[\sum s_i^2 - \frac{(\sum s_i)^2}{K}\right]\left[\sum s_i'^2 - \frac{(\sum s_i')^2}{K}\right]}}$$

où K est le nombre d'intensités mesurées, s_i la $i^{ème}$ intensité en dB du spectre S et s', la $i^{ème}$. intensité en dB du spectre S' (toutes les sommations sont effectuées pour toute valeur entière de i, de 1 à K).

Nos expériences précédentes ont permis de mettre en évidence des valeurs moyennes de corrélation interspectrale en fonction des types de comparaisons spectrales impliquées. Ces valeurs moyennes, qui constituent les prémisses d'une sémiologie de la corrélation interspectrale, sont rappelées, à titre de référence, au tableau 2.

Types de comparaisons	corrélation moyenne
intra sujet intra texte	.93
intra sujet inter texte	.91
intra sujet inter langue	.88
inter sujet intra texte	.80

TABLEAU 2 : Valeurs moyennes des coefficients de corrélation interspectrale en fonction du type de comparaison [1,2].

3. RESULTATS

Nous avons, dans un premier temps, comparé, pour chaque sujet, le SMLT du corpus produit en condition 1 avec celui du corpus produit en condition 2. Les résultats de cette confrontation figurent au tableau 3. Dans la suite, nous appellerons ces comparaisons "comparaisons intercondition".

Sujet	Corrélation interspectrale
BOR	.7400
BRA	.9543
DEL	.9054
HAR	.9607
LUX	.8612
MAR	.9024

TABLEAU 3 : Coefficients de corrélation interspectrale obtenus à l'occasion des comparaisons intercondition.

Dans un second temps, nous avons tenté d'apprécier dans quelle mesure les voix de nos sujets sourds pouvaient se différencier des voix de sujets normaux. Nous avons, à cet effet, eu recours aux enregistrements réalisés en vue de recherches précédentes : grâce à ceux-ci, nous avons établi, pour chacun de dix sujets normaux, un SMLT moyen réalisé à partir de 10 productions du corpus-test. Chacun des 10 SMLT moyens ainsi calculés a été confronté aux 12 SMLT établis à partir des productions des six sujets sourds. Il en est résulté 120 coefficients de corrélation interspectrale. Cependant, afin d'alléger la présentation des données, nous ne reproduisons au

tableau 4 que 12 moyennes calculées sur ces 120 coefficients. Chaque moyenne y représente 10 valeurs de corrélation interspectrale : les dix coefficients correspondant aux 10 comparaisons entre les SMLT moyens des 10 sujets normaux et le SMLT du sujet sourd considéré sous la condition considérée.

Sujet	Condition 1	Condition 2
BOR	.5467 (.1002)	.7160 (.0452)
BRA	.8473 (.0579)	.8352 (.0433)
DEL	.6474 (.0713)	.6886 (.0471)
HAR	.6003 (.0657)	.5922 (.0661)
LUX	-.0694 (.1090)	.1413 (.0812)
MAR	.4451 (.0512)	.4625 (.0662)

TABLEAU 4 : Moyennes et écarts types (entre parenthèses) des coefficients de corrélation interspectrale entre SMLT de sujets sourds (en condition 1 et en condition 2) et SMLT moyens de 10 sujets normaux.

4. DISCUSSION

Il semble évident, au vu du tableau 3, que le changement de condition d'élocution n'agit pas de manière homogène sur l'ensemble de nos sujets. Deux de ceux-ci (BRA et HAR) présentent des valeurs de corrélation intercondition supérieures aux corrélations intrasujet intratexte moyennes relevées sur les sujets normaux. Deux sujets (BOR et LUX) présentent par contre des valeurs de corrélation intercondition incompatibles avec les corrélations intrasujet (LUX), voire même intersujet (BOR) caractéristiques de sujets normaux. Deux sujets, enfin, (MAR et DEL) présentent des corrélations intercondition médiocres (.90 environ) mais proches de celles que l'on obtient lorsque l'on compare les SMLT provenant de corpus différents produits par un même sujet.

La comparaison avec les voix normales indique, elle aussi, une grande hétérogénéité de réactions au sein de l'échantillon. Un seul sujet (BRA) s'avère pouvoir être confondu avec un sujet normal, qu'il soit appareillé ou non. Les autres sujets présentent sous les deux conditions des valeurs de corrélation qu'il serait peu probable (HAR, DEL, BOR) voire très improbable (MAR, LUX) d'observer lors de comparaisons intersujet impliquant des sujets normaux. En outre, les deux sujets qui présentent les corrélations intercondition les plus faibles (MAR, LUX) présentent une très nette diminution de la corrélation avec les voix normales lorsqu'ils sont dépourvus de leurs prothèses (Δ_r = .21 pour LUX et Δ_r = .17 pour BOR). Les deux sujets présentant des corrélations intercondition médiocres (HAR, DEL) montrent, en condition 1, un éloignement par rapport aux voix normales (Δ_r = .02 pour MAR et Δ_r = .04 pour DEL) assez faible pour qu'on puisse douter de sa signification. Enfin, les voix des deux sujets présentant des corrélations intercondition élevées (BRA et HAR) sont caractérisées par une très faible augmentation de la corrélation avec les voix normales lorsque les sujets sont privés de leurs prothèses (Δ_r = -.01 et Δ_r = -.008); ici, plus encore, une très grande méfiance s'impose vis-à-vis de la signification statistique de cette différence.

Il n'est par ailleurs pas inintéressant de rapprocher ces observations de celles relatives à l'intelligibilité subjective. On peut ainsi constater que les deux sujets dont la voix est fortement modifiée par la privation des prothèses sont considérés comme des locuteurs peu intelligibles par leur entourage éducatif. Par contre, les deux sujets ne présentant pas de différence vocale intercondition sont considérés comme moyen ou bon locuteurs.

L'interprétation de l'hétérogénéité des réactions vocales de nos sujets à la privation d'appareillage prothétique est évidemment très délicate. A priori, il est en effet malaisé de différencier ces six individus, qui ont été sélectionnés en vue de constituer un groupe homogène. Nous avancerons cependant l'hypothèse explicative suivante : certains sujets auraient déjà atteint un stade d'équilibre fonctionnel de la phonation, d'autres pas. Les premiers présentent généralement une bonne intelligibilité, leur voix n'est pas très éloignée d'une voix normale et, finalement, une privation momentanée de leur prothèse auditive ne modifie guère la qualité de leur émission vocale: les comportements phonatoires, étant bien installés, ne nécessitent que peu de contrôle auditif. Chez les seconds, les comportements phonatoires sont mal stabilisés, ils nécessitent beaucoup de régulations et toutes les sources de contrôle sont donc importantes : la privation de la prothèse constitue donc un handicap qui fait chuter les performances vocales; ces sujets sont ceux qui présentent, surtout non appareillés, une mauvaise intelligibilité et des caractéristiques acoustiques vocales éloignées de celles des sujets normaux.

5. CONCLUSIONS

Au terme de cette courte étude, nous désirons insister sur son caractère exploratoire: la taille de l'échantillon testé n'autorise bien sûr aucune généralisation des constatations effectuées.

Certaines tendances apparaissent cependant assez nettement. Ainsi, la prothèse auditive semble agir de manière très hétérogène sur les voix des sujets de notre échantillon. Pour deux d'entre eux, en effet, elle semble n'exercer aucune action; deux autres y semblent au contraire très sensibles; deux enfin semblent montrer un faible effet. En outre, lorsque la voix du sujet sourd est modifiée sous l'effet de l'appareillage prothétique, elle l'est dans le sens d'un rapprochement en direction des voix normales. Les observations acoustiques semblent en outre objectivement liées avec l'appréciation subjective de l'intelligibilité.

Nous retiendrons, en bref, que la méthodologie mise en oeuvre semble se montrer bien adaptée à l'étude des voix de sourds. En particulier, l'idée d'une mesure objective de l'effet de l'appareillage prothétique sur la qualité vocale semble trouver ici un début de concrétisation. La notion même d'intelligibilité semble elle aussi pouvoir être abordée avec moins de subjectivité.

Finalement, les importantes variations du SMLT constatées nous semblent apporter une justification supplémentaire à l'étude acoustique des variations vocales induites par des sources de variations extérieures aux mécanismes phonatoires.

6. REFERENCES

[1] HARMEGNIES (B.), "Traitements et utilisations des spectres vocaux moyens", Actes des 13èmes J.E.P., Bruxelles, 1984.

[2] HARMEGNIES (B.), LANDERCY (A.), "Language features in the long-term average spectrum", Revue de Phonétique Appliquée, 73-74-75, 1985.

[3] NICKERSON (R.S.), STEVENS (K.N.), BROOTHROY (A.), ROLLINS (A.), "Some observations on timing in the speech of the deaf and hearing speakers", BBN report 2905, 1974.

[4] BOONE (D.R.), Modification of the voices of deaf children, Volta Review, 68, 686-694, 1966.

[5] MARTONY (J), "On the correction of the voice pitch level for the severely hard of hearing subjects", American Annals of the Deaf, 113, 195-202, 1968.

[6] ENGEN (T), ENGEN (E.A.), CLARKSON (R.L.), BLACKWELL (P.M.), "Discrimination of intonation by hearing - impaired children", Applied Psycholinguistics, 4, 149-160, 1983.

MODELISATION ARTICULATOIRE DU CONDUIT VOCAL. EXPLORATION ET EXPLOITATION.

PERRIER Pascal, BOE Louis-Jean, MAJID SHIHAB Rajaa & GUERIN Bernard

INSTITUT DE LA COMMUNICATION PARLEE (LA CNRS 368)
I.N.P.G. - 46 Avenue Felix Viallet
38031 GRENOBLE CEDEX

ABSTRACT

Our ambition is to use on articulatory model of the vocal tract as a tool of description and specification of vocalic systems. We have choosen the model proposed by S. MAEDA (1) : its has been established taking into account a priori articulatory knowledge. To explore the model 371293 configurations have been generated . Here we analyze the global and phonetic aspects in terms of articulatory-acoustic relations.

INTRODUCTION.

Bon nombre de problèmes concernant les systèmes vocaliques restent en suspens et certaines données nécessaires à leur analyse demeurent difficiles à obtenir. L'utilisation d'un modèle articulatoire de production comme instrument de prédiction peut prolonger l'analyse qui a été à l'origine de son élaboration. Mais pour espérer, de cette façon, valider des hypothèses, il faut préalablement s'assurer du bon comportement du modèle adopté.

Dans cette étude, nous avons constitué un "dictionnaire" associant paramètres articulatoires de commande et caractéristiques acoustiques résultantes, en vue d'une étude systématique globale et par classes phonétiques vocaliques.

Dans l'état actuel des connaissances, tous les problèmes du passage de l'articulatoire à l'acoustique ne sont pas encore résolus. On peut, par exemple, citer la simulation de l'effet de radiation aux lèvres (mais où donc s'arrête le conduit vocal ?), l'estimation du passage de la coupe sagittale à la fonction d'aire, l'appréciation fine des pertes tout au long du tractus. Mais c'est "l'inversion du conduit vocal", c'est à dire le retour de l'acoustique à l'articulatoire, qui suscite le plus de questions, et qui, par cela même, pourra se révéler le plus riche en interprétations. C'est en effet la clé de l'étude de tous les phénomènes de compensation : il peut exister de nombreuses configurations articulatoires associées à un même point de l'espace acoustique. Et l'on sait, par ailleurs, que les phénomènes de coarticulation engagent, dans leur processus, de nécessaires compensations acquises au cours de l'apprentissage de la langue (cf. les expériences sur les bite blocks).

II. LE MODELE DU CONDUIT VOCAL.

Nous avons choisi le modèle articulatoire de MAEDA (1), qui offre l'avantage d'être plus que fonctionnel puisque physique. Il repose sur une analyse en composantes linéaires, guidée par une connaissance a priori du comportement des articulateurs (2)). La coupe sagittale est ainsi générée à partir de cinq paramètres associés à la mâchoire (Ma), à l'apex (L3), au dos (L2) et au corps (L1) de la langue, et aux lèvres (Le). Ces paramètres sont normalisés par rapport à leur moyenne (m) et à leur écart-type (σ); leur croissance correspond à une diminution de la section du conduit vocal.

A sa version de base nous avons apporté quelques modifications :

* Articulatoires :

- Nous avons légèrement modifié la position du larynx pour disposer d'une plus grande dynamique pour le formant F2 des voyelles antérieures fermées.
- Les coefficients de passage de la coupe sagittale à la fonction d'aire ont été remplacés par ceux obtenus par moulage (3). Il s'agit d'une amélioration qui n'est cependant pas encore totalement satisfaisante : la morphologie du conduit vocal est certes respectée (division en cinq zones anatomiques), mais les coefficients appliqués sont constants dans chacune des régions, quelle que soit la dimension sagittale, ce qui induit peut être des erreurs pour les faibles sections.

* Acoustiques :

- A la modélisation de KELLY & LOCHBAUM (4), nous avons préféré celle de FLANAGAN (5), qui permet une meilleure prise en compte des différentes pertes (le logiciel utilisé a été développé par DEGRYSE et adapté par SANCHEZ).

III. LA GENERATION DU DICTIONNAIRE.

Nous avons choisi de décrire l'espace acoustique à partir d'un maillage relativement serré du domaine de commande articulatoire : treize valeurs pour chaque paramètre entre $m \pm 3\sigma$, soit 371293 combinaisons (A six secondes sur un LSI 11/73, cela représente un certain nombre d'heures de calcul!!!). Rappelons que ATAL (6), dans son étude fondamentale, avait généré 30720

configurations et que MAEDA (7) en avait simulées 16151.

Le dictionnaire ainsi constitué renferme pour chaque entrée :

- les valeurs des cinq paramètres de commande
- les fréquences, les bandes passantes et les amplitudes relatives des cinq premiers formants.

Une élimination des configurations occlusives, une limitation de la section minimale à 1 mm^2 et une série de tests (F1>180 Hz, 1.84F1+F2< 3285 Hz) ont permis de ne retenir que les réalisations vocaliques (soit 120317 selectionnées)

IV. RESULTATS.

1. Résultats globaux.

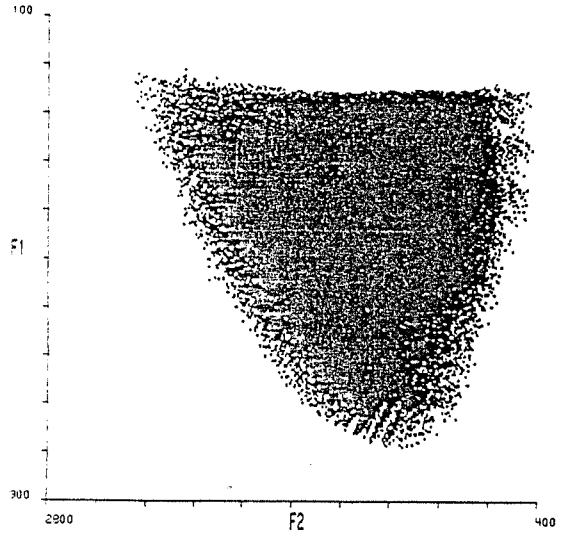


Figure 1.a : L'espace généré; Plan F1/F2

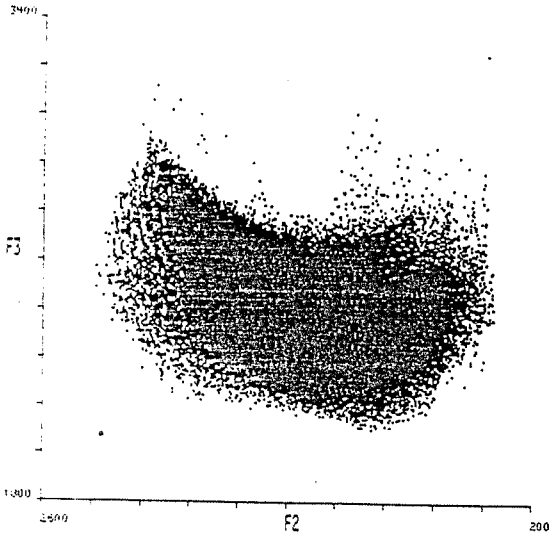


Figure 1.b : L'espace généré; Plan F2/F3

L' espace vocalique maximum apparaît comme parfaitement décrit dans le plan F1/F2 (Figure 1a) et sa représentation dans F2/F3 est conforme à ce que l'on pouvait prévoir à partir des valeurs moyennes généralement citées (Figure 1b).
En utilisant une technique de grisé (huit niveaux obtenus à partir des caractères spéciaux de FENG), nous mettons en évidence les densités dans le plan F1/F2, par pavés de 16 Hz/48 Hz. On peut immédiatement observer :

- une zone de forte densité correspondant aux voyelles antérieures [e,ɛ,a] ;
- de faibles densités pour les voyelles postérieures [ɔ,o,u].

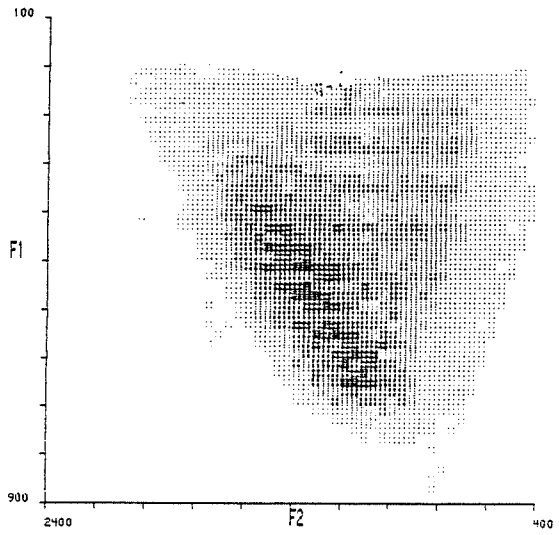


Figure 2 : Les densités dans le plan F1/F2.

Que le [i] soit de faible densité ne nous étonnera pas, le modèle ayant des difficultés à générer des configurations avec un deuxième formant élevé. Remarquons enfin que la zone de faible densité définie par MAEDA comme la zone nasale ne se retrouve pas. Notre plus grand nombre de simulations suffit-il à expliquer cette absence ? (encore qu'ATAL avec un plus faible nombre ne l'ait pas non plus générée..)

2. Sous espaces vocaliques (acoustiques)

Les besoins de l'analyse, les exigences de la synthèse de qualité, la mise en oeuvre de modèles perceptifs (8), ont mis en évidence l'importance des cinq premiers formants.

	n	F1	F2	F3	F4	F5
i	764	250	2200	2110-4050	2880-4390	3850-5190
e	1533	350	2000	2165-2990	2875-4300	3920-5200
ɛ	6922	510	1900	2140-2830	2660-4280	3675-5040
a	22085	725	1250	2120-2720	3630-4246	4000-5030
ɔ	22674	510	1100	2040-2900	2520-4720	3760-5150
o	5401	350	870	2050-3060	2760-4690	3740-5070
u	2914	250	750	2050-3200	3580-4270	3610-5045
y	1230	250	1800	2130-2870	3150-4430	3880-5195
ø	16323	350	1600	2090-3000	3080-4430	3650-5080
œ	19478	510	1400	2090-2805	3080-4280	3680-5095

A partir de plausibles ellipses de dispersion (m,σ,R) (9,10) choisies dans le plan F1/F2, nous avons sélectionné pour dix voyelles du français les candidats de notre dictionnaire. Le tableau ci-dessus présente les bornes de variation des formants supérieurs : les valeurs des formants sont exprimées en Hz, et n indique le nombre de configurations contenues dans chaque ellipse.

3. Sous espaces vocaliques. (acoustiques et articulatoires)

Nous avons porté une attention particulière aux voyelles cardinales [i,a,u] pour lesquelles le sous espace articulatoire de commande a été représenté. Nous avons tracé les dix plans articulatoires correspondants, soit Ma, Le, L1, L2, L3, pris deux à deux.

Le modèle permet de retrouver des données de production bien connues ou déjà établies :

- impossibilité de prononcer un [i] avec une mâchoire très ouverte;
- pas de [a] avec des lèvres fermées (Figure.3);
- pas de [u] avec des lèvres ou une mâchoire très ouvertes.

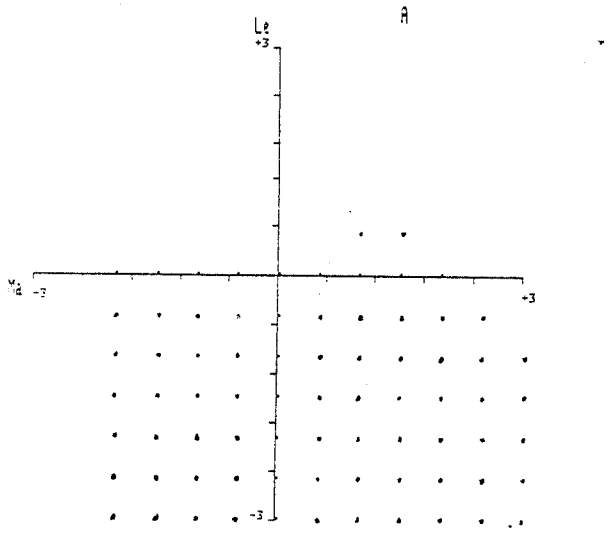


Figure 3 : Plan articulatoire de production Ma/Le pour le [a] .

Ces interdictions sont largement contrebalançées par d'énormes possibilités compensatoires, par exemple.:

- très grande latitude de la langue pour [a] et [u] , du dos et de l'apex de la langue pour [i] .

Une première analyse permet de mettre en évidence les compensations suivantes :

- abaissement de la mâchoire compensé par une avancée de la langue pour [i] ;
- recul de la langue compensé par un aplatissement pour [i] (Figure 4);
- fermeture de la mâchoire compensée par un recul de la langue pour [a] .

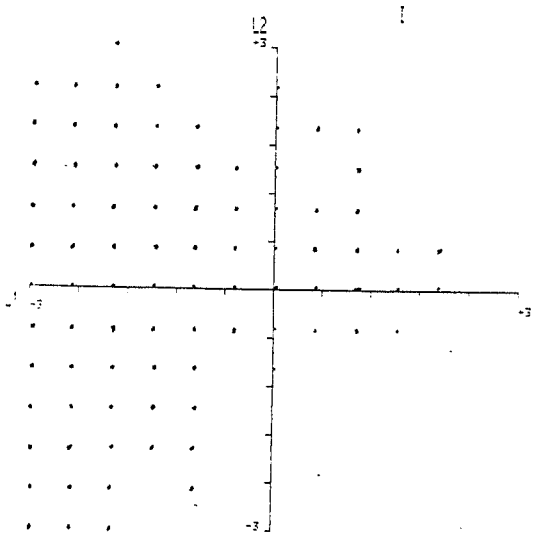


Figure 4 : Plan articulatoire de production L1/L2 pour le [i] .

V. PERSPECTIVES.

Si certains points peuvent être améliorés (problème du F2 pour le [i]), introduction d'un deuxième paramètre pour les lèvres (l1), pondération du coefficient de forme avec la dimension sagittale, le bon comportement du modèle nous permet d'envisager des études fondamentales. Il nous semble indispensable de pouvoir disposer de fonctions de sensibilité associées aux paramètres articulatoires pour les configurations vocaliques du français et de préciser en détail les possibilités compensatoires reliées à des phénomènes précis de coarticulation.

Remerciements

Nous remercions Shinji MAEDA, Christian ABRY et Gang FENG pour leurs critiques et suggestions.

REFERENCES.

(1) S. Maeda ,"An Articulatory Model Of the Tongue Based on a Statistical Analysis", J. Acoust. Soc. Am. 65 S1, S22(A), 1979.

(2) B.E.F. Lindblom & J.E.F. Sundberg, "Acoustical Consequences of Lip, Tongue, Jaw and Larynx Movements", J. Acoust. Soc. Am. 50, 1166-1179, 1972.

(3) H. Sanchez & L. J. Boë, "De la coupe sagittale à la fonction d'aire du conduit vocal", 13èmes J.E.P. G.C.P. du G.A.L.F.,pp. 23-25, 1984.

(4) J.L.Jr Kelly & C. Lochbaum, "Speech Synthesis", Speech Comm. Seminar 2, F7, R.I.T. Stockholm, 1962.

(5) J.L. Flanagan, "Speech Analysis Synthesis and Perception", 2nd Ed. Springer Verlag. New York, 25-69, 1972.

(6) B.S. Atal, J.J. Chang, M.V. Mathews & J.W. Tukey, "Inversion of Articulatory-to-acoustic transformation in the Vocal Tract by a Computer-Sorting Technique", J. Acoust. Soc. Am. 63, 1535-1555, 1978.
(7) S. Maeda, "Une paires de pics spectraux comme corrélat acoustique de la nasalisation des voyelles", 13èmes J.E.P. G.C.P. du G.A.L.F., 223-224,1984.
(8) A. Bladon , "Two-Formant Models of vowel Perception : Shortcomings and Enhancements, Speech Comm. 2, 305-313, 1983.
(9) G. Chollet, "Computer Signal Analysis and Variability of Acoustic Parameters in Phonetics", Ph. Diss. Computer Sci. and Linguistics, U.C.L.A., 1978.
(10) C. Abry, L.J. Boë & R. Descout, "(i,a,u) ? Pas si fou ? Ou les lèvres maximisent elles les voyelles ?", 13èmes J.E.P. G.C.P. du G.A.L.F., 205-207, 1984.
(11) C. Abry, L.J. Boë, P. Corsi, R. Descout, M. Gentil & P. Graillot, "Labialité et phonétique", Publication de l'Université des Langues et Lettres de GRENOBLE, 1980.

CODAGE VECTORIEL ET MODELISATION ARTICULATOIRE DYNAMIQUE

ROCHETTE Denis, PERRIER Pascal & BOE Louis-Jean

INSTITUT DE LA COMMUNICATION PARLEE (LA CNRS 368)
I.N.P.G. - 46, avenue Felix Viallet
38031 GRENOBLE CEDEX

ABSTRACT

MAEDA's articulary model allows one to get a dictionary of all five-parameters vocal tract shapes and corresponding acoustic outputs. We propose a strategy to choose articulatory trajectories mapping into specific variations in the acoustic space of the two first formants, those variations being obtained from cineradiographic transitions V-C-V.

I. INTRODUCTION.

Un grand nombre de modèles a été étudié dans le domaine de la production articulatoire (1),(2). Cette dernière décennie a vu le développement d'études statistiques à partir de cinéradiographies donnant la coupe sagittale du conduit vocal (3); les résultats de ces mesures permettent de réduire à quelques unités le nombre des paramètres de commande de ces modèles qualifiés d'"abstraits", par opposition aux analogues du conduit vocal qui utilisent directement une fonction d'aire discrétisée et approchée par une vingtaine de sections cylindriques (4). Une fonction plus ou moins simplifiée permet le calcul de la réponse acoustique pour une disposition donnée, c'est-à-dire de passer de l'espace articulatoire à l'espace acoustique. Ce dernier sera typiquement l'espace des trois ou quatre premiers formants, puisque ces derniers suffisent à caractériser le signal de parole au plan acoustique.

Deux problèmes sont cependant soulevés :

- Hormis les bornes de variation des paramètres, fixées par la conception même du modèle, le domaine de définition de la fonction sus-citée n'est, a priori, pas connu. En effet, à l'intérieur de ces bornes, toutes les combinaisons des paramètres articulatoires supposés indépendants ne conduisent pas nécessairement à un équivalent "réaliste" dans l'espace acoustique. D'où une corrélation a posteriori des paramètres de commande, exprimée par le domaine de définition de la fonction, ce domaine ne pouvant être évalué par une étude exhaustive.
- La fonction est surjective sur son domaine de définition. Afin de déterminer dans l'espace articulatoire l'ensemble des points associés à un même point de l'espace acoustique, ATAL (5) a proposé une méthode permettant de calculer les "fibres", sous-variétés de niveau associées à la fonction, en approximant au premier ordre ces sous-variétés par leurs espaces tangents. Néanmoins, la détermination de trajectoires articulatoires correspondant à des transitions acoustiques données consisterait à ajuster pas à pas les paramètres du modèle en fonction des paramètres acoustiques.

Dans cette étude, nous adopterons l'hypothèse optimiste selon laquelle il serait possible de définir une stratégie d'évolution des paramètres articulatoires, intrinsèque au modèle et telle que la transition associée dans le plan (F_1,F_2) soit proche des transitions "réelles" obtenues à partir de cinéradiographies (6). Ceci revient à définir :

- Un système de commande (au sens dynamique du terme) du modèle.
- Un critère dont l'optimisation déterminera la (ou les) trajectoire(s) pertinente(s) entre deux points fixés dans l'espace articulatoire.

II. LE MODELE ARTICULATOIRE DE MAEDA

Pour cette étude, nous disposons d'un "dictionnaire" établi à partir du modèle articulatoire du conduit vocal de MAEDA (3), qui associe les paramètres de commande du modèle (constituant ce que l'on appellera l'espace articulatoire de commande) aux caractéristiques acoustiques de la fonction de transfert ainsi déterminée.

Le modèle de MAEDA est le résultat d'une analyse en composantes linéaires, menée sur 400 radiographies d'un conduit vocal de femme, et guidée par des connaissances a priori (7) sur le comportement respectifs des différents articulateurs (1), ce qui permet de supposer qu'il approchera au mieux la réalité physique. Grâce au système de commande, consistant en cinq paramètres associés à la mâchoire, au corps, au dos et à l'apex de la langue, et, enfin, aux lèvres, on peut faire évoluer la coupe sagittale du conduit vocal.

Une étude approfondie des capacités de ce modèle, légèrement modifié en ce qui concerne le passage de la coupe sagittale à la fonction d'aire (8), et la modélisation acoustique (9) a été entreprise (10), grâce à une exploration fine de l'espace articulatoire de commande : chaque paramètre évolue linéairement entre ses extrêmes et peut prendre treize valeurs. Le "dictionnaire"

ainsi obtenu contient 371293 configurations, associant les valeurs des cinq paramètres de commande aux fréquences, bandes passantes et amplitudes relatives des cinq premiers formants correspondants.

III. LA STRATEGIE PROPOSEE

Comme nous l'avons dit précédemment, notre approche consiste à choisir des trajectoires dans l'espace articulatoire de commande, indépendemment de l'espace acoustique , c'est à dire sans contrôle adaptatif par référence à l'évolution simultanée des formants. Formulons donc le problème de la façon suivante :

* Soient (Flo,F2o) et (Fl_f, $F2_f$), le point de départ et le point d'arrivée de la transition considérée dans le plan (F1,F2), et soient Yo et Y_f les points correspondant dans l'espace articulatoire (remarquons au passage que le choix de Yo et de Y_f est arbitraire du fait de de la surjectivité de la fonction qui les associe aux valeurs formantiques).
* Posons : $Yo = (Yi)o$ i=1,5
  et
  $Y_f = (Yi)_f$ i=1,5
  Pour décrire une trajectoire reliant Yo à Y_f, il faut appliquer à chaque point intermédiaire Y(t) une commande désignée par :
  $U(t) = (Ui(t))$ i=1,5 avec $|Ui(t)| \leq Mi$
  les bornes Mi exprimant des contraintes d'inertie sur les cinq paramètres (essentiellement dérivées de mesures de vitesses d'articulateurs (11), (12))

L'équation d'évolution s'écrit alors :

$$Y(t) = Y(t-1)+U(t)$$

Pour l'instant, nous ne souhaitons faire aucune hypothèse a priori sur le modèle, ce qui nous amène à proposer deux critères d'optimisation pour le choix de la trajectoire (Yo, Y_f).

Le premier consiste à choisir la commande U à chaque instant telle que la distance entre Y(t) et Y_f soit minimale. En d'autres termes cela revient à minimiser $Jl = (d(Y(t),Y_f))^2$ sous la contrainte $|Ui(t)| \leq Mi$, i=1,5.

Pour le deuxième on cherchera à minimiser l'écart entre la trajectoire produite et une trajectoire que l'on qualifiera de "souhaitée", par exemple comprise dans un "couloir", entre Yo et Y_f. Nous poserons donc :

$$J2 = \frac{\sum (Ui+Yi).((Yi)_f-(Yi)o)}{(Ui+Yi)^2}$$

Le problème revient à maximiser J2 sous les contraintes Mi. Contrairement à Jl, J2 ne vise pas l'obtention d'une trajectoire décrite en un temps minimum.

PERSPECTIVES ET EVOLUTIONS

Par la suite nous serons amenés éventuellement à considérer des critères plus restrictifs, représentatifs du poids respectif des paramètres dans le processus de l'articulation, et des énergies potentielles et cinétiques liées à la position de l'articulateur.

Les résultats porteront sur les transitions maximales entre les trois voyelles cardinales (i,a,u), insérées dans des logatomes de type VCV.

REFERENCES

(1) B.E.F. Lindblom & J.E.F. Sundberg, "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement", J. Acoust. Soc. Am. 50 4, 1166-1179, 1971.

(2) P. Mermelstein, "Articulatory Model for the Study of Speech Production", J. Acoust. Soc. Am. 53 4, 1070-1082, 1973.

(3) S. Maeda, "An Articulatory Model of the Tongue Based on a Statistical Analysis", J. Acoust. Soc. Am. 65 S1, S22(A), 1979.

(4) M. MRAYATI, "Contribution aux Etudes sur la Production de la Parole. Modéles électriques du Conduit Vocal avec Pertes, du Conduit Nasal et de la Source Vocale. Etude de leurs Interactions. Relation entre Dispositions Articulatoires et Caractéristiques Acoustiques.", Thèse d'Etat, I.N.P. Grenoble, 1978.

(5) B.S. Atal, J.J. Chang, M.V. Mathews & J.W. Tukey, "Inversion of Articulatory-to-acoustic transformation in the Vocal Tract, by Computer-sorting Technique", J. Acoust. Soc. Am. 63, 1535-1555, 1978.

(6) H. Sanchez, T. Gay & L.J. Boë, "Maximal Transitions (VCV) from Radiography Data to Formantic Trajectories", Sem. Franco-Suédois, Grenoble, 1985.

(7) J.E. Overall, "Orthogonal Factors and Uncorrelated Factors Scores", Psychological Report 10, 651-662,1962.

(8) H. Sanchez & L.J. Boë, "De la coupe sagittale à la fonction d'aire du conduit vocal", 13èmes J.E.P. G.C.P. du G.A.L.F., 23-25, 1984

(9) J.L. Flanagan, "Speech Analysis Synthesis and Perception", 2nd Ed. Springer Verlag. New York, 1972

(10) P. Perrier, L.J. Boë, R. Majid Shihab & B. Guérin, "Modélisation articulatoire du conduit vocal. Exploration et Exploitation",14èmes J.E.P. G.C.P. du G.A.L.F., 1985.

(11) D.P. Kuehn & K.L. Moll, "A Cineradiographic Study of VC and CV ariculatory velocities", J. Phonetics 4 4,1976.

(12) S. Masaki, K. Shirai, H. Imagawa & S. Kiritani, "Jaw Opening and the Time Constant of Jaw Movements in the Production of Sequences of Stationary Vowels and Vowel Sequence Words", Japan. J. Logopedics Phoniatrics, 18, 1984.

SYSTEME D'EXPLOITATION AUTOMATIQUE DE LABIO-FILMS

J.P. GOURRET J. PAILLÉ L.J. BOÉ R.DESCOUT

LTSN/ENSP Marseille I. de Phon. Grenoble CNET Lannion

Résumé- La modélisation articulatoire de la parole nécessite l'analyse d'un grand nombre de données physiques et physiologiques recueillies sur l'être humain. Les informations relatives aux positions des lèvres proviennent de l'enregistrement synchrone du signal acoustique et du mouvement des lèvres, vues de face et de profil, par cinéma rapide.

Le système que nous avons réalisé permet d'extraire automatiquement les paramètres dimensionnels des lèvres à partir des images d'un labiofilm. Il comprend un banc de défilement pour film 35 mm, une caméra noir et blanc associée à un numériseur d'images 512x512 points codés sur 8 bits, et un miniordinateur PDP 11/10.

L'exploitation du film s'effectue en deux étapes. On entre tout d'abord la liste des images à traiter, puis dans une deuxième étape entièrement automatique, chaque image est positionnée sous la caméra, numérisée et traitée. Pour chaque image les contours représentatifs des lèvres inférieure et supérieure vues de face et de profil sont recherchés et approximés par un polynôme. Après divers tests de fiabilité, les coefficients des polynômes ainsi que les paramètres géométriques des lèvres sont stockés dans un fichier. Ce fichier permet ultérieurement de restituer l'allure des contours ou encore de tracer l'évolution des divers paramètres.

INTRODUCTION

La modélisation articulatoire de la parole nécessite l'analyse d'un grand nombre de données physiques et physiologiques recueillies sur l'être humain. La plupart des informations concernant l'intérieur du conduit vocal sont issues de mesures cinéradiographiques, celles relatives aux commandes des différents muscles sont obtenues à partir d'enregistrements électromyographiques, et celles relatives aux positions des lèvres proviennent de l'enregistrement synchrone du signal acoustique et du mouvement des lèvres par cinéma rapide (50 images par seconde). Les enregistrements cinématographiques obtenus dans ce dernier cas sont appelés des labiofilms et comportent généralement entre 3000 et 9000 images (soit 1 à 3 minutes d'enregistrement).

Le problème de la détection du contour des lèvres et d'extraction des paramètres dimensionnels à partir des images d'un labiofilm a été abordé par LOUVION (4). Il a proposé de numériser chaque image du labiofilm et d'utiliser des algorithmes numériques pour détecter le contour des lèvres et en extraire les paramètres représentatifs. Ces traitements, effectués sur quelques images, donnent des résultats de bonne qualité et nous ont permis d'envisager la conception d'un système entièrement automatique.

C'est ce système que nous proposons ici. Compte tenu de l'acquit en matériel spécifique (numériseur d'images) et en logiciel (logiciel interactif de traitement d'images) du LTSN nous avons réalisé un système qui permet de dépouiller, en un temps raisonnable, les quelques milliers d'images qui constituent un film. Ce projet s'insère donc tout naturellement dans le projet plus vaste de synthèse articulatoire de la parole.

DESCRIPTION DU MATERIEL

On a représenté Figure 1 le schéma synoptique du système réalisé.

Il comprend un numériseur d'images, un système informatique et un système mécanique d'entraînement du film.

Le signal vidéo issu de la caméra est échantillonné par le numériseur et envoyé par Accès Direct Mémoire à l'ordinateur. Par construction (2) le numériseur permet d'acquérir des images 512x512 points. Chaque point est codé sur un octet (256 niveaux de gris).

L'avancement du film doit être controlé avec une très grande précision. Pour résoudre le problème de la bobine débitrice et de la fragilité du film nous avons utilisé trois moteurs et des capteurs de position comme indiqué Figure 1. Un premier moteur pas à pas couplé directement sur une roue dentée jouant le rôle de cabestan entraine effectivement le film avec précision. Deux autres moteurs couplés respectivement avec les bobines débitrices et réceptrices assurent une tension convenable du film. Coté roue débitrice, si la tension est trop forte le moteur est actionné pour dérouler un peu de film. Coté roue réceptrice le moteur est actionné pour assurer l'enroulement tant que la tension n'est pas trop élevée.

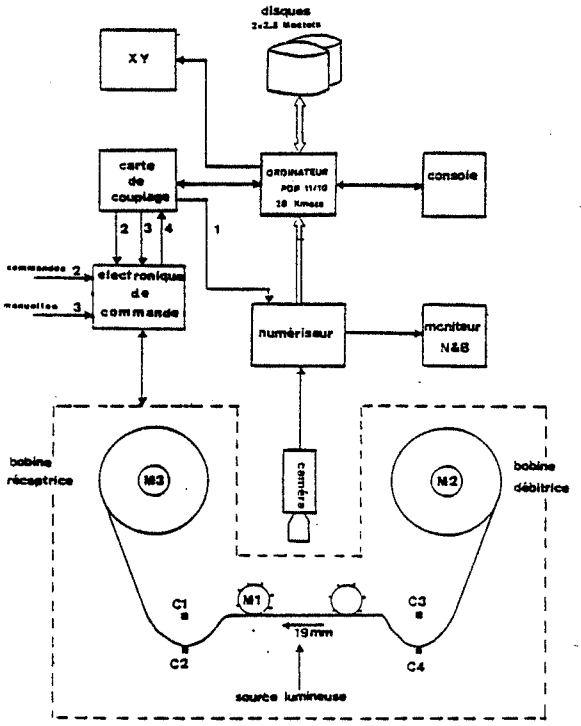


Figure 1 . 1 commande numériseur
2 commande avance d'une image
3 commande du sens de déplacement
4 signal d'arrêt sur image

L'avancement normal du film se fait par translation de 19 mm correspondant à la largeur d'une image. La commande des différents moteurs est assurée par l'électronique de commande qui reçoit des informations des capteurs de tension du film d'une part et du numériseur et de l'ensemble informatique d'autre part. Elle comprend la commande du moteur pas à pas M1, la commande des moteurs associés aux bobines débitrice et réceptrice, M2 et M3, et la logique de changement de sens.

La liaison avec le système informatique est assurée par une carte de couplage (type DR11C par exemple) permettant le mode interruptionnel. Le fonctionnement est parfaitement symétrique. Il est possible de revenir en arrière pour une nouvelle numérisation d'une ou plusieurs images si le besoin s'en fait sentir. Ce retour peut être effectué manuellement ou commandé par l'ordinateur.

TRAITEMENT DES SIGNAUX NUMERIQUES BIDIMENSIONNELS

Pour ce qui concerne la partie traitement des signaux numériques bidimensionnels, le principal problème consiste à choisir et à mettre en oeuvre des algorithmes rapides de segmentation et de recherche de contours afin de diminuer au mieux le temps de traitement. Les traitements ont été réalisés sur des images similaires à celle de la Figure 2.

Figure 2 .

Chaque image comporte une vue de face et une vue de profil obtenues par un jeu de miroir et un premier travail consiste à la morceler en deux sous images de 128x128 points chacune, l'une correspondant à la vue de face et l'autre à la vue de profil (Figure 3).

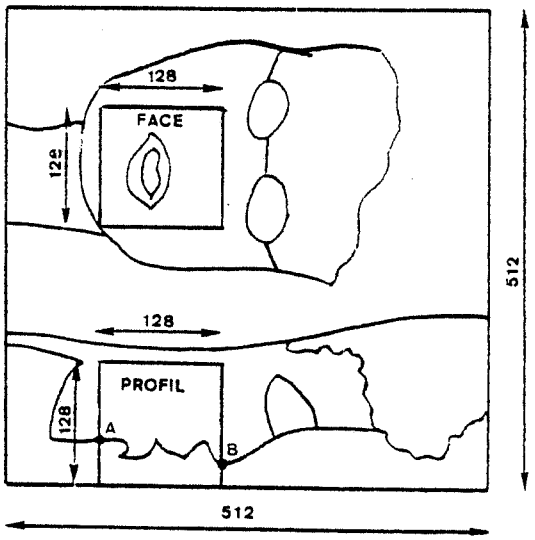


Figure 3 .

Chacune des sous images est ensuite codée sur deux niveaux de gris après détermination d'un seuil S. Une deuxième étape consiste alors à traiter différemment les images de face et de profil pour extraire les contours qui nous interessent.

Pour l'image de face, après avoir répertorié tous les contours dans la sous image au moyen d'un algorithme classique, de nombreux tests sont mis en oeuvre pour obtenir le contour relatif à la

bouche. Ces tests sont destinés à éviter l'accrochage sur un autre contour que celui représentatif de la bouche, suite à un déplacement de la personne filmée ou à des ombres portées. Ils prennent en compte la position des contours trouvés et/ou leur périmètre.

Pour l'image de profil, il suffit de décrire la frontière A,B entre les lèvres et le fond pour répertorier tous les points du contour profil.

Les contours face et profil sont approximés par quatre polynômes déterminés par le critère des moindres carrés (Figure 4).

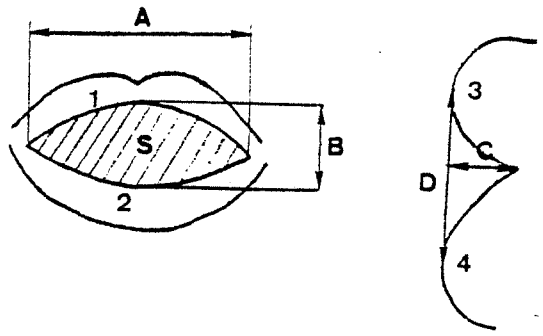


Figure 4 .

Cette étape consiste à rechercher les commissures des lèvres puis à approximer les contours face 1 et 2 par deux polynômes de degré 4, approximer les contours profil 3 et 4 par deux polynômes de degré 3 et à calculer les paramètres A,B,C,D et S.

Les 19 paramètres sont stockés sur disque dans un fichier au format RT-11. Ce fichier permet ultérieurement de restituer l'allure des contours ou encore de tracer l'évolution des paramètres.

Malgré le grand nombre de précautions prises lors de la prise de vues et du traitement il est possible que les coefficients relatifs à certaines images soient erronnés. Ces erreurs proviennent généralement de la présence des dents du sujet filmé (Figure 5).

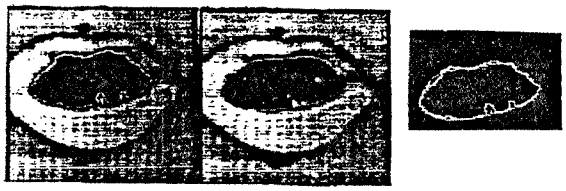


Figure 5 .

Pour éviter ce genre de problème il est possible, après seuillage de l'image face, de rajouter une étape de dilatation-érosion au moyen d'un élément structurant.

UTILISATION DU SYSTEME

Le logiciel de gestion du banc d'acquisition, de traitement des sous images et de visualisation des courbes a été inséré dans le système GIMAGE (2). Ce logiciel utilise l'ordinateur en mode conversationnel et est fortement orienté vers l'utilisateur non spécialiste de la programmation. Il fait appel à une soixantaine de commandes pour l'acquisition, la visualisation, le traitement ou la manipulation d'images numériques.

Trois nouvelles commandes ont été rajoutées pour l'acquisition automatique du film :

- une commande pour créer le fichier de stockage des paramètres et entrer le numéro des images à acquérir,
- une commande pour lister les numéros des images acquises et/ou à acquérir ainsi que leurs paramètres,
- une commande pour lancer l'acquisition et le traitement des images. Cette commande peut être stoppée à n'importe quel moment et reprise ultérieurement. Il est également possible de reprendre certaines acquisitions lorsque les tests de fiabilité ont éliminé les images erronnées.

A titre d'exemple on donne dans la figure 6 l'allure des paramètres A, B, C, D et S pour la transition $\int u$.

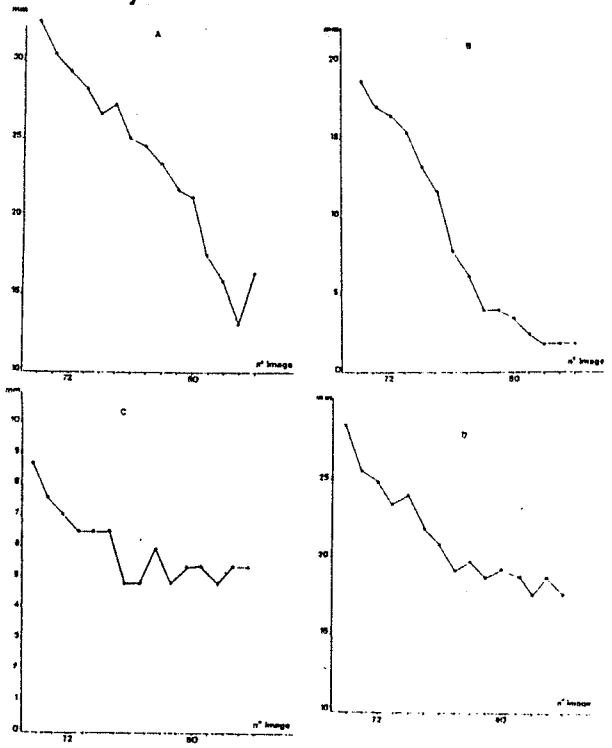


Figure 6 .

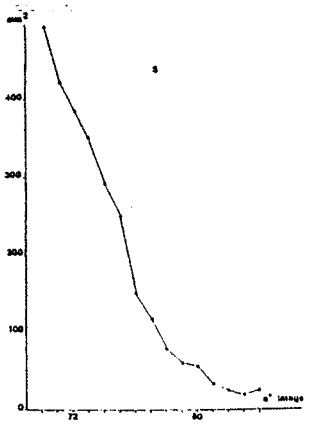


Figure 6 (suite) .

CONCLUSION

Les perspectives offertes par la synthèse articulatoire qui prend en compte les contraintes de production des organes de l'articulation vont necessiter l'acquisition de nombreuses données physiologiques (3). L'utilisation de techniques automatiques de dépouillement va permettre une exploitation qui jusque là était limitée par le coût en temps.

Le système que nous avons réalisé permet d'extraire automatiquement les paramètres dimensionnels des lèvres à partir des images d'un labio-film. Il permet de dépouiller en un temps raisonnable (une minute par image) les quelques milliers d'images qui constituent le film.

BIBLIOGRAPHIE

(1) GARRIC PH. Numériseur d'images. DEA Signaux et Systèmes spatio-temporels. Laboratoire de Traitement du Signal Numérique. 1980

(2) GOURRET J.P., PAILLE J. Système d'exploitation pour l'acquisition, le traitement et la visualisation des images numériques. TSI vol.2,n°1,1983

(3) HESSELMAN N.L. Structure analysis of leep contours for isolated spoken vowels using Fourier descriptors. Speech communication, vol.2, pp 327-340, 1983

(4) LOUVION J.R. Détection du contour de lèvres et extraction de paramètres constitutifs. Application à l'analyse automatique de labio-films, Thèse de 3 ème cycle, Université de Rennes 1, A-620-92, dec. 1980.

****
* ANALYSE *
****

ANALYSE A TRES COURT TERME DE LA PAROLE
Un outil et quelques directions de recherche

LIENARD Jean-Sylvain (*)

LIMSI - CNRS BP 30 91406 ORSAY Cedex France
et AT&T BELL LABS MURRAY HILL NJ 07974 USA

ABSTRACT

In the present paper we will describe a non-causal, adjustable filterbank, that allows for reconstruction of the input signal through the mere summation of the outputs. We will then present the principles and some results of a representation of the speech signal by a set of discrete events. We will look also for a finer decomposition into elementary waveforms that could be thought of as the impulse response associated to each spectral maximum. This will give us the opportunity to study synchronisation phenomena between output signals of adjacent channels. Finally we will discuss the relations of our ideas and some contemporary auditory models.

INTRODUCTION

Depuis très longtemps, l'une des idées de base de l'analyse acoustique de la parole est la séparation fonctionnelle entre la source et le système de transfert. Pour évaluer le spectre à court terme on considère le signal comme stationnaire dans un intervalle de 20 à 50 ms, incorporant plusieurs périodes de vibration des cordes vocales. Il est évident que cette approche n'est pas adaptée à l'analyse des évolutions spectrales rapides qui se produisent dans certaines consonnes (nasales, occlusives) ou semi-voyelles, dont la représentation est ainsi fortement atténuée alors même que leur pertinence perceptive ne peut être mise en doute.

Nous esquissons ici une méthode d'analyse qui évite de faire d'emblée une distinction entre le spectre à court terme et sa modulation par le pitch. Cette méthode doit aussi fournir une description de l'onde acoustique qui soit interprétable en termes de production ou de perception et, autant que possible, permettre une reconstruction du signal qui soit indiscernable de l'original.

(*) Une partie importante de ce travail a été effectuée pendant l'été 84, lors d'un stage dans le Département de Recherches Acoustiques des laboratoires AT&T BELL, que nous remercions ici.

LE BANC DE FILTRES

Le banc de filtres a été conçu avec le souci de simplicité, de respect des rapports de temps entre les canaux, et de reconstruction de la forme précise du signal analysé. La simplicité du filtrage fait que les filtres se comportent de manière aussi neutre que possible en régime transitoire, et entraîne des calculs plus rapides. Comme nous sommes intéressés par les rapports de temps des divers composants du signal, il est essentiel que tous les filtres réagissent avec le même retard, de manière à n'introduire aucune distorsion temporelle entre canaux. Ceci peut être obtenu avec des filtres à phase linéaire (filtres FIR symétriques, non-récursifs) ou avec des filtres non-causaux, en utilisant la propriété de réversion du temps. Dans ce dernier cas, on peut utiliser des filtres IIR, qui entraînent des temps de calcul beaucoup plus faibles.

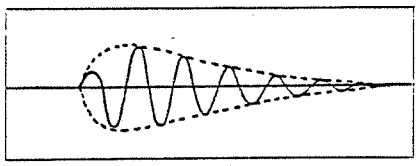
Nous avons donc choisi de mettre en oeuvre de simples resonateurs, comportant des zéros placés de telle sorte que le gain à la résonance soit indépendant de la fréquence centrale. En appliquant ce filtre deux fois de suite au même signal, mais en inversant son sens de défilement lors de la seconde passe, on obtient une fonction de transfert globale réelle, ce qui signifie que le spectre de phase est nul en tout point. Les gains individuels sont ajustés de manière que la fonction de transfert composite (somme des fonctions de transfert individuelles) soit aussi proche que possible de l'unité, tout au long de l'intervalle de fréquence couvert par le banc. Ainsi le signal initial peut être reconstruit au moyen d'une simple sommation des sorties. Les fréquences centrales des canaux sont répartis soit linéairement, soit selon une échelle de BARK.

La fig 1 montre une sinusoïde wobulée, analysée selon l'échelle de BARK. Le signal initial est reporté au dessous des sorties des 16 canaux. Au dessous, on trouve le signal reconstitué par sommation. Il est clair que celui-ci est très proche de l'original, et cette propriété est vérifiée quelle que soit la forme du signal analysé, pourvu que sa bande passante soit inférieure à celle couverte par le banc.

d'énergie dans diverses zones spectrales ne se produisent pas exactement au même instant. Notre schème de mesure sera le suivant: entre deux minimums successifs de la FC, on cherche dans chaque canal la valeur du maximum d'énergie. L'ensemble de valeurs ainsi obtenues est alors affecté à l'instant de l'événement et en constitue la représentation spectrale. Cette représentation est compacte, surtout si l'on ne retient que les maximums fréquentiels de chaque événement (fig 4), et elle conserve les éléments caractérisant le voisement, le pitch, les impulsions isolées. Le débit d'information nécessaire pour coder la parole à ce niveau devrait être de l'ordre de 50 bits par événement, avec 100 à 200 événements par seconde, ce qui placerait cette approche dans la gamme des systèmes de transmission à 10 kbauds. On notera les différences entre cette approche et les méthodes d'analyse "synchrone au pitch", qui requièrent la détection préalable du pitch et reviennent à une analyse à intervalles fixes en dehors des segments voisés.

DECOMPOSITION EN FORMES D'ONDE ELEMENTAIRES

Nous souhaitons maintenant examiner de plus près les problèmes posés par les structures temporelles internes à l'événement. Pour cela nous allons considérer le signal comme formé de "grains" spectro-temporels, ou formes d'ondes élémentaires, à décroissance exponentielle, qui peuvent occasionnellement apparaître de manière quasi- synchrone et former des événements. Nous n'émettons pas d'hypothèse stricte sur la forme de l'attaque, mais il est probablement peu réaliste de la considérer comme infiniment abrupte (fig 5).

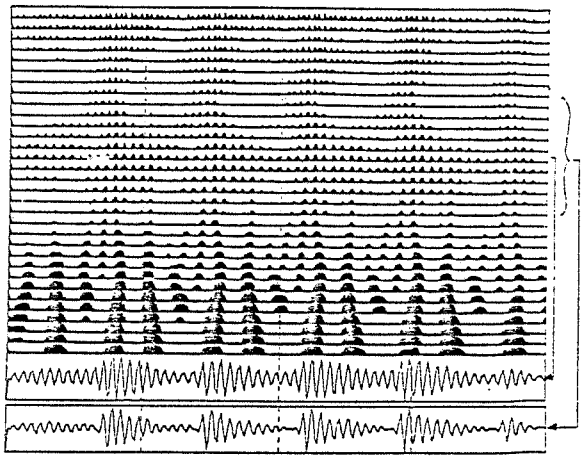


- fig 5 -

Ces "Formes d'Onde Elémentaires" FOE sont très proches conceptuellement des "Formes d'Ondes Formantiques" utilisées avec succès par X.RODET pour la synthèse de voix chantée [6]. Notre hypothèse est que l'oreille n'est sensible qu'à un très petit nombre de paramètres caractérisant les FOE, par exemple l'énergie, le calage temporel, la fréquence porteuse et peut-être un facteur de forme.

Notre banc de filtre peut être utilisé pour observer les FOE. Comme le signal peut être reconstruit par sommation de tous les canaux, nous pouvons décomposer l'ensemble en sous-groupes de canaux adjacents, en faisant en sorte que chaque sous-groupe corresponde à une seule composante du signal. La somme des canaux d'un sous-groupe donnera alors la FOE cherchée, et la somme des FOE redonnera le signal lui-même.

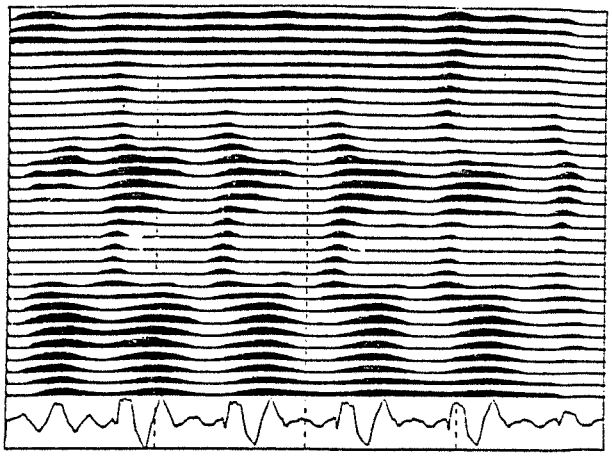
La fig 6 détaille la reconstruction des FOE correspondant au second formant d'un segment voisé.Le signal du canal 19, centré sur F2, a été



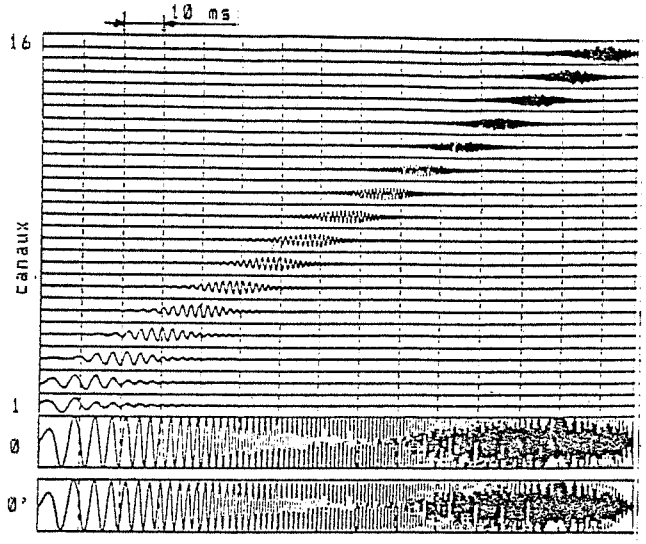
- fig 6 -

placé dans le canal 0. Au dessous, en 0', nous avons placé la somme des canaux 14 à 24, qui donnent tous une image plus ou moins atténuée et déformée temporellement de ce formant. La comparaison des signaux placés en 0 et 0' montre que la dissymétrie de l'enveloppe a été fortement amplifiée, et que la décroissance de la forme d'onde en 0' est maintenant très proche d'une exponentielle. L'instant où se produit la relance de l'excitation à chaque période de pitch est nettement marqué par une discontinuité. Les formes d'onde observables en 0' sont largement indépendantes du banc de filtres, et revèlent un phénomène (ici résonance formantique) intrinsèque au signal.

Pour repérer les groupements de canaux et leur durée, nous utiliserons une propriété de notre banc de filtres : lorsque plusieurs filtres répondent en phase, c'est qu'ils sont excités par une même composante fréquentielle du signal. L'évaluation de la simultanéité des canaux adjacents est faite en comptant 1 s'ils ont la même polarité et 0 dans le cas contraire, et en intégrant sur un court intervalle de temps la fonction ainsi définie, associée à chaque canal. La fonction résultante, définie et positive en chaque point du plan t-F, conduit à une représentation qui ressemble beaucoup à un spectrogramme (fig 7), mais qui a une tout autre



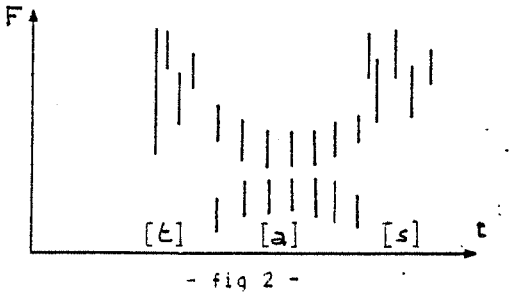
- fig 7 -



- fig 1 -

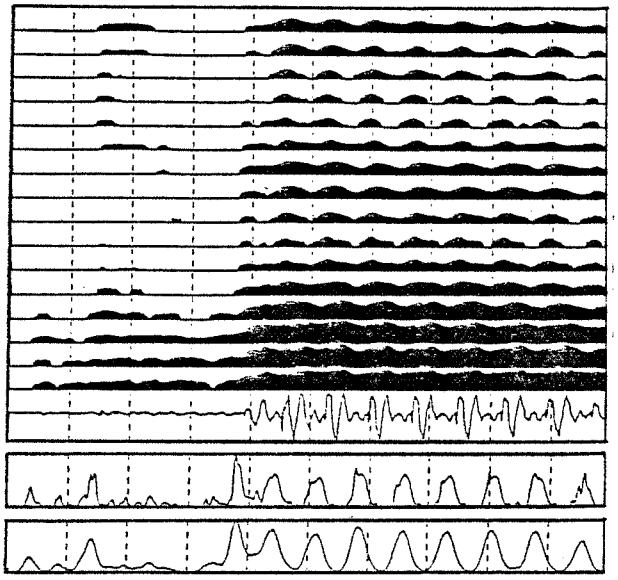
DECOMPOSITION EN EVENEMENTS

A l'examen d'un spectrogramme en bande large on peut voir des stries verticales dans les segments voisés, de même que dans les segments fricatifs - mais dans ce dernier cas avec une texture différente, essentiellement non-périodique. La situation est représentée schématiquement sur la fig 2, où l'on distingue trois types d'évenements selon leurs relations mutuelles : ceux qui sont répétés à intervalles réguliers, avec un spectre lentement variable, déterminent les qualités subjectives de voisement et de pitch; ceux qui apparaissent irrégulièrement ou changent brusquement de spectre se trouvent dans les segments fricatifs; et ceux qui sont isolés se trouvent habituellement à l'attaque de certaines consonnes.



- fig 2 -

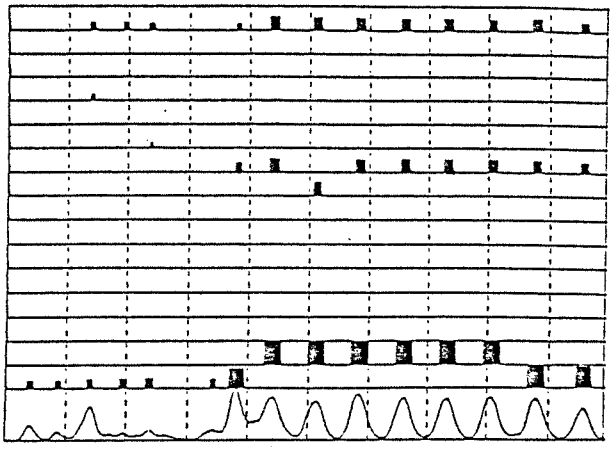
Pour localiser les événements nous avons défini une fonction, nommée fonction de cohérence FC [3, 5], de la manière suivante. En premier lieu le signal de sortie de chaque filtre est redressé et filtré passe-bas à 400 Hz, puis transformé selon une échelle quasi-logarithmique. Cette enveloppe est ensuite dérivée, et seuillée (on n'en garde que la partie positive, représentant la vitesse de l'accroissement de niveau dans le canal considéré). La FC est obtenue en sommant à chaque instant les fonctions ainsi disponibles dans les canaux. Positive par construction, elle possède habituellement un pic bien marqué au début de chaque événement. Comme le processus de dérivation est générateur de bruit, il est nécessaire d'appliquer un filtrage passe-bas pour lisser la fonction et la rendre utilisable pour repérer les événements (une fréquence de coupure de 400 Hz est convenable). La fig 3 montre un segment de parole analysé en 16 canaux (échelle linéaire, 0-3300 Hz, largeur 300 Hz) et sa FC, avant et après lissage. On constate aisément que les maximums de la FC indiquent le début des divers types d'événements que nous cherchons à caractériser : explosion du [t], pics d'énergie dans le bruit d'aspiration, et impulsions glottiques.



- fig 3 -

La fonction de cohérence pourrait être utilisée pour mesurer le pitch, même lorsque le fondamental n'est pas présent dans le signal (cas du téléphone). Cependant cet aspect de l'analyse n'était pas notre but principal et nous l'avons provisoirement laissé de côté.

Il s'agit maintenant d'associer une représentation spectrale à chaque événement. La difficulté vient du fait que les concentrations



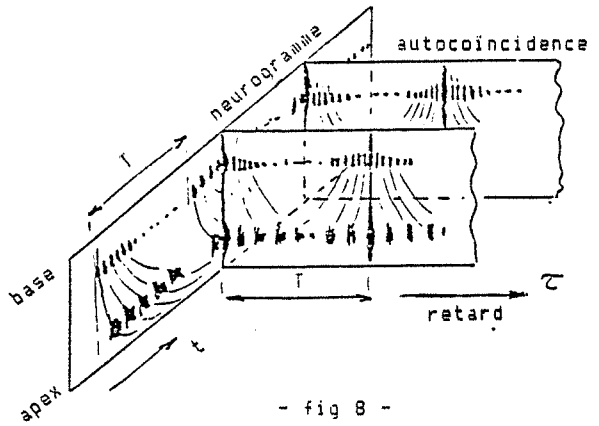
- fig 4 -

signification : elle indique, indépendamment de l'amplitude spectrale, le degré de ressemblance temporelle d'un canal avec ses voisins, ceci à chaque instant. Nous avons pu constater que les maximums fréquentiels de cette fonction donnaient une estimation des fréquences formantiques souvent plus sûre que celle qui utilise les maximums du spectre, en particulier pour le premier formant.

Un travail en cours cherche à mettre en oeuvre la notion de simultanéité entre canaux adjacents pour transformer automatiquement l'ensemble des signaux issus des filtres en un petit nombre de formes d'ondes élémentaires.

RELATION AVEC CERTAINS MODELES AUDITIFS

Notre choix de filtres à phase nulle peut sembler contradictoire avec la modélisation du système auditif, car il existe dans la cochlée des délais de propagation et de filtrage qui altèrent à priori toute évaluation de simultanéité entre canaux. En fait cette différence est superficielle, si l'on considère que le système auditif peut être assimilé à un banc de filtres plus ou moins complexes, chacun étant suivi d'une ligne à retard et d'un dispositif de comparaison. Cette conception est présente dans plusieurs travaux contemporains [1, 4, 7]. En particulier R.LYON, dans son modèle très complet, est amené à associer a son "neurogramme" (histogramme évolutif des décharges nerveuses à la sortie des cellules ciliées) une "fonction d'autocoïncidence", qui restitue la notion de simultanéité, mais dans la dimension de retard et non dans la dimension de temps. La fig 8 représente schématiquement les relations entre neurogramme et fonction d'autocoïncidence.



- fig 8 -

Les notions que nous avons présentées plus haut sont liées à la possibilité d'évaluer la simultanéité des composantes spectrales. Personne n'exclut actuellement que l'audition puisse faire une large part au codage temporel [2]. Bien que l'on ne dispose pas encore de données très fournies sur les phénomènes de synchronisation entre fibres voisines, rien ne permet d'infirmer l'idée que la comparaison temporelle fine des décharges dans des fibres voisines puisse faire apparaître des entités plus globales associées à nos formes d'onde élémentaires, et que la comparaison de celles-ci entre elles puisse donner naissance à des entités correspondant à nos événements.

CONCLUSION

Le travail présenté ici repose sur l'utilisation de l'information délivrée par un banc de filtres non-causaux, dont la principale propriété est de permettre la reconstruction du signal initial par une simple sommation de tous les signaux de sortie. Ce système, qui présente certaines analogies avec l'audition humaine, permet d'examiner en détail la simultanéité des sorties des filtres à différents niveaux. Au niveau des enveloppes de ces signaux, la simultanéité conduit à la notion d'événement, qui s'applique aux concentrations locales d'énergie dans les segments voisés ou bruités; les événements peuvent être repérés au moyen de la fonction de cohérence et caractérisés spectralement. Au niveau des passages à zéro des signaux de sortie, la simultanéité conduit à une décomposition du signal en un ensemble de formes d'onde élémentaires caractérisées par un petit nombre de paramètres. Ces formes d'onde élémentaires conservent la propriété de pouvoir redonner par addition le signal initial, ce qui permet d'envisager des applications dans le domaine du codage de la parole.

REFERENCES

1 - DE CHEVEIGNE A. " A Shift-and-comparison Model of Pitch Perception". Trans. of the Committee on Speech Research of the Ac. Soc. of Japan, S-84-38, 295-302, 1984.

2 - DELGUTTE B. "Codage de la parole dans le nerf auditif". Thèse es-sciences, Univ. de Paris VI, 1984.

3 - LIENARD J.S. "Analyse impulsionnelle de la parole", JEP GALF, Montréal, 1981.

4 - LYON R.F. "Computational Models of Neural Auditory Processing", IEEE ICASSP, San Diego, 1984.

5 - MANCERON F. "Contribution à l'analyse spectro-temporelle du signal de parole considéré comme une suite d'impulsions acoustiques", thèse D.I., Univ. de PARIS XI, 1982.

6 - RODET X. "Time-domain Formant Wave-function Synthesis", Computer Music Journal, vol 8 (3), 1984.

7 - SENEFF S. "Pitch and Spectral Estimation of Speech based on Auditory Synchrony Model" IEEE ICASSP, San Diego, 1984.

COMPRESSION SPECTRALE DU SIGNAL VOCAL PAR MODIFICATION DU MODELE AUTOREGRESSIF

L.G.P. Meloni, M. Lamotte, M.J. Vigneron

Université de Nancy I - Faculté des Sciences
Centre de Recherche en Automatique de Nancy - LEA (UA 821)
B.P. 239, 54506 Vandoeuvre-lès-Nancy

ABSTRACT - *We describe a technique for improving speech reception to persons with high frequencies neurosensorial hearing impairments. This technique is pitch invariant and it is based on a modification of autoregressive model, which transpose the formants to low frequencies. The transposition and bandwidth reduction is non uniform and atteined by the poles displacement in the complex plane. The signal synthesis is made from error signal by the modified autoregressive model. This signal passes through a fixed filter which alignes the spectral peaks of the model to those of modified one. This scheme causes little distortion to rhythmic patterns, fundamental frequency contours and segment durations. We make some considerations for real time implementation and present some spectrographics exemples for illustration of the system capability.*

INTRODUCTION

Dans de nombreux cas les pertes des propriétés acoustiques peuvent être compensées par la lecture labiale, l'amplification linéaire ou l'amplification combinée avec compression en amplitude. Cependant, si les pertes dans les hautes fréquences sont totales ou sevères, la sensibilité de l'auditeur aux variations du signal en fréquence, amplitude ou temps est réduite. Il a été proposé dans ces cas, qu'on peut améliorer la reception des sons et aider à la production de la parole au moyen de la compression du spectre vers les basses fréquences.

De nombreux efforts ont été fait dans ce sens, mais le mieux que l'ont puisse dire c'est que les résultats ont été tout juste satisfaisants. Les causes des résultats négatifs ne sont pas bien connus. Beaucoup de méthodes qui ont été évaluées ont introduit d'autres modifications du spectre instantané en plus de la transformation desirée du signal[1,5]. L'approche présentée dans cet article a comme base l'estimation du modèle autorégressif (AR) suivi d'une modification convenable du filtre de synthèse. Ce système réduit les fréquences et les largeurs de bande des formants de façon non-uniforme, en gardant les caractéristiques temporelles du signal, telles que les traits rythmiques, la durée des segments, le contour de la période fondamentale. L'avantage d'une méthode du type vocodeur réside dans la souplesse des modifications possibles des paramètres soit du modèle du conduit vocal, soit de la source d'excitation.

CONSIDERATIONS DANS LA MODELISATION DU SIGNAL

Nous avons utilisée le modèle autorégressif, correspondant à l'équation récurrente:

$$y_t + a_1 y_{t-1} + \ldots + a_p y_{t-p} = e_t \qquad (1)$$

Les différentes approches qui permettent d'estimer les paramètres du modèle sont couramment décrites dans la littérature[6,7,10]. On traite le signal vocal par segments de courte durée, pendant lesquels on considère le signal comme quasi-stationnaire. Pour le signal de la parole une durée comprise entre 20 ms et 40 ms donne de bons résultats. Ces segments sont pris avec un recouvrement partiel, et non disjoints.

Les expériences de Rabiner et al.[11], sur l'analyse de la variation de l'erreur de prédiction par rapport à la position de la fenêtre d'analyse, ont montré l'importance de l'utilisation d'une fonction de pondération pour la méthode d'autocorrelation et aussi d'un filtre de pré-emphase. Cette procédure donne de meilleurs résultats dans la modélisation. Le filtre proposé par Rabiner et al. a la fonction de transfert suivante: $H(z)=1+0.95\ z^{-1}$. L'utilisation d'un filtre différentiateur pur donne des résultats équivalents, avec un coût de calcul plus réduit.

En résumé nous présentons les paramètres fixés dans la phase d'analyse:

- fenêtre de Hamming
- fréquence d'échantillonnage: 10 kHz
- filtre anti-recouvrement: 4.8 kHz
- pré-emphase: $H(z)=1+z^{-1}$

- longueur de la fenêtre: 256 échantillons
- recouvrement: 50%
- ordre du modèle AR: 10
- méthode d'autocorrélation.

L'ALGORITHME DE COMPRESSION SPECTRALE

Nous présentons à la fig. 1 le schéma de l'algorithme de compression de la parole. Pour la détermination des pôles du modèle, nous avons utilisé la méthode de Bairstow[2] qui est basée sur la méthode itérative de Newton.

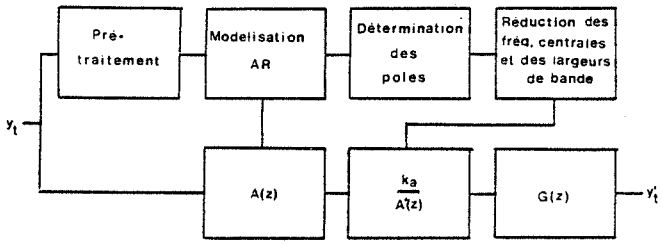


Fig. 1 *Schéma de l'algorithme de compression de la parole.*

La fréquence et la largeur de bande des formants sont liées aux pôles par les rélations:

$$F_c = \frac{f_e}{2\pi}\,\phi \qquad (2)$$

$$B = -\frac{f_e}{\pi}\,\ln\rho \qquad (3)$$

où f_e est la fréquence d'échantillonnage, ρ et ϕ sont respectivement le module et l'argument des pôles. La compression spectrale s'opère par transposition des formants et par réduction de leurs largeurs de bande selon les relations:

$$f_m = k_1\, f + k_2(1-\exp(-k_3\, f)) \qquad (4)$$

$$c_b = 1/(1 + k_4\, f^2) \qquad (5)$$

et se traduit par les courbes de la fig. 2 pour les facteurs de compression de 1/3, 1/2 et 2/3. Ces courbes ont été choisies

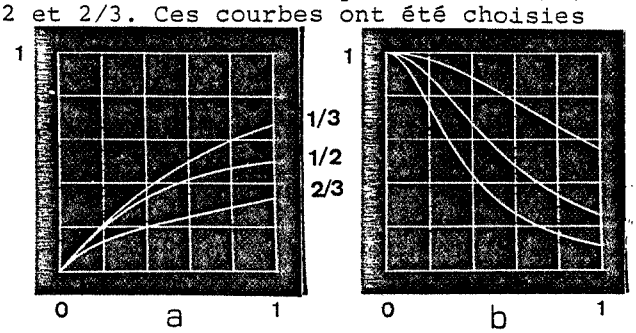


Fig. 2 *a) Courbes de réduction de fréquences centrales et b) de largeurs de bande*

de façon à laisser les basses fréquences pratiquement inchangées et à avoir une compression de plus en plus accentuée vers les hautes fréquences.

Nous pouvons voir à la fig. 3 le spectre du modèle AR de la voyelle /a/ (voix féminine). Quand on déplace les pôles on observe les effets suivants:

i) la dynamique spectrale augmente (fig. 3b). Cela est dû au fait que tous les pôles sont plus concentrés dans le 1er et 4ème quadrant et qu'il existe un vide dans les hautes fréquences. Par conséquent, la valeur continue du spectre d'amplitude est plus elevée,

ii) l'intensité des pics d'ordre plus elevé est plus forte. Celà est determiné par les courbes de déplacement de fréquence des formants et de réduction des largeurs de bande,

iii) le creux entre deux pics consécutifs diminue.

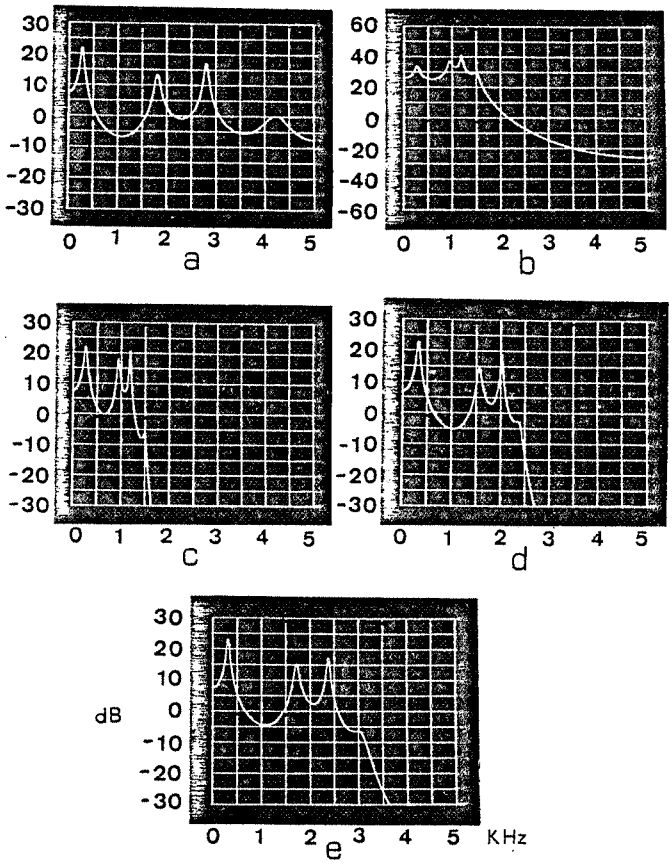


Fig. 3 *a) Spectre d'amplitude logarithmique du modèle AR original, b) spectre obtenu par déplacement des pôles, c) après égalisation du gain et filtrage, d) idem, compression de 1/2 et e) de 1/3.*

Pour aligner les amplitudes des pics spectraux sur celles du modèle original nous avons d'abord introduit un facteur d'égalisation de la valeur continue du spectre d'amplitude:

$$k_a = \frac{1 + \sum_{i=1}^{p} a_i}{1 + \sum_{i=1}^{p} a'_i} \qquad (6)$$

où les a'_i sont les coefficients du filtre obtenu par déplacement des pôles. Cette

constante est utilisée comme facteur de gain du filtre de synthèse.

Le deuxième effet peut être compensé au moyen d'un filtre fixe d'affaiblissement des formants elevés. Ce filtre obéit à des gabarits qui présentent des réponses plates jusqu'au 1/3 de la largeur de bande utile et des pentes d'affaiblissiment qui sont indiqués dans la fig. 4. Les pentes sont reglées pour atténuer suffisamment le 3ème formant sans trop affaiblir le 4ème. Dans les courbes 3c-e on observe que si on augmente trop le taux d'atténuation le 4ème formant perd beaucoup de son importance. Les filtres utilisés sont du 4ème ordre type tous-pôles.

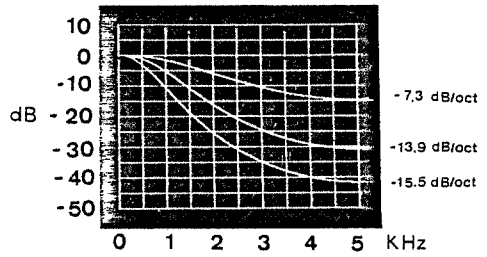


Fig. 4 *Réponses en amplitude des filtres d'affaiblissement des formants elevés.*

Le signal d'excitation du modèle modifié est le résidu, obtenu par le filtrage inverse. Nous n'avons donc pas réduit le débit du signal résiduel, comme il est pratiqué dans les vocodeurs à LPC. Ce signal répresente une estimation de la source d'excitation du conduit vocal. Nous l'utilisons pour la synthèse, en reproduisant un signal de période fondamentale inchangée.

Jusqu'ici nous n'avons pas parlé des pôles réels. Souvent les pôles du modèle AR sont des pôles réels. Nous avons pris la décision de ne pas changer la position de ces pôles. On sait que les pôles réels ont pour effet de changer la pente globale du spectre d'amplitude. On ne doit pas s'attendre à des pôles réels négatifs très proches du cercle unité, car cela signifierait un mauvais filtrage anti-repliement du signal. Quant aux pôles réels positifs, on a vu que les courbes de compression laissent inchangées les basses fréquences, ce qui justifie leur non déplacement.

Dans le cas de la compression de 2/3, la dynamique spectrale du modèle AR est très forte, et donne naissance à un coefficient k_a, définie par (6) très faible. Cela pose des problèmes d'instabilité numérique du filtre de synthèse. Nous avons résolu ce problème en décomposant le filtre AR du dixième ordre en cinq cellules du deuxième ordre et effectué l'atténuation du signal à chaque sortie des cellules.

Une autre possibilité pour l'égalisation des gains est d'utiliser un rapport entre les énergies des réponses impulsionnelles du modèle AR original et du modèle modifié:

$$k'_a = \left(\prod_{i=1}^{p} \frac{1 - k'^2_i}{1-k_i^2}\right)^{1/2} \qquad (7)$$

L'expression de la réponse impulsionnelle du modèle AR est obtenue à partir de l'algorithme de Levinson[7], où les k_i sont les coefficients de réflexion du modèle modifié. Les résultats obtenus à l'aide de cette relation ou de la relation (6) sont comparables. Nous avons choisi la relation (6) qui présente un coût de calcul plus faible.

Les coefficients $k_{1\to4}$ utilisés dans nos essais se trouvent dans la Table I. Les fréquences dans les équations (4) et (5) sont normalisées entre 0 et 1.

Table I *Coefficients des courbes de déplacements des pôles.*

Facteur de compression	k_1	k_2	k_3	k_4
1/3	-0.080	1.105	1.136	0.8
1/2	-0.100	0.711	1.860	3.0
2/3	0.213	0.120	7.463	8.0

Pour l'implantation de la méthode dans des processeurs de traitement de signal, nous devons utiliser des algorithmes avec faible coût de calcul et qui peuvent être exécutés aisément en virgule fixe. Tel est le cas de l'algorithme de Le Roux[13] pour la modélisation AR. Dans la chaîne de traitement schématisée dans la fig. 1, la partie la plus coûteuse en calcul est celle de la détermination des pôles par la méthode de Bairstow. Il existe dans la littérature plusieurs algorithmes qui permettent d'extraire les informations des formants, soit à partir du spectre d'amplitude logarithmique et de sa dérivée [3], soit à partir de la dérivée première ou troisième du spectre de phase [12]. L'algorithme de Gan et al.[4] permet le calcul rapide de la FFT des modèles ARMA (autoregressifs à moyenne ajustée). Nous avons estimé que deux processeurs TMS 32010 [8] peuvent réaliser les traitements nécessaires à la compression spectrale de la parole.

QUELQUES RESULTATS

Nous présentons à la fig. 5 les spectrogrammes du mot "assommer" original et des compressions de 1/2 et 2/3. La fréquence fondamentale et la structure temporelle sont bien préservées. Dans les spectogrammes des mots traités on observe du bruit dans les hautes fréquences, ceci

pourrait être dû à une petite imprécision dans la détermination des pôles du modèle de certaines fenêtres et aussi tout simplement au bruit de calcul.

Bien que nous n'ayons pas fait d'études de discrimination, les résultats obtenus sur quelques dizaines de mots traités nous ont montré une bonne intelligibilité ainsi qu'une bonne conservation des structures temporelles telles que les traits rythmiques, la durée des segments, le contour de la période fondamentale.

Cette méthode devrait être évaluée auprès d'enfants handicapés auditifs pour l'aide à la production vocale au Service d'Oto-rhino-laryngologie du Centre Hospitalier Régional de Nancy.

REMERCIEMENTS

Nous remercions vivement le concours de la Convention CNPq/Brésil et CEFI/France.

BIBLIOGRAPHIE

[1]L.D.Braida,N.I.Durlach,R.P.Lippmann, B.L.Hicks, R.M.Rabinowitz et C.M.Reed, "Hearing aids: a review of past research on linear amplification, amplitude compression and frequency lowering", ASHA Monography, Vol.19, 1979.
[2]C.Charet,"Cours d'analyse numérique", Sedes Informatique, 1975.
[3]R.L.Christensen, W.J.Strong et E.P. Palmer,"A comparison of three methods of extracting resonance information from predictor-coeffcient coded speech", Vol. 24, n°1, pp 8-14, Février 1976.
[4]R.G.Gan, K.F.Eman et S.M.Wu,"An estended FFT algorithm for ARMA spectral estimation", IEEE Trans. Acoust., Speech, Signal Processing, Vol.32, n°1, pp 168-170, Février 1984.
[5]B.L.Hicks, L.D.Braida et N.I.Durlach, "Pitch invariant frequency lowering with non uniform spectral compression", Proc. of th IEEE, Vol.1, pp 121-127, Avril 1981.
[6]J.Makhoul,"Linear prediction: a tutorial review", Proc. of the IEEE, Vol.63, n°4, pp 561-580, Avril 1975.
[7]J.D.Markel et A.H.Gray,"Linear prediction of speech", Springer-Verlag, Heidelberg, Alemagne, 1976.
[8]K.C.McDonough,"A single chip microcomputer archtecture optimized for signal processing", Proc. ICASSP 82, Paris, France, 3-5 Mai 1982.
[9]A.V.Oppenheim et D.H.Johnson,"Discret representation of signals", Proc. of the IEEE, Vol. 60, n°6, pp 681-691, Juin 1972.
[10]L.R.Rabiner et R.W.Shafer,"Digital processing of speech signals", Prentice-Hall, Inc., EUA, 1978.
[11]L.R.Rabiner, B.S.Atal et M.R.Sambur, "LPC prediction error. Analysis of its variation with the position of the analysis frame", IEEE Trans. Acoust., Speech, Signal Processing, Vol.25, n°5, pp 434-442, Octobre 1977.
[12]N.S.Reedy et M.N.S.Swamy,"High-resolution formant extraction from linear-prediction spectra, IEEE Trans. Acoust., Speech, Signal Processing, Vol.32, n°6, pp 1136-1144, Décembre 1984.
[13]Le Roux et C.Gueguen,"A fixed point computation of partial correlation coefficients", IEEE Trans. Acoust., Speech, Signal Processing, Vol.25, pp 257-259, Juin 1976.

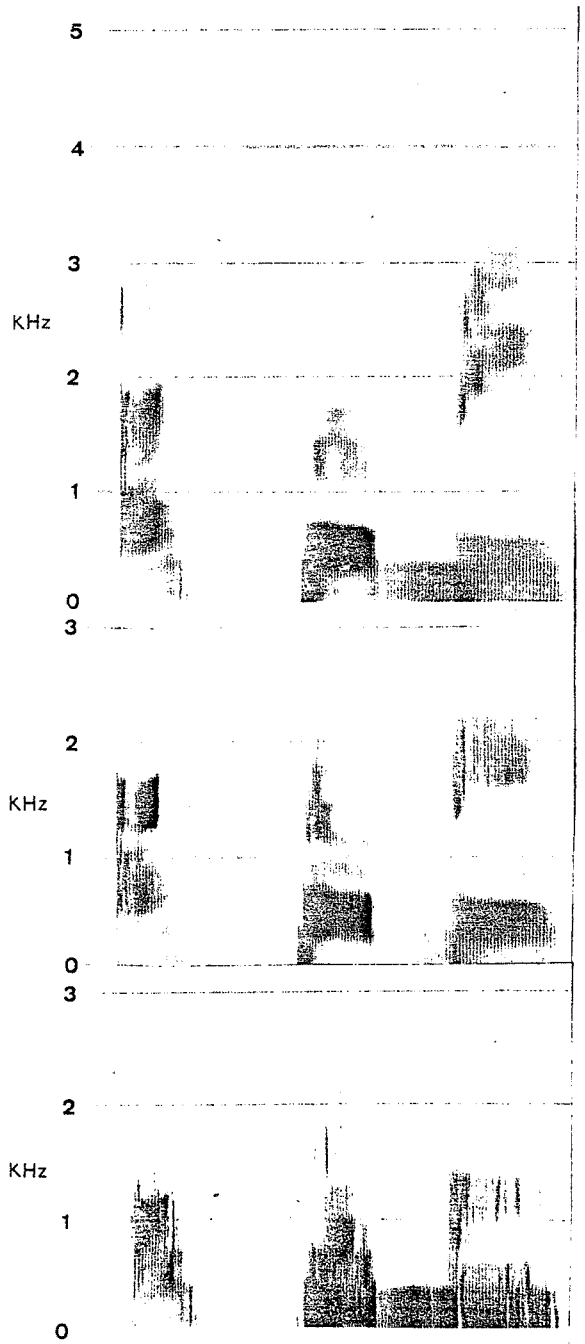


Fig. 5 *Spectrogrammes du mot "assommer" original, b) compression de 1/2 et c) compression de 2/3.*

LA NASALITE VOCALIQUE ENTRE DEUX "PÔLES" ACOUSTIQUES ...
DU CONDUIT NASO-PHARYNGAL AU CONDUIT ORAL

G. Feng, Ch. Abry, B. Guérin

Institut de Phonétique de Grenoble
Institut de la Communication Parlée, I.N.P., Grenoble

The aim of this paper is only to point up the fact that the rather complex acoustic structures of nasal vowels are "between" two simplest configurations, namely the naso-pharyngeal tract and the oral one. The effect of such an "interpolation" are hence examined for the cardinal extreme vowels, providing for each a domain of nasality to be validated.

INTRODUCTION

Les nombreuses études menées sur la nasalité vocalique n'ont pas jusqu'à présent livré de corrélats acoustiques simples et relativement indépendants de la nature des voyelles.

On a pu ainsi faire l'inventaire d'un ensemble de pôles, de zéros (ou de paires pôle-zéro) qui tous sont de bons candidats ... mais les plus importants ? La réponse n'est pas actuellement disponible - même si l'on n'en juge que d'un seul point de vue acoustique ou perceptif.

Dans ces conditions, il ne faut donc pas s'étonner que la représentation des voyelles nasales dans les espaces vocaliques classiques (F_1-F_2, F_1-F'_2, etc.) ne soit pas satisfaisante.

Simulant la production de l'espace acoustique des voyelles, avec un modèle articulatoire statistiquement réaliste, S. MAEDA [9] a mis en évidence la présence d'une faible densité des produits acoustiques situés sur le plan F_1-F_2, dans la région 250-1000 Hz, soit entre [ɯ] et [u]. Ce "vide" dans les productions du conduit oral peut-il être comblé par son couplage avec les cavités nasales ? C'est la réponse que suggère S. MAEDA. De ce point de vue la paire de pics 250-1000 Hz serait la meilleure des candidats ...

Malheureusement, une longue expérience en analyse acoustique des voyelles nasales révèle qu'il n'est - pour le moins - pas toujours très facile de retrouver dans le signal cette paire de pics pour chaque voyelle. En effet, celui vers 250 Hz est souvent absent dans les voyelles de type [a]. Bien entendu, ce premier pic est en quelque sorte "caché" pour les voyelles hautes [i] et [u] par leur premie formant. Et il en va de même pour celui vers 1000 Hz, que "cachent" tant soit peu les premiers formants de [a].

Mais si ces corrélats sont difficiles à repérer pour les voyelles, il est bien connu, par contre, depuis les premières études acoustiques sur la nasalité des consonnes, que 250 et 1000 Hz sont pratiquement des "invariants" du murmure nasal*.

Il semble ainsi plus commode, au moins dans un premier temps, de raisonner sur la nasalité vocalique à partir de la nasalité consonantique, raisonnement qui reproduit, dans une certaine mesure, le processus historique le plus courant : les voyelles nasales proviennent généralement de l'influence des consonnes nasales sur les voyelles orales.

LE CONDUIT NASO-PHARYNGAL

De toutes les consonnes nasales, nous ne considérerons que la plus simple d'un point de vue acoustique, [ɳ], celle qui approche le mieux un tube unique : le conduit naso-pharyngal.

La simulation acoustique de ce conduit a pu être bien étudiée par ailleurs [10] et la question à laquelle nous nous limiterons est la suivante : quelles sont les conditions requises d'un tel conduit pour qu'il produise deux premières résonances à 250 et 1000 Hz.

La lecture des nomogrammes de G. FANT [2], nous montre clairement qu'un tel tube nous donnera sa fréquence de résonance la plus basse avec une constriction suffisamment petite à son extrémité. Plus cette constriction sera loin de la source (plus le volume sera grand), plus basse sera la résonance. Avec une telle condition à l'extrémité du tube, le pic de 250 Hz est relativement peu sensible; et cela reste vrai même si une autre constriction est déplacée à l'intérieur du tube.

La réalité anatomique de la sortie nasale, reconsidérée à la lumière de son équivalence acous-

*. Le plus invariant des deux étant bien entendu le premier [3].

tique - en tant que sortie d'un analogue du conduit nasal - semble offrir la possibilité d'une telle condition*.

En ce qui concerne le pic vers 1000 Hz, on peut obtenir une telle résonance - simultanément avec le 250 Hz - si on a une longueur de tube suffisante. Ce second pic est par contre plus sensible que le premier; en particulier lorsqu'une autre constriction est produite dans la partie centrale du tube : plus la constriction est petite, plus il sera abaissé.

Ces résultats sur la sensibilité différente des deux pics de la nasalité, sont en accord avec le fait que l'indice le plus important de [nasal] pour les consonnes est bien la fréquence la plus basse (avec une amplitude relativement grande [12]). C'est elle qui sera la plus résistive aux changements de forme du conduit nasopharyngal, par clôsion du velum ou par changement de la taille du pharynx.

LE BRANCHEMENT DU CONDUIT BUCCAL SUR LE CONDUIT NASOPHARYNGAL.

Il est courant d'étudier les effets du couplage en partant des structures acoustiques orales, considérant ainsi qu'elles sont "perturbées" par leur connexion avec les cavités nasales. Nous avons dit plus haut qu'il était difficile de retrouver les caractéristiques orales des voyelles nasales. Sera-t-il plus facile de récupérer les corrélats acoustiques des consonnes nasales dans ces voyelles ? Ou encore - moins ambitieusement - serait-il plus facile ainsi d'expliquer pourquoi il est plus ou moins difficile de retrouver ces corrélats selon la nature des voyelles ?

Nous avons choisi pour point de départ de simuler un conduit naso-pharyngal réaliste approchant au mieux nos conditions pour produire 250 et 1000 Hz. Cette configuration articulatoire correspond approximativement à la coproduction de [ɲ] avec [i], qui présente un grand pharynx.

Cette configuration de départ est ensuite modifiée pas à pas pour atteindre la configuration entièrement orale de [i].

Au départ, le velum reposant sur la langue, ferme la sortie buccale. Son élévation va accroître le degré de couplage de la partie orale et décroître celui de la partie nasale (Fig. 1-4).

Ceci correspond acoustiquement à la transition entre deux spectres sans zéros* : sur le premier on a 260-1090-2440 Hz en tant que premières résonances; sur le second, une structure de type [i], soit 240-2270-3240 Hz (Fig. 5).

Pendant l'évolution du couplage une première paire de pôle-zéro apparaît dans la gamme 250-2000 Hz. La distance entre ce pôle et ce zero décroît et leurs valeurs tendent vers celle du premier formant de [i].

Dans cette transition, les connaissances en analyse et en perception peuvent nous permettre de valider un domaine où la voyelle est nasale.

En ce qui concerne la production de [u], nous partons d'une configuration à 310-940-2180 Hz (Fig. 6). Le premier pôle reste stable durant toute l'évolution vers la voyelle orale. Le second pic se déplace vers le second formant de [u] et la première paire pôle-zéro évolue entre ces deux pôles.

Nous pouvons obtenir, là aussi, une assez large zone de validation de la nasalité de la voyelle.

L'évolution pour la voyelle [a] part d'un conduit nasopharyngal produisant 350-1040-1890 Hz comme premières résonances (Fig. 7). Nous remarquons que, s'il y a bien toujours un pic dans la région 1000-1300 Hz, le 350 Hz de la configuration initiale tend à s'accroître. Comme dans les transitions précédentes, une première paire pôle-zéro apparaît : elle évolue de la région des 1000 Hz vers celle du premier formant de [a].

Les données d'analyse (en particulier les données de balayage fréquentiel [4]), confirment ici que le domaine de la nasalité vocalique pour [a] est plutôt proche des configurations qui présentent l'ouverture velopharyngée maximale - contrairement à [i] et [u] qui offrent un plus grand domaine de validation**.

---

*. En l'absence de données statistiques sur l'anatomie du conduit nasal, les valeurs données dans la "littérature" varient de 1 à 10 ([5] et [1]).

*. Le spectre nasopharyngal ne pourrait en effet présenter de zéros que si l'on introduisait le couplage avec les cavités des sinus. L'effet de celles-ci existe [6], lorsque les canaux pneumatiques qui les relient aux fosses nasales ne sont pas trop petits ou ... bouchés, ce qui est fréquent. Il n'existe pas, à notre connaissance, d'étude contrôlée sur les effets (y compris compensatoires) de leur absence/présence sur la production des nasales. On ne peut donc que constater que les études qui les introduisent dans la simulation des voyelles nasales [8] obtiennent des résultats proches de ceux que nous présentons, avec une condition de sortie petite.

**. Ceci correspond au fait bien connu que les voyelles basses exigent une ouverture vélopharyngée plus grande pour être catégorisées comme voyelles nasales [11].

CONCLUSION

L'approche que nous avons proposée semble offrir une perspective qui nous permet de situer clairement la nasalité des voyelles entre ses bornes, ou si l'on veut, entre ses deux "pôles", l'oropharyngalité et la nasopharyngalité.

Partant de la structure nasale la plus simple à contrôler, celle des consonnes, il est par là-même plus facile de suivre la persistance ou la disparition des corrélats majeurs de la nasalité.

La démarche qui a consisté, jusqu'à aujourd'hui à examiner directement les résultats de la complexification apportée par la nasalité dans la riche gamme de structures déjà possibles pour les voyelles orales, nous semble au contraire avoir empêché un suivi clair des phénomènes principaux dûs à ce type de couplage.

Nous pensons, qu'à partir de ce jalonnement s'offrent plusieurs directions de recherche exploitables; notamment en ce qui concerne :

- l'étude de la dynamique des voyelles nasales [7];
- la tendance à la nasalisation en fonction de la hauteur des voyelles [11];
- la neutralisation des consonnes nasales dans le soi-disant appendice [ŋ].

Remerciements

Pour leurs commentaires (formels ou informels) : à S. MAEDA, L.J. BOË, D. AUTESERRE, P. BADIN.

Pour leur aide : à F. CHARPENTIER, B. LHERM, H. SANCHEZ. Manuscrit : D. VUILLET.

References

[1] G. Bjuggren & G. Fant, "The nasal cavity structures", STL-QPSR 4, pp 5-7, 1964.

[2] G. Fant, "Acoustic theory of speech production", The Hague, 1960.

[3] O. Fujimura, "Analysis of nasal consonants", JASA 34, pp 1865-1875, 1962.

[4] O. Fujimura & J. Lindqvist, "Sweep-tone measurements of vocal tract characteristics", JASA 49, pp 541-558, 1971.

[5] A.S. House & K.N. Stevens, "Analog studies of the nasalization of vowels", Journal of Speech and Hearing Disorders 21/2, pp 218-232, 1956.

[6] J. Lindqvist & J. Sundberg, "Acoustic properties of the nasal tract", STL-QPSR 1, pp 13-17, 1972.

[7] P. Linthorst, "Les voyelles nasales du français. Etude phonétique et phonologique", Groningen, 1973.

[8] S. Maeda, "The role of the sinus cavities in the production of nasal vowels", Proc. IEEE/ICASSP, Paris, pp 911-914, Mai 1982.

[9] S. Maeda, "Une paire de pics spectraux comme corrélat acoustique de la nasalisation des voyelles", 13e JEP du Galf, Bruxelles, pp 223-224, Mai 1984.

[10] B. Merlier, "Etudes sur le rayonnement externe des conduits oral et nasal", Thèse de Docteur-Ingénieur, INP Grenoble, 1984.

[11] P. Reenen (Van), "Phonetic feature definitions. Their integration into phonology and their relation to speech. A case study of the feature NASAL", Dordrecht, Cinnaminson, 1982.

[12] K.N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds", JASA 69 (Suppl. 1), 1981.

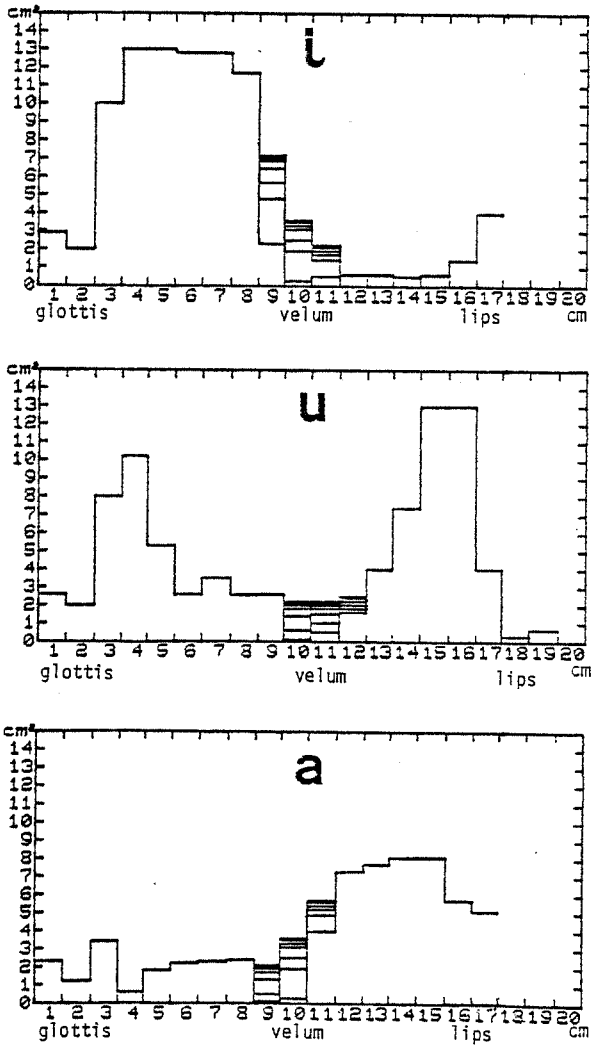


Fig. 1-3. Variations des fonctions d'aire pour [i], [u] et [a] (partie oro-pharyngale).

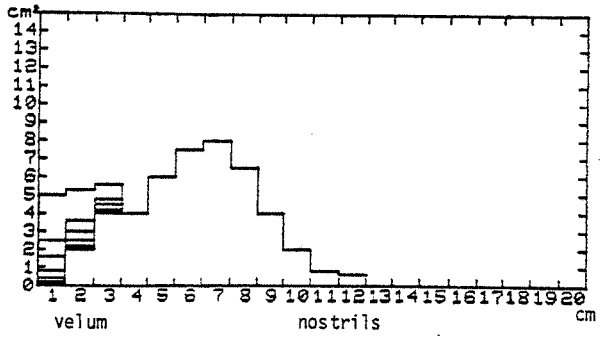
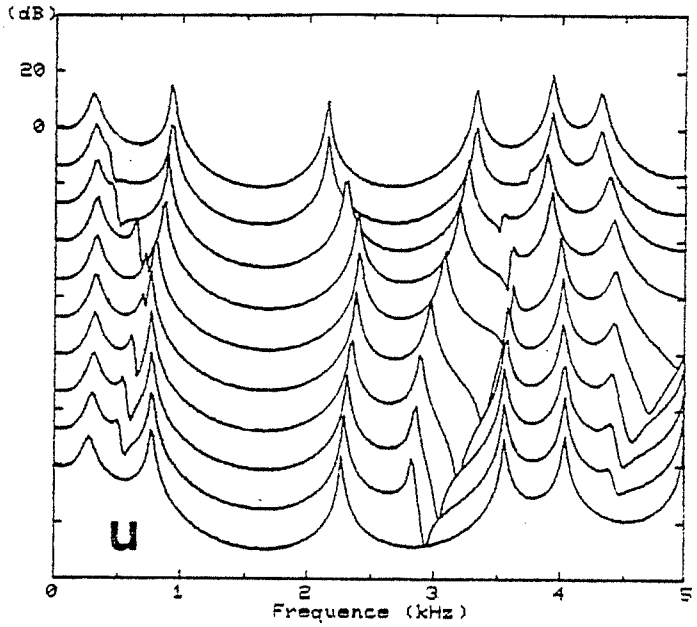
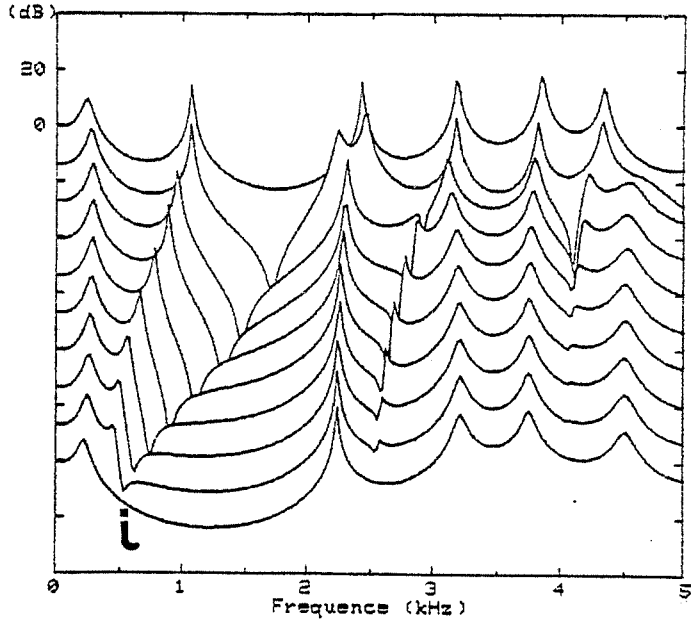


Fig. 4. Variations de la fonction d'aire du conduit nasal.





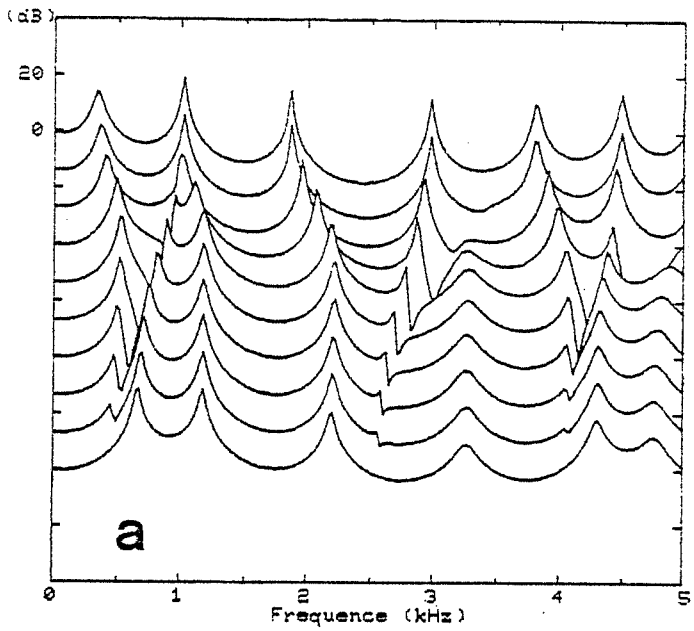


Fig. 5-7. Du conduit naso-pharyngal (en haut) au conduit oro-pharyngal (en bas) : pour [i], [u] et [a].

UNE EVALUATION COMPARATIVE DE TROIS METHODES D'EXTRACTION ET DE SUIVI AUTOMATIQUE DE FORMANTS.

J. Génin*, G. Feng+ & L.J. Boë+

*Centre National d'Etudes des Télécommunications
38243 Meylan

+Institut de la Communication Parlée
Institut de Phonétique de Grenoble
38400 St Martin d'Hères

ABSTRACT.

Three methods for automatic formant extraction and tracking have been developped for different goals.

Two of them aim to speech analysis/synthesis (cepstrum and LPC).

The last one has been studied in connection with sensory auditory aids for deaf people.

We compare the results for 54 non-sense words uttered by a female speaker. The speech material had been choosen to explore boundaries of the vocalic system and we had choosen a female speaker because of her a priori poorer spectral definition.

I. INTRODUCTION.

La description des systèmes vocaliques, l'évaluation des distances et l'élaboration de modèles perceptifs, la description des phénomènes de coarticulation et leur vérification par la synthèse, nécessitent l'obtention des formants vocaliques (F1 à F3, voire F4).

Le peu d'information que l'on peut fournir à un patient sourd total au moyen d'une prothèse auditive sensorielle, les interactions électriques inévitables entre les électrodes implantées, conduisent à rechercher une représentation du signal de parole de type formantique plutôt que sous forme de canaux.

Comme c'est le cas pour l'évaluation des méthodes de mesure de F0, il n'existe pas de procédure "type" pour comparer différentes méthodes de détection de formants.

Notre démarche a consisté à explorer les bornes du système vocalique d'un locuteur féminin dont on sait, a priori, que la quantification spectrale est moins fine, à cause de la valeur plus élevée du fondamental, si on la compare à celle d'un locuteur masculin.

Quel que soit le contexte consonantique on s'attend à un certain nombre de difficultés :

- pour les voyelles fermées /i,y,u/ il sera difficile de séparer F0 et F1 ;

- F1 et F2 sont très voisins pour les voyelles ouvertes /a/, /ɑ/ et /ɔ/ ;

- F3 de /u/ présente une amplitude très faible (-40 dB par rapport à F1), qui le rend peu visible même sur des représentations de type sonagraphique ;

- les voyelles nasales se caractérisent par un aplatissement du spectre (par élargissement des bandes passantes) et l'apparition de formants de nasalité (250 Hz et 1000 Hz).

C'est dans les syllabes /pa/, /si/, /fu/ que les 3 voyelles extrêmes se trouvent déplacées aux bornes de leur espace de réalisation (Abry & Boë, 1984).

- dans le cas du /i/, l'influence de la consonne /s/ élève F2, F3 et F4, ce qui n'entraîne aucune difficulté de mesure ;

- pour /a/, la consonne /p/ augmente F1, l'amenant ainsi plus près de F2 ;

- pour /u/, l'écart d'amplitude entre F1/F2 et F3, déjà élevé (Fant, 1956), est accru par coarticulation endolabiale, le /f/ de /fu/ diminuant F2.

Nous avons choisi de tester nos trois procédures sur ces points sensibles plutôt que de nous livrer à une évaluation globale qui, dans son traitement statistique, n'aurait pas permis de mettre en évidence ces types de difficultés (voir figure 1).

II. LES METHODES UTILISEES.

Le signal de parole a été numérisé* à 10 kHz et codé sur 12 bits.

* Nous remercions R. Descout Départ. RCP CNET Lannion qui a mis en place les procédures d'enregistrement de la base de données des sons du GRECO Parole.

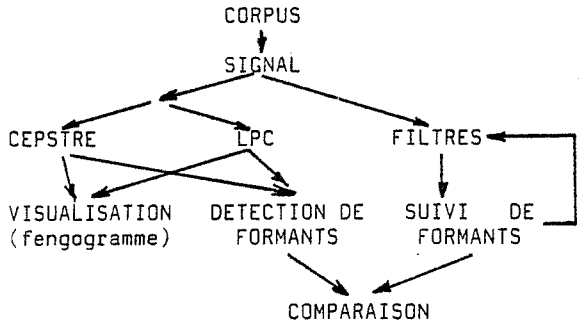


Figure 1. Procédure utilisée.

Les deux premières méthodes ont été réalisées sous forme modulaire :

. Analyse par cepstre et LPC (M1 et M2) ;

. Visualisation spectrale, détection et suivi de formants.

1. Le cepstre et le codage prédictif.

1.1 Le cepstre.

Pour cette méthode bien connue, nous avons testé certains points délicats :

. La fenêtre d'analyse est une fenêtre de HAMMING de 25,6 ms (soit 256 points), que l'on déplace de 6,4 ms pour chaque trame d'analyse.

. La préemphase est réalisée par un filtre ne comportant qu'un zéro de forme :

$$1 - \mu 3^{-1} \quad \text{avec} \quad \mu = 0.75$$

. Le choix de la fenêtre cepstrale est de loin le plus délicat : il s'agit de séparer l'apport de la source et celui du conduit vocal.

On doit tenir compte de la différence homme/femme. Nous avons retenu une fenêtre de type semi-HAMMING de longeur 4 ms pour les locuteurs masculins et 3,2 ms pour les féminins. Une pondération particulière a été apportée pour réduire l'influence du niveau du global signal (Feng, 1983).

1.2 Le codage prédictif.

La méthode d'autocorrélation a été adoptée (Markel & Gray, 1976) : plus stable que la méthode de covariance, elle se prête mieux à un suivi de formants. On a conservé les mêmes conditions d'analyse que pour le cepstre (longueur de fenêtre, pas de déplacement et préaccentuation).

Un nombre de coefficients égal à 14 paraît un bon compromis : au delà on aboutit à une "surdétection" de formants et au-dessous à des manques manifestes.

1.3 Visualisation "spectrale".

Afin de pouvoir évaluer globalement les qualités de l'analyse et des résultats du suivi formantique, il était indispensable de présenter des tracés de type sonagraphique : les "fengogrammes" réalisés à partir des analyses cepstrales ou LPC.

1.4 Détection et suivi des formants.

Une première étape consiste à éliminer des maxima locaux et à retenir des candidats potentiels. Ensuite la cohérence des structures des candidats détectés dans 3 trames (Feng, 1984) successives permet de sélectionner les formants.

2. L'analyse temporelle (M3).

Pour chaque formant, on adopte la procédure suivante inspirée de Watanabe (1980) :

- Filtrage passe-bande pour une gamme de fréquences dans laquelle on s'attend à voir évoluer le formant : 300 - 950 Hz pour F1, 700 - 2800 Hz pour F2 et 2000 - 4000 Hz pour F3. Les gabarits des filtres retenus ici sont du type TCHEBYTCHEFF avec une ondulation de 3 dB dans la bande passante.

- Filtre coupe-bande piloté par la fréquence détectée pour chaque formant contigu. On utilise ici une cellule de degré 2 dont la largeur de bande fixe correspond, en moyenne, à celle du formant à inhiber (50, 75, 150 Hz pour F1, F2, F3).

- Mesure de la période "dominante" du signal ainsi obtenu après utilisation d'un seuil proportionnel à la moyenne de la valeur absolue. On impose à la valeur de la fréquence ainsi évaluée de ne pas sortir de la gamme qui a été choisie initialement pour le filtre passe-bande correspondant (figure 2).

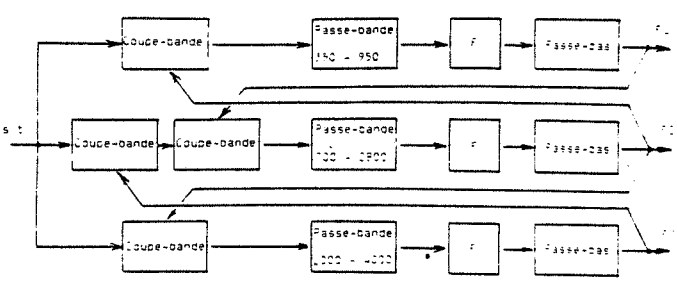


Figure 2. Schéma de la méthode temporelle.

III. RESULTATS.

Nous allons reprendre une par une les difficultés pour les trois méthodes M1, M2 et M3.

. Pour /i/ :

Séparation F0/F1

Le cepstre et le LPC donnent un résultat qui n'est pas influencé par F0. Avec la méthode temporelle, la détection du premier formant risque d'être perturbée dans ce cas là. Aussi nous envisageons de détecter F0 et d'utiliser un filtre réjecteur supplémentaire. On constate que F2 et F3 sont bien détectés.

. Pour /a/ :

Le voisinage de F1/F2 pose à M1 et M2 de sérieux problèmes de séparation, car ces méthodes reposent sur la détection de maxima alors qu'ici il faudrait rechercher plutôt une inflexion ; M3 semble mieux s'en sortir grâce aux filtres réjecteurs dont les largeurs de bandes sont égales à celles des formants (en moyenne).

. Pour /u/ :

Les méthodes spectrales achoppent au problème du peu d'énergie au delà de 1500 Hz. M3 s'en tire beaucoup mieux car cette méthode cherche un formant dans la bande 2200 - 4000 Hz.

. Pour /ã/ :

On commence à connaître la répartition formantique des voyelles nasales (Feng & al. 1985). Compte tenu de l'aplatissement spectral, pour la voyelle ouverte on peut dire que les méthodes spectrales nous proposent un choix relativement satisfaisant (250, 850, 1000 et 2200 Hz) ; la méthode temporelle, qui n'est prévue que pour détecter trois formants va "hésiter" entre les quatre valeurs possibles.

IV. CONCLUSION.

Les méthodes spectrales permettent en moyenne une bonne précision dans la mesure des formants, elles ne nécessitent aucune fourchette a priori pour détecter un maximum dans la fonction de transfert du conduit vocal.

Le LPC est un peu moins performant que le cepstre car il rajoute assez souvent des extra-formants entre des maxima éloignés (cas du /i/). On peut d'ailleurs s'en rendre compte d'une façon globale sur les représentations de type sonagraphique.

Pour le /a/, F1 et F2 ne se manifestent pas par la présence de 2 pics mais d'un seul accompagné par un point d'inflexion.

La voyelle nasale /ã/, qui semblait la plus délicate à traiter ne pose pas trop de difficultés.

La méthode temporelle est relativement complémentaire ; elle offre les avantages de ses inconvénients : suivi "consciencieux" des formants, au risque d'un mauvais choix initial.

Par rapport aux buts fixés, ces trois méthodes semblent bien utilisables.

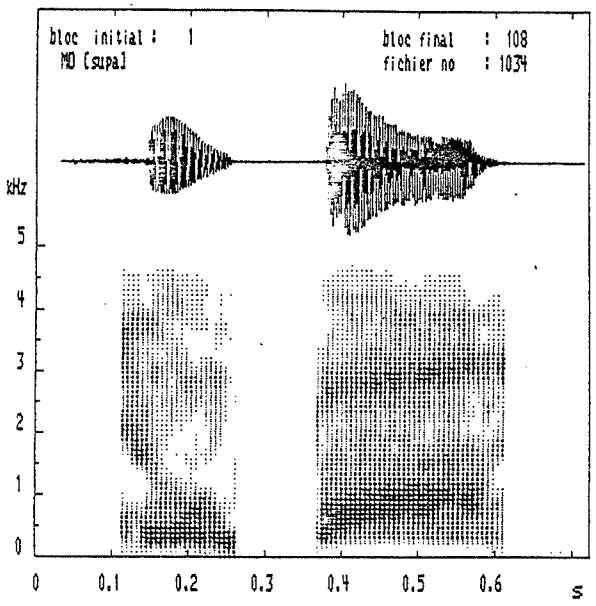


Figure 3. Fengogramme à partir du cepstre. La compacité de la voyelle /a/ va rendre difficile la séparation F1/F2.

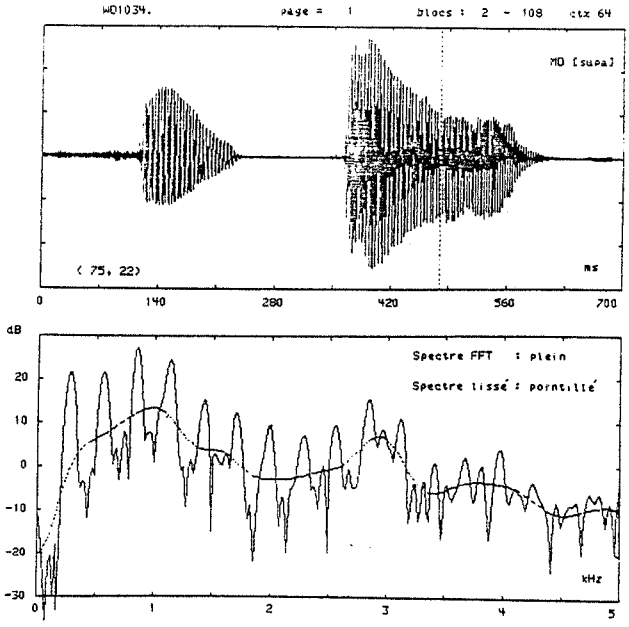


Figure 4. Coupe cepstrale dans la voyelle /a/. Mise en défaut de la détection des maxima (sur F1).

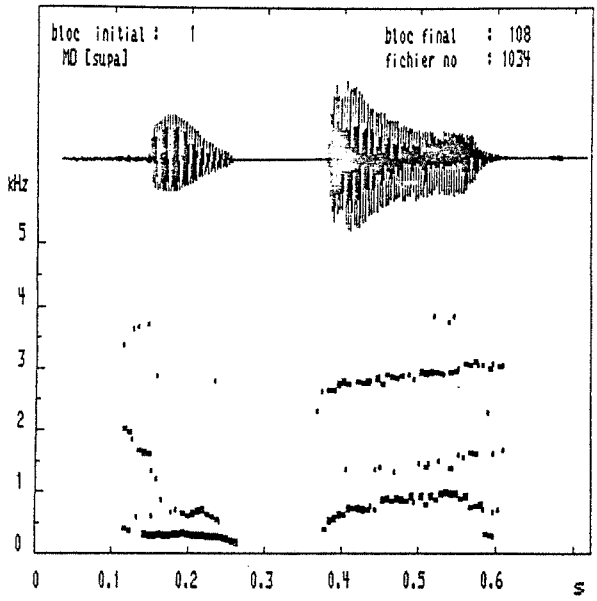


Figure 5. Suivi formantique à partir du cepstre.

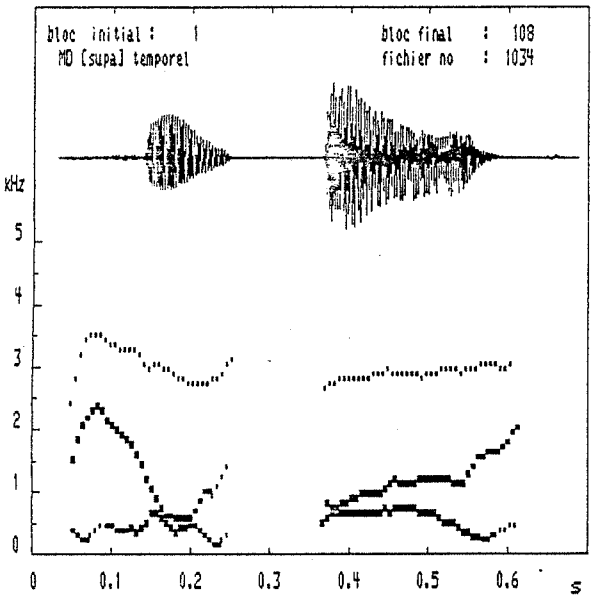


Figure 6. Suivi formantique. Méthode temporelle.

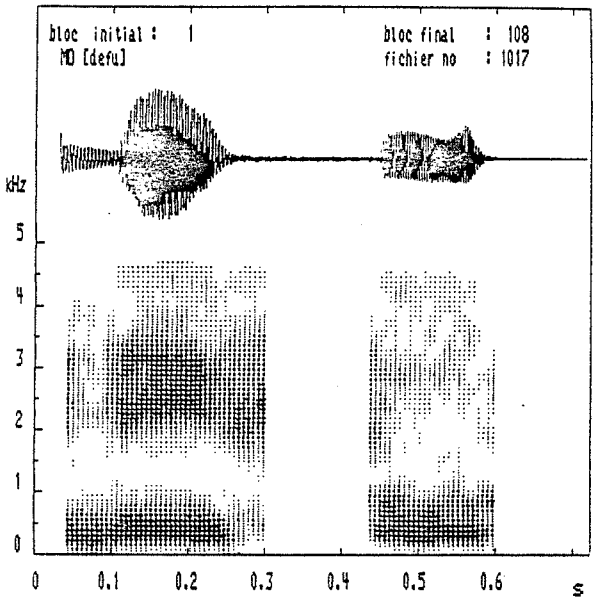


Figure 7. Fengogramme. Difficulté de mise en évidence de F2 et F3 de /u/.

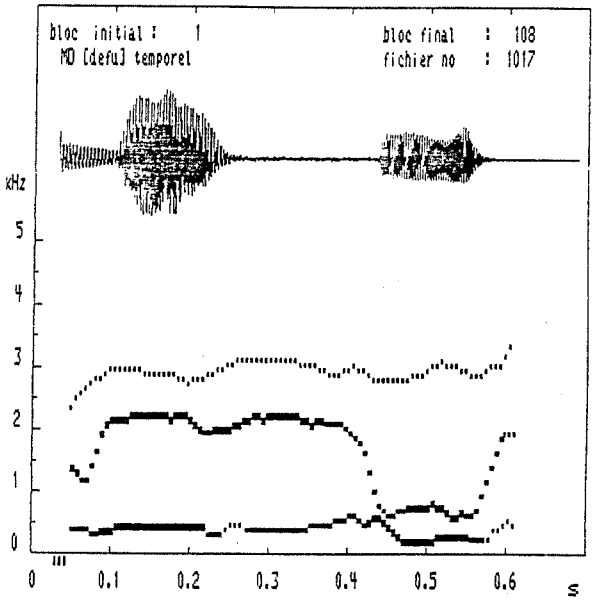


Figure 8. Bon comportement de la méthode temporelle sur /u/.

REFERENCES.

C. Abry & L.J. Boë, "(i, a, u) ? Pas si fou ? Ou les lèvres des consonnes maximisent-elles l'espace acoustique des voyelles ?", 13èmes JEP, GCP-GALF pp. 205-207, 1984

C.G.M. FANT, "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies", Readings in Acoustic Phonetics, Ed. by E. Lehiste, pp. 44-56, M.I.T. Press, 1957.

G. Feng, "Détection et mesure numériques de la fréquence fondamentale et des formants du signal de parole", DEA CEPHAG Inst. Phonét. Grenoble, 119 p. 1983.

G. Feng, "Analyse cepstrale, visualisation sonagraphique et détection des formants", Séminaire Traitement du Signal de Parole, GALF GRECO pp. 207-216, 1984.

G. Feng, C. Abry & B. Guérin, "How to Cope with Nasal Vowels? Some Acoustic Boundary Poles", Symposium Franco-Suédois, Grenoble 1985.

J.D. Markel & A.H. Jr. Gray, "Linear Prediction of Speech", Springer Verlag N.Y., 1976.

A. Watanabe, "A Real-Time Formant Tracker using Inverse Filters", STL QPSR 3-4, pp. 1-30, 1979.

APPRENTISSAGE BINAIRE DES NOYAUX STATIONNAIRES

H. PIGOT P. DELEGLISE

UNIVERSITE PIERRE ET MARIE CURIE - INSTITUT DE PROGRAMMATION -
Laboratoire de Reconnaissance de Formes et de Parole -
4, Place Jussieu - 75230 PARIS CEDEX 05

ABSTRACT

This study is focused on the steady states segments recognition to be used as anchor segments. The corresponding windows are automatically extracted from continuous speech by an algorithm based on the energy variation. The L.P.C. spectrum is coded as a binary description in two ways :
1 - description of the L.S.P.,
2 - description of the spectral maxima.
A logic learning procedure, which uses an automatic construction of adapted boolean expressions, is applied in order to classify the steady segments into vocalic groups.

INTRODUTION

Nous cherchons à reconnaître des parties stables du signal de parole continue. Ces segments serviront ultérieurement de points d'ancrage à des algorithmes de reconnaissance prenant en compte les phénomènes de coarticulation. Nous proposons ici une méthode pour extraire de telles parties de manière automatique et pour les reconnaître au moyen d'un apprentissage binaire.

I - EXTRACTION ET ETIQUETAGE

Les études sur la segmentation par l'énergie (1) ont montré que de telles parties se trouvaient le plus souvent dans les maxima de l'énergie du signal.

Pour exploiter ce critère nous avons découpé le signal en tranches de 10 ms, et détecté sur chaque fenêtre le maximum de l'énergie lissée sur une durée de 1 ms. L'abscisse temporelle de ce maximum nous donne la position d'un début de cycle de voisement, s'il y en a un dans la fenêtre. L'amplitude du maximum caractérise l'énergie de la fenêtre.

Sur ces paramètres nous effectuons un filtrage pour ne sélectionner que les maxima vérifiant les propriétés suivantes :
- Amplitude supérieure à un seuil fixé à l'avance.
- Entre deux maxima sélectionnés il existe au moins un minimu situé à plus de 40 ms de chacun d'eux.

La sélection de tels maxima a été réalisée au moyen d'un jeu de règles de réécriture programmées avec prédicats (3).

Pour évaluer ce traitement et étiqueter les fenêtres en vue de l'apprentissage, nous avons écouté les noyaux stationnaires dans un voisinage de 0,1 seconde autour des fenêtres choisies. On constate :
- une partie de ces fenêtres sélectionnées (32 %) n'est identifiable à l'oreille que dans un contexte plus long,
- cette identification n'est pas améliorée par une une prononciation plus lente
- les parties centrales des fricatives peuvent être sélectionnées si le seuil d'amplitude est faible.
- quelques erreurs se produisent dans le cas de transition semi-voyelles voyelles.

Du résultat de cette étude nous avons retiré un corpus qui comprend 524 noyaux stationnaires identifiables comme fricatives, extraits de 3mn30 de parole continue monolocuteur. Ce corpus est divisé en deux sous ensembles disjoints :

- l'ensemble d'apprentissage
- l'ensemble de reconnaissance.

Sur l'ensemble du corpus nous affectuons :

1 - une analyse L.P.C. à 16 coefficients avec extraction des maxima du spectre dont les caractéristiques (amplitude, fréquence, largeur de bande) sont obtenues par une analyse géométrique de celui-ci (2).

2 - une analyse L.P.C. à 12 coefficients avec dilatation spectrale et extraction des lignes spectrales (4). La dilatation spectrale est utilisée pour mettre un poids plus élevé sur les basses fréquences que sur les hautes fréquences dans la résolution du modèle L.P.C. (5).

II - APPRENTISSAGE LOGIQUE

1 - Principe

Nous voulons reconnaître l'étiquette phonétique des noyaux stationnaires (les objets). L'étiquetage d'un noyau stationnaire est parfois incertain quand la voyelle est très coarticulée avec les phonèmes qui l'encadrent, aussi les classes ne sont pas toujours séparables et un objet peut n'être assimilable ni à une classe ni à son contraire.

Nous avons utilisé une méthode d'apprentissage binaire qui décrit chaque classe par un ensemble de formules logiques (7,8). Les formules établies sont de type discriminatif entre deux classes, c'est-à-dire qu'une formule est sélectionnée dans une classe si elle décrit bien la classe et si elle est peu vérifiée par les objets de l'autre classe. On ne cherche pas à ce que chaque formule soit vérifiée par tous les objets d'une classe mais à ce que l'ensemble des formules recouvre la classe entière.

2 - Méthode

a - Description des objets

Chaque objet est codé sur une suite de descripteurs binaires. Pour notre application, un descripteur décrira par exemple une bande de fréquences du spectre : il sera égal à 1 si l'enveloppe spectrale présente un maximum dans cette bande.

b - Sélection des formules logiques

Les seuils de sélection des formules sont fixés en fonction de la complexité de la formule pour chaque type de formule. Une formule sera sélectionnée dans une classe si le % d'objets qui la vérifient dans cette classe est supérieur au seuil de sélection et si peu d'objets de l'autre classe la vérifient.

Les formules des 2 classes sont étudiées simultanément, en 3 étapes, successivement pour 1, 2 et 3 descripteurs. Pour limiter la combinatoire de construction des formules, un descripteur choisi pour une formule à la première étape est éliminé pour la construction des formules à 2 et 3 descripteurs, et une paire de descripteurs choisie à la deuxième étape est éliminée pour la construction des formules à 3 descripteurs.

Sur toutes les formules sélectionnées, un algorithme de compression est appliqué dans chaque classe pour éliminer les formules redondantes (6).

c - Reconnaissance

Un objet est reconnu par une classe si le pourcentage de formules de cette classe qu'il vérifie est supérieur à celui des formules vérifiées dans l'autre classe. Si la différence n'est pas significative alors l'objet sera considéré bruité.

La reconnaissance est d'abord effectuée sur l'ensemble d'apprentissage pour évaluer la validité de l'apprentissage puis sur l'ensemble de reconnaissance pour évaluer sa généralisation.

III - APPLICATION

1 - Codage des maxima du spectre

a - Codage par filtres

Les maxima du spectre sont décrits sur 35 descripteurs, chacun codant une bande de fréquence ou d'amplitude. Les bornes des filtres sont calculées à partir des objets des deux classes à discriminer ce qui présente l'avantage d'une adaptation au locuteur et aux segments de parole étudiés (par exemple, les filtres des fricatives seront déplacés vers les hautes fréquences par rapport à ceux des voyelles).

Tout d'abord, le spectre est découpé en 16 bandes fréquentielles auxquelles sont associés 16 descripteurs égaux à 1 si au moins une proéminance se trouve dans la bande correspondante.

De plus, les trois proéminances de fréquence plus faible sont décrits sur 19 descripteurs : la fréquence centrale est quantifiée sur 3 bits en fonction de seuils établis sur l'ensemble d'apprentissage, de même l'amplitude et la largeur de bande sont quantifiées chacune sur 2 bits.

Les spectres étant normalisés sur l'énergie maximale, la première proéminance n'est pas codée sur les 2 descripteurs de l'amplitude.

b - Stratégie de reconnaissance

Nous avons testé notre codage en discriminant une voyelle contre les autres. Nous obtenons respectivement pour l'ensemble d'apprentissage et de reconnaissance 94 % et 87 % de noyaux stationnaires reconnus comme conformes à leur étiquette vocalique, mais à cette étape chaque objet est reconnu par un trop grand nombre d'étiquettes.

Nous effectuons sur le même ensemble d'apprentissage un nouvel apprentissage qui, au lieu de discriminer chacune des 12 voyelles contre le reste de l'ensemble, discrimine une voyelle contre l'autre. Pour chacune de ces 66 discriminations, l'objet à reconnaître sera codé sur les filtres calculés à partir des 2 voyelles discriminées. Cela permet d'axer la reconnaissance sur les spécificités séparant 2 voyelles entre elles.

Un objet après avoir été reconnu globalement dans la discrimination d'une voyelle contre l'ensemble obtient une liste d'étiquettes vocaliques possibles.

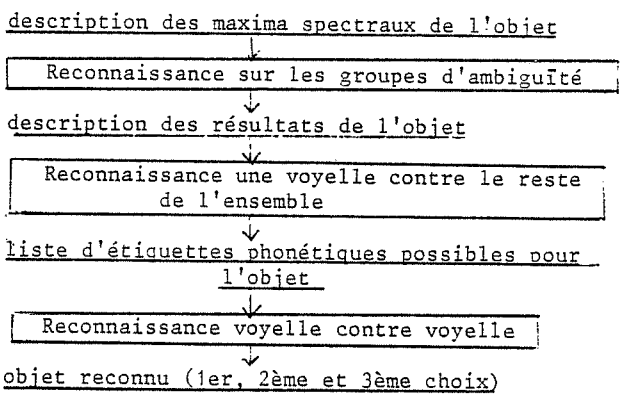
Pour supprimer les ambiguïtés, nous discriminons l'objet sur les paires d'étiquettes constituant cette liste. Le nombre d'ambiguïtés diminue et nous obtenons les résultats suivants :

% d'objets reconnus	1°choix	2°choix	3°choix
ensemble d'apprentissage	79 %	90,3 %	91,3 %
ensemble de reconnaissance	61 %	79,6 %	82,3 %

Les résultats sont très sensibles à la généralisation sur l'ensemble de reconnaissance. Ceci tient au fait que l'apprentissage global (une voyelle contre le reste de l'ensemble) ne tient pas suffisamment compte des différences séparant les voyelles entre elles ; la classe des autres voyelles est trop

hétérogène pour que chaque voyelle soit bien décrite. Pour éviter cela, nous substituons à l'apprentissage une voyelle contre le reste de l'ensemble, un apprentissage en deux étapes. La première étape tient compte des différences locales entre les voyelles ; la seconde réapprend chaque classe en fonction du comportement des objets sur ces différences locales. Pour ce faire, l'ensemble d'apprentissage est divisé en plusieurs partitions d'après les traits acoustiques et les confusions obtenues lors du test précédent.

La stratégie de reconnaissance est expliquée par le schéma suivant :



2 - Codage des lignes spectrales

Comme les lignes spectrales plongent notre espace de représentation dans un espace euclidien de dimension 12, nous effectuons un pré-apprentissage à l'aide de la méthode de Représentation par Sous-Espaces Vectoriels (R.S.E.V.) (9).

Cette méthode consiste dans le cas général à construire des s.e.v. représentant un concept défini par des exemples. On considère pour cela le nuage de points formé par les éléments de l'ensemble d'apprentissage vérifiant le concept C. Puis on construit la suite des s.e.v. emboîtés $S_1,\ldots,S_i,\ldots,S_{12}$ tels que S_i soit le s.e.v. de dimension i qui porte le plus d'inertie du nuage parmi les s.e.v. de même dimension. On fait de même avec le concept $\tilde{C}$, défini comme la négation du concept C, et on obtient alors $\tilde{S}_1\ \tilde{S}_2\ldots\tilde{S}_{12}$.

Pour tout point x, on considère le rapport $\dfrac{||P_{S_i} c\ (x)||}{||P_{S_i} \tilde{c}\ (x)||}$ où $P_{S_i} c(x)$ désigne la projection de l'objet x sur l'espace de dimension i représentant le concept C. Ce rapport définit une similarité entre l'objet x et le concept C.

Nous avons appliqué cette méthode en prenant comme concept "être étiqueté par une voyelle ou une fricative". Nous n'avons conservé dans chaque cas que les 5 premiers rapports, car les s.e.v. de dimension 5 obtenus portent plus de 90 % de l'inertie du nuage à représenter.

A chaque rapport correspond un descripteur binaire. Celui-ci est égal à un, si le rapport correspondant est supérieur à un. Comme à l'issue de ce pré-apprentissage la classe des fricatives est parfaitement caractérisée par ces cinq descripteurs, nous n'avons fait un sur-apprentissage binaire que sur les 12 voyelles. Chaque élément est décrit par 60 descripteurs binaires provenant des 12*5 rapports.

IV - RESULTATS

Reconnaissance	1°choix	2°choix	3°choix
ens d'apprentissage	72 %	88 %	93 %
ens de reconnaissance	58 %	76 %	84 %

Résultats de reconnaissance de la description par lignes spectrales.

Reconnaissance	1°choix	2°choix	3°choix
ens d'apprentissage	78 %	92 %	93 %
ens de reconnaissance	63 %	86 %	87 %

Résultats de reconnaissance de la description des maxima spectraux.

Les résultats présentés dans les deux tableaux ci-dessus montrent de meilleurs taux de reconnaissance pour la description des maxima en généralisation. Ces derniers sont améliorés par l'étape de pré-apprentissage qui permet de différencier les voyelles localement. La représentation par sous-espace vectoriel comporte elle aussi une étape de pré-apprentissage, mais comme celui-ci est de type statistique, il ne convient pas à notre problème de reconnaissance car nous possèdons peu d'objets par classe. Particulièrement, les classes contenant moins de 15 exemples dans l'ensemble d'apprentissage, obtiennent de très faibles résultats de reconnaissance.

Les résultats du 1er choix de la description des maxima pourraient être améliorés en forçant la décision de reconnaissance. En effet, seuls les objets reconnus par une étiquette sont comptés dans ce premier résultat, les objets bruités étant donc assimilés à des contestés. En fait, seulement 20 % des objets sont reconnus par une autre étiquette lors du premier choix. La reconnaissance sur les maxima spectraux présente l'avantage de limiter la hauteur du treillis phonétique en moyenne à 1,5 étiquette vocalique par objet, limitant ainsi la combinatoire lors de la reconnaissance.

CONCLUSION

Cette étude montre que la description des classes vocaliques par des formules logiques donne de bons résultats de reconnaissance sur une description simple du spectre. Une description binaire d'autres segments de la parole pourrait conduire à construire un système de reconnaissance acoustique par formules logiques.

Cet apprentissage offre l'avantage de nécessiter peu d'objets par classe à apprendre. Il est rapide pour la reconnaissance (vérification de 20 formules booléennes par classe), mais lent à l'apprentissage. L'adaptation au locuteur peut être facilement incluse dans l'apprentissage (cf.III.1).

Une autre utilisation de cette méthode est l'étude de critères de reconnaissance. Pour réaliser cela, plusieurs méthodes sont envisageables : expliciter les formules, examiner les descripteurs équivalents, ceux qui sont souvent employés ou ignorés dans les formules, regrouper les objets en divers groupes, vérifier la pertinence des indices acoustiques. Une première approche de cette utilisation a été montrée au § III.1 pour l'évaluation de groupes d'ambiguïté.

BIBLIOGRAPHIE

(1) BAUDRY M. : "Etude du signal de parole dans sa représentation amplitude - temps" Thèse d'Etat - Paris VI - 1978.

(2) CARATY M.J. : "Distance perceptive interspectrale" 14ème JEP - Paris - Juin 1985.

(3) DELEGLISE P. : "Paramétrisation et détermination des noyaux stationnaires en vue de la reconnaissance de la parole continue" Thèse 3ème cycle - Paris VI - 1983.

(4) ITAKURA F. et AL : "A hardware implementation of new narrow medium band speech coding" ICASSP 1982.

(5) OPPENHEIM A.V. and JOHNSON A.H. : "Discrete representation of signals" Proc. IEEE - Vol. 60 - June 1972.

(6) PIGOT H. : "Paramétrisation de la parole continue. Apprentissage binaire et reconnaissance des ilots vocaliques". Thèse 3ème cycle - Paris VI - 1985.

(7) SALLANTIN J. : "Méthodologie de l'apprentissage pour des variables binaires". Outils pour l'apprentissage. Coll. Publ. CNRS GR22 1983.

(8) SINGER D. : "Contributions à l'expression logique des mécanismes d'apprentissage" Thèse 3ème cycle - Paris VI - 1984.

(9) WATANABE S. : "Methodologies of pattern recognition". London Acad. Press. 1969.

DISTANCE INTERSPECTRALE A CRITERES PERCEPTIFS

M.J. CARATY X. RODET

UNIVERSITE PIERRE ET MARIE CURIE - INSTITUT DE PROGRAMMATION
Laboratoire de Reconnaissance de Formes et de Parole
4, Place Jussieu - 75230 PARIS CEDEX 05

ABSTRACT

Study of an interspectral distance based on perceptual criterions.

Spectral maxima, automatically extracted from LPC spectrum, parametrised by frequency, bandwith and magnitude, are used to estimate the interspectral distance.

An "interpeak" partial distance is computed from psycho-acoustic data, based on perceptual differential limens. The interspectral distance is then a linear combination of partial distances, defined according to a minimum spectral distortion criterion.

Application : continuous speech segments recognition.

I - CADRE D'ETUDE

La présente étude s'inscrit dans le projet d'un système de reconnaissance de parole continue.

Le projet est un système de reconnaissance basé sur des entités phonétiques VCV, considérant que les effets de coarticulation sont largement inclus dans ces entités.

Les états stables déterminent des points d'ancrage stratégiques, destinés à réduire la combinatoire des modèles VCV à comparer à l'entité à reconnaître.

Les états stables sont automatiquement extraits du signal de parole continue par un algorithme basé sur la variation d'énergie du signal. Ils sont étiquetés à l'écoute des courts segments de parole les contenant (5. PIGOT, DELEGLISE).

Ainsi identifiés sur une durée de parole de 3 mn 30 d'un locuteur-homme, 524 états stables (de type noyaux vocaliques) constituent notre corpus où l'on distinguera deux sous-ensembles disjoints :
- ensemble d'apprentissage : à partir duquel sera défini l'ensemble des "formes de référence"
- ensemble de reconnaissance : constitué des "formes inconnues" à reconnaître parmi les formes référencées.

II - PARAMETRISATION DU SIGNAL

1 - Choix de la représentation

Nombreuses sont les paramétrisations liées aux opérateurs de reconnaissance de type mesure de distance. Pour citer les plus classiques : valeurs de l'enveloppe spectrale, coefficients de prédiction linéaire, coefficients cepstraux, coefficients de réflection, etc...

D'autres paramétrisations s'orientent, et ce sera notre cas, vers des mesures basées sur la perception.

De nombreuses études soulignent l'importance des formants dans la perception des voyelles. Si une étude rapide de l'enveloppe spectrale ne permet pas de distinguer maxima spectraux et formants, la notion de proéminence spectrale n'en est pas moins intéressante de par son effet direct sur la perception du son. Des études montrent que les vallées spectrales ont peu d'effet sur la perception tandis que les maxima, tout du moins dans les "zones basse fréquence" ont un effet important (4. Klatt).

C'est ainsi que nous nous orientons vers une représentation spectrale basée sur la détermination des caractéristiques des maxima spectraux : fréquence centrale, largeur de bande et amplitude par analogie aux formants.

2 - Détermination des paramètres spectraux

Les caractéristiques des maxima spectraux sont déterminées à partir de l'analyse par prédiction linéaire d'ordre 16 du signal.

a - Fréquence centrale :
Un balayage du spectre LPC permet l'extraction des maxima $P^k, k \in \mathbb{N}$, et la détermination de f, leur fréquence centrale .

b - Largeur de bande à 3 dB :
A partir d'un maximum P^k est déclenchée, une recherche systématique gauche et droite des points d'inflexion (I_g^k, I_d^k) et des points spectraux (L_g^k, L_d^k) correspondant à une chute de 3 dB relative à P^k.
On distinguera alors les cas suivants :
- L_g^k défini, L_d^k défini : la largeur de bande l est automatiquement déduite.
- L_g^k défini, L_d^k indéfini (resp. L_g^k indéfini, L_d^k défini) ; on considèrera que le point L_g^k (resp. L_d^k) définit la demi-largeur de bande, l/2.
- L_g^k indéfini, L_d^k indéfini : la largeur de bande l sera déterminée à partir de l'interpolation parabolique passant par les trois points (L_g^k, P^k, L_d^k).

c - Amplitude:
L'amplitude a est calculée en dB, en prenant

pour référence le maximorum du spectre à 0 dB.

Cet algorithme a été appliqué à un suivi de trajectoire de formants à l'IRCAM (6. RODET, DEPALLE).

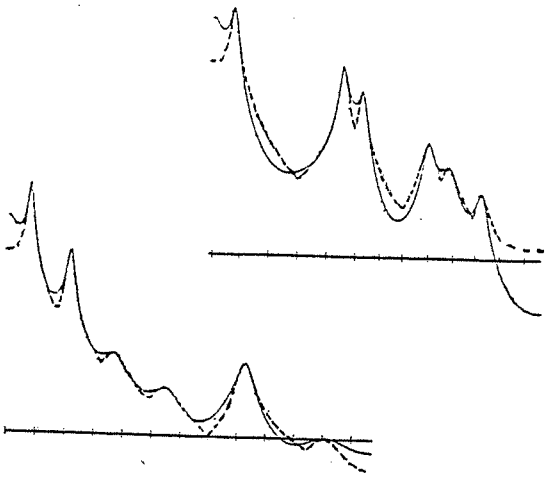


Figure 1 — Spectre LPC
-- Simulation du spectre à partir des fonctions d'onde formantique déduites des caractéristiques spectrales.

III - DEFINITION DE LA MESURE DE DISTANCE PERCEPTIVE

La mesure de distance va nous permettre de mesurer à partir des maxima, la déformation spectrale minimale du "spectre référence" S_r, que l'on considèrera si on admet S_r comme représentant du "spectre test" S_t.

Les deux spectres à comparer S_r et S_t sont définis par l'ensemble des caractéristiques de leurs maxima :

$$S_r=\{P_r^j(f_r^j,l_r^j,a_r^j),j \in \mathbb{N}\},S_t=\{P_t^k(f_t^k,l_t^k,a_t^k),k \in \mathbb{N}\}$$

L'idée est de chercher une déformation minimale de S_r, établissant une correspondance de ses pics à ceux de S_t. Une mesure de distance "interpic" est donc introduite faisant intervenir des mesures de déformation liées aux écarts en fréquence, largeur de bande et amplitude (e_f,e_l,e_a) :

$$d(P_t,P_r)=p_f e_f(P_t,P_r)+p_l e_l(P_t,P_r)+p_a e_a(P_t,P_r)$$

avec :

$$e_f(P_t,P_r)=2\left|\frac{f_r-f_t}{f_r+f_t}\right|,e_l(P_t,P_r)=2\left|\frac{l_r-l_t}{l_r+l_r}\right|,$$

$$e_a(P_t,P_r)=|a_r-a_t|$$

Les coefficients de pondération p_f,p_l,p_a sont respectivement liés à la fréquence, largeur de bande et amplitude. Ils résultent du produit de deux sortes de facteurs :

- $(\omega_f,\omega_l,\omega_a)$: tenant compte des connaissances sur la perception sonore.
- (ρ_f,ρ_l,ρ_a) : à titre de normalisation de l'influence respective de la fréquence, de la largeur de bande et de l'amplitude sur la distance globale.

Ainsi :

$p_f=\rho_f\omega_f$, $p_l=\rho_l\omega_l$ et $p_a=\rho_a\omega_a g(a)$
(cf. §c pour g(a)).

a - Détermination de ω_f :
Les proéminences des "spectres référence" sont réparties dans des classes F_1,F_2,F_3,F_4 correspondant approximativement à la numérotation habituelle des formants, et auxquelles sont attribuées des valeurs reflétant leur importance perceptuelle supposée. Nous avons choisi :

P^r_ϵ	F1	F2	F3	F4
ω_f	8	10	3	1

b - Détermination de ω_a :
Tenant compte des seuils différentiels perceptuels (2. Flanagan) et en extrapolant ces valeurs pour F_3 et F_4, nous choisissons :

P^r_ϵ	F1	F2	F3	F4
ω_a	1/1.5	1/3	1/7.5	1/15

c - Détermination de g(a) :
On accordera plus d'importance à la déformation des proéminences dont l'amplitude est élevée, d'où notre choix pour g(a) : g(a)=1+(a/50)

d - Détermination de ω_l:
L'amplitude d'un formant ayant tendance à varier comme l'inverse de sa largeur de bande, nous prendrons $\omega_l = \omega_a$

e - Détermination de ρ_f,ρ_l,ρ_a :
Nous prenons ρ_f,ρ_l et ρ_a de telle sorte que dans chaque cas une variation égale au seuil différentiel perceptuel apporte la même contribution à la distance "interpic".

ρ_l étant choisi arbitrairement à 1, nous obtenons :

P^r_ϵ	F1	F2	F3	F4
ρ_f	8	7	22	67
ρ_a	0.13	0.13	0.13	0.13
ρ_l	1	1	1	1

D'où les pondérations p_f,p_a et p_l :

P^r_ϵ	F1	F2	F3	F4
p_f	67	67	67	67
p_a	1.3	0.65	0.26	0.13
p_l	10	5	2	1

La distance interspectrale

A tout pic de S_r, on fait correspondre le pic de S_t, pour lequel la distance interpic est minimale. Une correspondance de même type est établie pour tout pic de S_t non associé.

La distance interspectrale finale est alors la moyenne d'une sélection des distances partielles précédemment déterminées.

On suppose que tout spectre S_r admet un "motif invariant" dans la classe qu'il représente, défini par la présence des n premiers pics de S_r.

La moyenne des distances partielles définie à partir du "motif invariant" est calculée (pénalisant ainsi apparition ou disparition de pic). Les distances partielles associées aux pics restants de S_r n'interviendront dans la distance interspectrale que si elles diminuent la moyenne déjà calculée.

IV - RESULTATS

Douze voyelles permettent d'étiqueter le corpus étudié :

code phonétique :

|ɛ| |e| |ã| |o| |õ| |ɛ̃| |a| |œ| |i| |o| |y| |u|

code machine :

( ) * + / > A E I O V W

Après analyse des histogrammes de fréquences, largeurs de bande et amplitudes des spectres de l'ensemble d'apprentissage, quatre classes ont été séparées. Leur histogramme de fréquences ne respectant pas une distribution gaussienne, on effectuera les séparations suivant un seuil de fréquence :

- du 2ème pic des classes |ã| , |õ| et |i|.
- du 1er pic de la classe |œ|.

Pour chaque classe, on définira un spectre moyen à partir de l'ensemble d'apprentissage, par moyenne statistique des caractéristiques des pics ayant même numéro. Ce spectre sera choisi comme spectre référence.

La "forme invariante" d'une classe sera déterminée en faisant varier n, le nombre de pics la définissant, et en choisissant pour n la valeur correspondant au taux le plus élevé de reconnaissance intraclasse et extraclasse.

Résultat du treillis de reconnaissance à l'ordre 1, 2 et 3 des ensembles d'apprentissage et de reconnaissance.

ORDRE	1	2	3
E. APP.	87.6%	95.9%	98.6%
E. REC.	80.3%	93.2%	98.3%

Matrice de reconnaissance à l'ordre 2 de l'ensemble de reconnaissance :

	R(()	R())	R(*)	R(+)	R(/)	R(>)	R(A)	R(E)	R(I)	R(O)	R(U)	R(W)
39(	38	1	-	-	-	-	2	29	-	-	-	-
	0.97	0.23	-	-	-	-	0.35	0.74	-	-	-	-
25)	3	23	-	-	-	-	-	7	14	-	9	-
	0.12	0.92	-	-	-	-	-	0.28	0.56	-	0.12	-
13*	-	-	12	8	5	-	-	1	-	-	-	-
	-	-	0.92	0.62	0.38	-	-	0.08	-	-	-	-
5+	-	-	-	5	-	-	-	1	-	4	-	-
	-	-	-	1.00	-	-	-	0.25	-	0.80	-	-
9/	-	1	4	-	5	-	-	-	2	3	-	1
	-	0.11	0.44	-	0.56	-	-	-	0.22	0.33	-	0.11
10>	5	-	-	-	-	8	5	2	-	-	-	-
	0.50	-	-	-	-	0.80	0.50	0.20	-	-	-	-
39A	9	-	3	11	-	3	39	13	-	-	-	-
	0.23	-	0.08	0.28	-	0.08	1.00	0.33	-	-	-	-
53E	23	11	-	8	-	1	-	51	-	4	8	-
	0.43	0.21	-	0.15	-	0.02	-	0.96	-	0.08	0.15	-
19!	-	3	-	-	-	-	-	-	18	-	12	5
	-	0.16	-	-	-	-	-	-	0.95	-	0.63	0.26
13O	-	-	-	1	-	-	-	3	1	11	-	10
	-	-	-	0.08	-	-	-	0.23	0.08	0.85	-	0.77
8U	-	-	-	-	-	-	-	2	6	-	7	1
	-	-	-	-	-	-	-	0.25	0.75	-	0.88	0.12
1W	-	-	-	-	-	-	-	-	1	-	-	1
	-	-	-	-	-	-	-	-	1.00	-	-	1.00

Interprétation du 1er élément de la matrice : sur 39 '(' contenus dans l'ensemble de reconnaissance, 38 sont reconnus comme '(', le taux de reconnaissance associé est 0,97.

On remarque une diminution du pourcentage de la reconnaissance globale sur l'ensemble de reconnaissance très sensible à l'ordre 1.

Le taux de reconnaissance du '/' : 0,22 à l'ordre 1, 0,56 à l'ordre 2, 0,89 à l'ordre 3 pénalise fortement la reconnaissance ; une étude particulière est à envisager pour trouver un meilleur représentant de la classe.

V - CONCLUSION

Les coefficients de pondération sont la base de la distance, et leur détermination est bien trop imprécise dans notre étude.

Des études de la perception sonore plus adaptées devraient nous permettre d'optimiser ces coefficients.

Le critère de sélection des distances partielles, lié au choix de la "forme invariante", est très discutable, et loin d'être optimal. Il en est de même pour le critère de sélection des références.

La distance présente néanmoins des avantages :

- facile à mettre en oeuvre,
- souplesse du choix des références (réelles ou synthétiques),
- peu de données sont nécessaires à l'adaptation au locuteur,
- permet une reconnaissance rapide grâce à la méthode et au nombre limité de références (12 classes, 16 représentants).

Diverses applications :

- étiquetage automatique des noyaux vocaliques extraits de la parole continue,
- identification du locuteur
- compression d'état stable de la parole.

BIBLIOGRAPHIE

(1) DELEGLISE P. : "Paramétrisation et détermination des noyaux stationnaires en vue de la reconnaissance de la parole continue" Thèse de 3ème cycle, Paris VI, 1983.

(2) FLANAGAN J.L. : "Speech analysis synthesis and perception" Springer Verlag Berlin Heidelberg New-York, 1972.

(3) GRAY R.M. : "Distortion measures for speech processing" IEEE Trans. Acoust. Speech and Signal Processing, août 1980.

(4) KLATT D. : "Prediction of perceived phonetic distance from critical-band spectra : a first step" IEEE ICASSP, Paris 1982.

(5) PIGOT H., DELEGLISE P., : "Apprentissage binaire des noyaux stationnaires" 14ème JEP Paris, juin 1985.

(6) RODET X., DEPALLE P. : "Synthesis by rule : formants and LPC diphones" IEEE ICASSP Tampa (Florida) mars 1985.

(7) VISWANATHAN R., MAKHOUL J., RUSSEL W. : "Towards perceptually consistent measures of spectral distance" ICASSP 1981.

LE POUVOIR DISCRIMINATIF D'INDICES DE DYSPHONIE CALCULES A PARTIR DE VOYELLES SOUTENUES ET DE PHRASES ISOLEES

Jean SCHOENTGEN

Institut de Phonétique, Université Libre de Bruxelles

RESUME

L'objectif de l'étude est de comparer les pouvoirs discriminatifs respectifs d'indices acoustiques extraits à partir de voyelles soutenues et de courtes phrases. A cet effet, un échantillon de la voyelle /a/ et un échantillon de la phrase /c'est une saga/ ont été numérisés pour chaque membre d'un ensemble de locuteurs normaux et dysphoniques. Les deux indices extraits sont le quotient de perturbation de la fondamentale (PPQ) et le facteur de perturbation de Lieberman (FPL). Le prétraitement appliqué est l'extraction du critère de variation d'amplitude (CVA). Les résultats indiquent que les indices extraits à partir de la parole continue ne sont pas systématiquement plus discriminatifs que ceux extraits à partir de voyelles soutenues. Par ailleurs, lorsque le seuil de décision intervenant dans le facteur de perturbation de Lieberman est modifié dans le sens d'une adaptation à la durée de la période fondamentale, son pouvoir discriminatif s'en trouve significativement amélioré.

INTRODUCTION

Cette étude concerne la description objective de l'incidence des pathologies laryngées sur la représentation acoustique du signal de parole. Cette description se veut quantitative, sur base d'un ensemble d'indices acoustiques extraits à partir du signal de parole. Ce genre d'analyse procède en quatre étapes successives :

1) l'acquisition et la numérisation du signal de parole
2) le prétraitement
3) l'extraction des indices de dysphonie
4) l'évaluation des performances de discrimination de ces indices

Les objets étudiés sont des échantillons de signaux de parole émis par des locuteurs normaux et dysphoniques. Le rôle du prétraitement est d'amplifier les aspects du signal de parole qui sont propres à la source vocale et d'atténuer ceux qui sont propres au conduit. Sur le plan pratique l'objectif de cette opération est d'améliorer le pouvoir discriminatif de la chaîne d'analyse.

Les indices extraits appartiennent à la famille des indices de microperturbations. Ils quantifient les déviations de l'amplitude et de la période du signal par rapport à la périodicité stricte. Comme tout travail touchant à la qualité de la voix, l'étude des pathologies laryngées au travers du signal de parole soulève un certain nombre de questions délicates quant à la constitution des corpus et au choix des signaux à analyser. En ce qui concerne des derniers, deux écoles de pensée coexistent. La première, largement majoritaire, estime que la voyelle isolée contient suffisamment d'informations et parmi les plus utiles (1,2,3). La seconde école, minoritaire, affirme que la parole continue est, de par son caractère dynamique, davantage en mesure de faire ressortir les anomalies de la phonation dans le signal. Dans son ouvrage de synthèse sur l'examen clinique de la voix, Hirano (4) note que les parties stationnaires de voyelles soutenues devraient servir comme échantillons standards pour l'analyse acoustique, mais que certaines conditions pathologiques seraient plus apparentes durant les phases de transitions de la phonation. Les observations que nous avons pu faire lors des contacts avec les patients, observations qui ne sont pas nécessairement représentatives, suggèrent que les deux situations opposées sont possibles : celle où la perturbation est plus manifeste lors de la production de la parole continue et la situation inverse où les voyelles soutenues apparaissent plus perturbées.

Dans le but de tester expérimentalement les arguments des deux écoles en présence, nous avons enregistré à la fois des voyelles soutenues (/a/ /e/ /i/) et une phrase isolée (/c'est une saga/) émises par le même ensemble de locuteurs normaux et dysphoniques. Nous pouvions ainsi comparer le pouvoir séparateur des indices obtenus sous les deux conditions. S'il est admis que les microperturbations apparaissent amplifiés dans les transitions voisées/non voisées, ou dans les transitions rapides entre formants, ceci n'implique pas nécessairement que les performances de discrimination soient également amplifiées. Le critère de variation d'amplitude a servi de prétraitement.

METHODE

Corpus

Tous les locuteurs normaux (21 ♂, 22 ♀), francophones, âgés d'environ 25 ans, étaient étudiants à l'Université Libre de Bruxelles. Ils ont soutenu les voyelles /a/, /e/, /i/ pendant une

durée de l'ordre de deux secondes et ont prononcé 2 phrases, dont /c'est une saga/(*). Ils avaient pour consigne d'articuler distinctement, d'une façon affirmative et sens coloration paralinguistique. Les locuteurs dysphoniques (16♂ , 21♀) étaient des patients, soit au Service O.R.L. de la Clinique Universitaire St Pierre à Bruxelles, soit du Centre d'Audiophonologie Paul Guns (Clinique Universitaire St Luc). Le diagnostic a été établi par les médecins attachés à ces Services. Les locuteurs pathologiques ont émis, dans des conditions semblables, les mêmes voyelles et les mêmes phrases que les locuteurs normaux.

L'approche la plus globaliste serait de considérer la manifestation des pathologies laryngées dans le signal de parole comme l'une des multiples facettes possibles de la qualité de la voix. Celle-ci est déterminée par des influences aussi diverses que le sexe du locuteur, son âge, des facteurs géographiques, linguistiques et sociologiques , dont(particulièrement importante dans le cadre de notre sujet) les habitudes de consommation d'alcool et de tabac. Il est donc illusoire de considérer un quelconque corpus comme un échantillon représentatif d'une population dont on serait bien en peine de la situer dans cet espace multidimensionnel.

En pratique, l'expérimentateur compare le groupe des malades à un échantillon de locuteurs normaux assez homogène et composé de sujets jeunes.

Acquisition et numérisation du signal

Les locuteurs normaux ont prononcé voyelles et phrases deux fois de suite. Les locuteurs dysphoniques trois fois de suite. La vitesse d'enregistrement était de 38 cm/sec. Le microphone (SHURE 5455) était placé à une distance de 40 cm de la bouche pour les locuteurs normaux et à une distance variant entre 20 et 40 cm pour les locuteurs dysphoniques. L'enregistreur était du type NAGRA 4.2 et les enregistrements ont été effectués en chambre audiométrique. Parmi toutes les voyelles /a/ prononcées par chaque locuteur, la plus longue a été retenue pour numérisation. Les voyelles d'une durée inférieure à 0.7 secondes ont été rejetées d'office. Avant le codage, les voyelles retenues ont été préfiltrées passe-bas à l'aide d'un filtre elliptique. La fréquence d'échantillonnage était de 10 kHz et le nombre de niveaux de discrétisation de 4096 (12 bits). Le niveau du signal a été ajusté de manière à utiliser toute la dynamique des convertisseurs. Pour chaque locuteur, un échantillon de la phrase /c'est une saga/ a été numérisé. Avant la numérisation il a été préfiltré à 3 kHz à l'aide d'un filtre elliptique. La fréquence d'échantillonnage était de 6666 Hz (12 bits).

(*) contient une majorité de sons voisés, notamment l'occlusive /g/, dont la production impose des contraintes aérodynamiques sévères au système vocal.

Prétraitement

Lors d'une étape antérieure de notre étude, le critère de variation d'amplitude (CVA) s'est révélé le prétraitement le plus efficace comparativement au signal de parole non traité et au signal résiduel. C'est donc ce type de prétraitement que nous avons appliqué aux échantillons retenus.

Rappelons que cette méthode (5,6,7) consiste à détecter des discontinuités dans le signal, liées au fonctionnement de la source laryngée. Le brusque accroissement de l'énergie sonore suite à la fermeture de la glotte et la décroissance du signal de parole en raison des pertes dans l'appareil vocal, impriment à l'enveloppe lissée du signal d'un son voisé une forme en dents de scie. Le principe de calcul du CVA est le suivant : le signal de parole est filtré passe-bande (Fc = fréquence centrale du filtre, largeur du filtre = 400 Hz), ensuite l'enveloppe est estimée, lissée et enfin différenciée. Le tableau ci-dessous reprend les valeurs retenues pour les différents paramètres :

	/a/		/c'est une saga/	
	masc.	fém.	masc.	fém.
fréquence centrale du filtre passe-bande (en Hz)	700	900	650	650
demi-longueur de la fenêtre de lissage (en nombre de pts)	17	17	15	9
intervalle de différenciation (en nombre de pts)	21	21	21	15

La figure 1 montre le CVA calculé à partir de la représentation acoustique de /c'est une saga/ pour un locuteur féminin normal. Un détecteur d'événements est ensuite appliqué. Il repère l'emplacement de toutes les crêtes significatives du CVA. Une crête est jugée significative lorsqu'elle signale le début d'un nouveau cycle glottique. L'ensemble des emplacements de toutes les crêtes détectées au sein d'un segment voisé, est ordonné dans un vecteur. La position dans le temps des crêtes est précisée à l'aide d'une interpolation parabolique et la différence entre l'emplacement de deux crêtes adjacentes est assimilée à la durée instantanée de la période fondamentale. Précisons que le but n'est pas de faire le suivi de la mélodie au sein de la phrase. L'objectif du traitement consiste plutôt en la détermination de la durée de chacune des périodes.

Indices acoustiques (8,9)

Pour la plupart, les indices acoustiques étudiés sur des phrases isolées concernent les microperturbations de la fondamentale. Concrètement, deux indices ont été retenus. Le premier est le quotient de perturbation de la fondamentale (PPQ). Cet indice tient compte de toutes les périodes au

sein d'un segment voisé et évalue leurs déviations instantanées par rapport à la mélodie. Le deuxième est le facteur de perturbation de Lieberman (FPL) qui comptabilise toutes les perturbations brutes plus importantes que 0.5 millisec. Comme on s'attend à observer les perturbations les plus importantes dans les transitions V/NV et dans les transitions rapides entre formants, le facteur de perturbation semble donc plus spécifiquement sensible au caractère dynamique de la parole. En effet, nous savons que les microperturbations dans la partie quasi-stationnaire d'une voyelle soutenue sont de l'ordre de 1 % ; une durée de 0.5 millisec, par contre, est de l'ordre de 5 à 10 % de la durée de la période fondamentale masculine. Le quotient de la perturbation de la fondamentale est calculé pour chaque segment voisé, dont le nombre varie entre 3 et 4, selon que l'algorithme a décidé que les vibrations des cordes vocales ont cessé lors de l'occlusion (/g/) ou non ; la médiane des 3 ou 4 valeurs a été retenue comme représentative.

Évaluation du pouvoir séparateur

Le pouvoir discriminatif est évalué par la valeur d'un coefficient de mérite. Le "facteur de qualité" proposé (10) admet des valeurs entre 0 et 1. Il est déterminé en plaçant une frontière de décision au sein des valeurs réalisées par un indice sur la totalité des locuteurs normaux et dysphoniques de manière à maximiser le facteur de qualité. Lorsque la séparation entre les deux groupes (normaux et dysphoniques) est parfaite, alors FQ = 1. Si tous les locuteurs sont regroupés dans une seule classe ou répartis de manière aléatoire entre les deux classes, alors FQ = 0.

RESULTATS

Le tableau 1 résume les résultats obtenus à la fois sur les voyelles et sur la phrase pour les locuteurs masculins et féminins (Rappel : la valeur du facteur de qualité sera d'autant plus élevée que l'indice est discriminatif).

- Analyse de la phrase /c'est une saga/

Les résultats obtenus sur le corpus féminin confirment nos hypothèses : l'indice FPL, particulièrement adapté à l'analyse de la parole continue, l'emporte en valeur absolue sur PPQ. Le FPQ est sensible aux perturbations importantes qui ont lieu lors des transitions entre les segments nv/v, tandis que le PPQ est calculé au sein de chaque segment voisé dont chaque période (se trouvant dans la zone stationnaire ou transitoire) contribue à la valeur de l'indice. Les résultats obtenus sur le corpus masculin sont par contre surprenants a priori. Ces résultats s'expliquent néanmoins si on prend en compte les résultats d'études entreprises après la publication de l'article de Lieberman en 1963. Ces études montrent, en effet, que les perturbations de la fondamentale sont plus importantes en valeur absolue pour les hommes que pour les femmes, du fait qu'elles évoluent avec la fréquence fondamentale. Etant donné que le FPL fait preuve de bonnes performances à partir de voix féminines, indication que le seuil est bien adapté, nous l'avons doublé (de 0.5 à 1 milliseconde) pour les voix masculines. Le coefficient de mérite associé à la suite de cette modification du seuil, s'est alors élevé de 0.39 à 0.61.

- Comparaison entre voyelle soutenue et phrase isolée

Le tableau 1 indique que les performances de l'indice PPQ (/a/) dépassent, au moins dans la plage de fréquence considérée, les performances des indices de perturbation calculés à partir de /c'est une saga/. En supprimant dans l'algorithme de suivi de la fondamentale, le contenu énergétique de l'intervalle d'analyse comme un des critères de détection d'un pic, le pouvoir discriminatif du FPL pour le corpus féminin augmente jusqu'à égaler celui de l'indice PPQ extrait à partir de la voyelle soutenue. Cependant, le pouvoir discriminant du PPQ (phrase) s'avère alors très sensible au choix des paramètres d'analyse (intervalle de différenciation et fenêtre de lissage) et les performances de discrimination chutent nettement par suite de nombreuses erreurs du type 1, c'est-à-dire que de nombreuses crêtes sont déclarées significatives, tout en ne l'étant pas. Il s'est avéré préférable d'implanter la version la plus robuste de l'algorithme de suivi de TO et de se contenter d'une discrimination légèrement réduite. Il ressort du tableau 1 que les algorithmes dont nous disposons (et dans le cas précis de nos corpus), n'ont pas réussi à extraire, à partir d'une phrase isolée des indices plus discriminatifs que ceux extraits à partir d'une voyelle soutenue. Nous pensons pouvoir extrapoler, sans trop de risques, les résultats de cette étude, à d'autres corpus et d'autres phrases. Bien sûr, nous ne prétendons pas qu'il soit exclu de développer un algorithme réalisant sur de la parole continue des performances systématiquement supérieures (on en trouve dans la littérature). Mais l'étude de ces travaux, comparativement à nos résultats, nous a convaincu qu'il est possible, pour une performance donnée sur de la parole continue, de concevoir un algorithme qui réalise les mêmes performances sur des voyelles, et ceci pour un choix de corpus raisonnable.

DISCUSSION ET RESUME

Nous pouvons conclure que nos résultats démentent les argumentations en faveur d'une utilisation préférentielle d'échantillons de parole continue. Ils infirment l'hypothèse d'Askenfelt et al (11) selon laquelle une voyelle soutenue n'est représentative de la qualité de la voix qu'en présence d'une pathologie grave. Une des principales raisons pour lesquelles nos conclusions ne rejoignent pas celles d'Askenfelt et al, est liée aux différences dans les procédures d'analyse : les procédures dont nous disposons permettent de rendre compte du fait que les perturbations détectables dans des voyelles sont en moyenne plus faibles en valeur absolue que celles exhibées par de la parole continue. Les valeurs de PPQ calculées partir d'une phrase et d'une voyelle, par exemple,

différent dans leur ordre de grandeur. Même notre algorithme de calcul du CVA, qui est le plus contraignant quant aux largeurs de bande, permet encore d'explorer des régions fréquentielles dont les bornes dépassent d'un facteur 10 celles qui peuvent être atteintes par les méthodes (assez extrêmes) de Gubrynowicz et al (12) ou Askenfelt et al. Rappelons enfin qu'une des raisons ayant motivé le choix de la phrase /c'est une saga) est la présence de l'occlusive /g/ qui impose des contraintes aérodynamiques sévères sur la source lors de sa production. Cependant, l'inspection des signaux de parole a montré que les locuteurs dysphoniques, dans l'ensemble n'ont pas rencontré de difficultés particulières à produire l'occlusive. Il semble que des difficultés de production accrues soient observées uniquement dans le cas où /g/ se trouve en position initiale, au début d'une syllabe par exemple. Dans un contexte intervocalique c'est l'occlusive non voisée /k/ qui présenterait un maximum de difficultés (13).

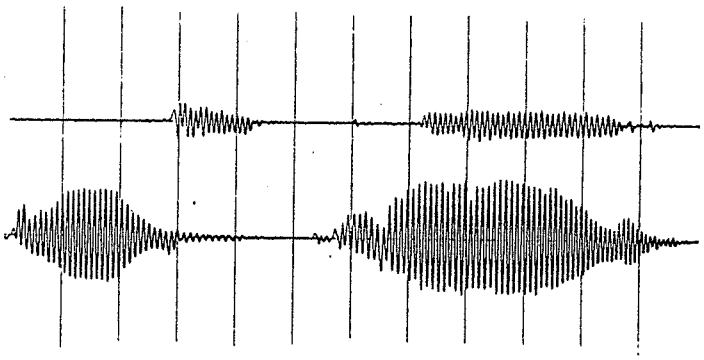


Figure 1: CVA (/c'est une saga/) locuteur féminin normal; durée entre deux marques de temps: 0.04285 sec

Tableau 1 : Facteurs de qualité

	/a/ (♀)	/c'est une saga/ (♀)	/a/ (♂)	/c'est une saga/ (♂)
PPQ	0.66	0.52	0.66	0.61
FPL	/	0.61	/	0.39

BIBLIOGRAPHIE

(1) Koike Y., "Application of some acoustic measures for the evaluation of laryngeal dysfunction", Studia Phonologica, VII, 1973, pp 17-23.

(2) Koike Y., Takahashi A., Calcaterra T.C., "Acoustic measures for detecting laryngeal pathology", Ann. Otol., 84, 1975, pp. 117-123.

(3) Kasuya K., Kobayashi Y., Kobayashi T., "Characteristics of pitch period and amplitude perturbations in pathologic voice", Proceedings ICASSP 83, Boston, pp 1372-1375.

(4) Hirano M., "The Clinical Examination of Voice", Springer Verlag, New York, 1981.

(5) Jospa P., "Critères de variation d'amplitude", Rapport d'activités de l'Institut de Phonétique de l'ULB, 17, 1982, pp 89-108.

(6) Jospa P. et Schoentgen J., "Signal acoustique, signal résiduel, critère de variation d'amplitude à court terme : trois supports au calcul d'indices de dysphonie", Proc. FASE/DAGA, Goettingen, 1982, pp 993-996.

(7) Jospa P., "Détection synchrone du pitch par le critère de variation d'amplitude", Comptes-rendus 13èmes Journées d'Étude du Groupe de la "Communication Parlée", GALF, 1984, Bruxelles, pp 161-162.

(8) Davis St. B., "Acoustic Characteristics of Normal and Pathological Voices", in "Speech and Language, Advances in Basic Research and Practice", N.J. Lass (Ed.), 1, Academic Press, New York, 1979, pp 273-335.

(9) Lieberman Ph., "Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathological Larynges", J. of the Acoust. Soc. of America, 35, 3, 1963, pp 344-352.

(10) Schoentgen J., "Quantitative Evaluation of the Discrimination Performance of Acoustic Features in Detecting Laryngeal Pathology", Speech Comm., 1, 1982, pp 269-282.

(11) Askenfelt A., Hammarberg B., "Speech Waveform Perturbation Analysis Revisited", STL-QPSR, 4, 1981, pp 40-49.

(12) Gubrynowicz R., Mikiel W., Zarnecky P., "An Acoustic Method for the Evaluation of the State of the Larynx Source in cases involving Pathological Changes of the Larynx", Archives of Acoustics, 5, 1, 1980, pp 3-30.

(13) Serniclaes W., Communication personnelle, 1984.

PROTHESE COCHLEAIRE : TRAITEMENT DE SIGNAL POUR UN SYSTEME MULTICANAL.

J. Génin* & R. Charachon+

*Centre National d'Etudes des Télécommunications 38243 Meylan

+Centre Hospitalier Régional et Universitaire 38043 Grenoble

ABSTRACT.

In a multi-channel cochlear prosthesis, several electrodes are inserted close to the patient's cochlea. The electrical stimulation of different parts of the cochlea evokes different sound sensations. This is used to provide the deaf patient with speech spectral information.

The most helpful spectral description is provided by filter bank analysis. Though, the electrical interactions between electrodes located close to one another bring severe limitations to signal level and spectral definition.

This work adds to common filter design some items improving it by forecasting electrical interactions.

I. INTRODUCTION.

La prothèse cochléaire s'adresse à des patients sourds totaux auxquels elle redonne des perceptions de nature auditive grâce à l'excitation électrique d'électrodes implantées à demeure au voisinage de terminaisons restées actives du nerf auditif.

Sans rentrer dans tous les aspects techniques de cette méthode thérapeutique, nous classerons les solutions mises en oeuvre par les différents laboratoires travaillant sur le sujet en deux groupes principaux.

- les systèmes à électrodes multiples utilisent le fait que l'excitation au moyen d'impulsions électriques de sections différentes de la cochlée ou de différents groupes de fibres du nerf auditif provoque chez le patient la perception de sonorités de hauteurs spectrales différentes. Ils visent ainsi à présenter au patient l'information spectrale portée par le signal acoustique (Fardeau, 1979). Les travaux déjà anciens effectués sur les codage, transmission, reconnaissance et synthèse de la parole amènent à penser qu'une analyse spectrale au moyen d'une douzaine de filtres passe-bande fournit une quantité d'information suffisante à l'intelligibilité de la parole.

Cependant, rien ne prouve a priori que le nerf auditif véhicule correctement l'information fournie par l'excitation d'une douzaine d'électrodes ni que le cerveau soit en mesure de l'exploiter. Sans aller plus avant, une limitation évidente de cette méthode apparaît dans les inévitables interactions électriques entre ces électrodes placées au voisinage les unes des autres dans un milieu conducteur. Ces interactions perturbent l'intensité des impulsions électriques émises, modifient leur équilibre relatif et peuvent rendre un signal complexe totalement inutilisable.

Pour contourner cette difficulté qui illustre une certaine impossibilité à "passer" sur ce canal toute l'information spectrale contenue dans le signal incident, on peut rechercher dans la représentation formantique, un codage plus efficace (Clark, 1982).

- Les systèmes à électrode unique refusent le risque de recontrer les défauts ci-dessus et se contentent de fournir au patient un signal unique lui apportant comme information l'intensité, peut-être la fréquence fondamentale, voire quelque indication sur le timbre du signal incident... (Casals, 1985) (Moore, 1984).

II. INTERACTIONS ENTRE ELECTRODES.

La mise en évidence expérimentale des phénomènes d'interaction entre les différentes électrodes intervient immédiatement lors de la mise en oeuvre d'un système de prothèse à électrodes multiples.

Du fait des innombrables facteurs impondérables qui interviennent lors de la mise en place chirurgicale des électrodes, du fait de la plus ou moins bonne "qualité" des fibres nerveuses rencontrées chez le patient, les électrodes peuvent présenter une efficacité très différente d'un patient à un autre et même pour les différentes électrodes d'un même patient. Compte tenu des phénomènes neurophysiologiques mis en jeu

et des contraintes techniques rencontrées, la plupart de ces systèmes utilisent des impulsion électriques courtes répétées à une fréquence de l'ordre de 300 Hz. Le paramètre "excitant" est alors la quantité d'électricité délivrée par chacune de ces impulsions. Les perceptions les plus faibles sont obtenues au voisinage du seuil pour des quantités d'électricité de 1 à 40 nC, les perceptions "inconfortables" pour des quantités d'électricité doubles ou triples.

Ces seuils de perception sont raisonnablement stables dans le temps mais sont considérablement réduits si deux ou plusieurs électrodes sont excitées simultanément (Chouard, 1981) (Génin, 1984).

La figure 1 montre les seuils de perceptior tels que l'on peut les obtenir en excitant simultanément deux électrodes. Le niveau d'excitation de l'une des électrodes est placé en abscisse, celui de l'autre est placé en ordonnée. Les points représentant les seuils de perception relevés se situent au voisinage d'une courbe rencontrant les axes aux points correspondant aux seuils de perception des électrodes lorsqu'elles sont excitées une seule à la fois.

Pour les "bonnes" paires d'électrodes, cette courbe s'approche des segments AC et BC, pour les "mauvaises", elle s'approche de la diagonale AB.

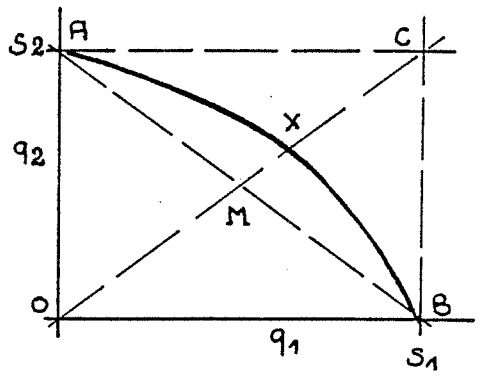


Figure 1. Etudes des interactions entre électrodes.

Ce résultat s'explique par le fait qu'électrodes et fibres nerveuses ne sont pas associées de façon exclusive mais par l'intermédiaire d'un milieu conducteur plus ou moins homogène et isotrope. En tenant le comportement électrique de ce milieu pour linéaire on peut considérer que, si deux électrodes 1 et 2 délivrent des impulsions électriques q_1 et q_2, les groupes de fibres auxquelles elles s'adressent reçoivent des "excitations" :

$$e_1 = A_{11} \cdot q_1 + A_{12} \cdot q_2$$

$$e_2 = A_{12} \cdot q_1 + A_{22} \cdot q_2$$

où A_{ij} sont tous des coefficients positifs.

Si les seuils des deux électrodes sont s_1 et s_2 respectivement, on obtient les seuils composites pour toutes valeurs q_1 et q_2 telles que :

$$A_{11} \cdot q_1 + A_{21} \cdot q_2 > A_{11} \cdot s_1$$

ou

$$A_{12} \cdot q_1 + A_{22} \cdot q_2 > A_{22} \cdot s_2$$

De toute évidence ce modèle ne rend pas compte de toute la réalité des phénomènes qui interviennent au voisinage du seuil de perception. Lorsque la quantité d'électricité émise par les électrodes se déplace de l'une à l'autre, les groupes de fibres qui répondent ne sont pas les mêmes mais se déplacent également d'une électrode à l'autre. Cependant on peut en première approximation tenir ce modèle pour représentatif de ce qui se passe aux niveaux d'excitation moyens et rechercher par les expériences de seuils une estimation des coefficients A_{ij} (les expériences de perception isosoniques à niveau moyen sont plus difficiles à mettre en oeuvre et leurs résultats sont beaucoup plus imprécis).

Supposons le système symétrique :

$$A_{11} = A_{22} = A \qquad \text{et} \qquad A_{12} = A_{21} = B$$

et normé par rapport à A (A=1). Appliquons le au seuil composite recherché sur la diagonale OC, on trouve :

$$B = s_1/q_1 - 1 = CX/OX$$

Nous n'avons pas pu à ce jour mener ce genre d'expériences auprès d'un grand nombre de patients. Cependant les relevés relativement complets effectués auprès d'un patient ont donné les résultats ci-dessous. Le tableau 1 présente les valeurs moyennes obtenues pour B en fonction de la distance qui sépare les électrodes des paires considérées. Cette distance est comptée en nombre d'électrodes intermédiaires. Pour les résultats obtenus par la moyenne de plus de trois paires différentes d'électrodes, le tableau donne la valeur de l'écart-type. Tous ces résultats sont donnés en pourcentage.

De ces résultats il ressort que les paires d'électrodes éloignées physiquement ne présentent pas un comportement beaucoup plus favorable du point de vue de ces interactions qui sont de toutes façons loin d'être négligeables.

Considérant encore le système comme linéaire on peut le décrire dans son intégralité par :

$$\begin{vmatrix} e_1 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{vmatrix} = X \cdot \begin{vmatrix} q_1 \\ \cdot \\ \cdot \\ \cdot \\ q_n \end{vmatrix}$$

où q1 ... qn sont les quantités d'électricité délivrées par les électrodes, e1 ... en les "excitations" reçues par les groupes de fibres auxquelles elles s'adressent et X une matrice symétrique construite à partir des A et B déterminés par les résultats de l'expérience précédente. La solution idéale de ce problème consiste à envoyer aux électrodes non pas les quantités d'électricité q1 ... qn demandées par les algorithmes de traitement de signal effectuées en amont par le système de prothèse mais :

$q'_1 \ldots q'_n$ tels que :

$$\begin{vmatrix} q'_1 \\ \cdot \\ \cdot \\ \cdot \\ q'_n \end{vmatrix} = X^{-1} \cdot \begin{vmatrix} q_1 \\ \cdot \\ \cdot \\ \cdot \\ q_n \end{vmatrix}$$

La limitation de cette méthode vient de ce que les coefficients non diagonaux de la matrice inverse X^{-1} sont négatifs. Pour compenser les phénomènes d'interactions il faut en effet diminuer les quantités d'électricité émises et ce d'autant plus qu'elles sont déjà plus faibles (accentuer les creux du spectre). Tous les procédés employés pour l'excitation effective des électrodes n'admettent pas que ces quantités d'électricité deviennent négatives, c'est-à dire inhibitrices.

III. SOLUTION PRATIQUE.

Pour mettre en oeuvre ces traitements nous nous trouvons en présence des deux méthodes habituelles de traitement de signal : numérique et analogique.

A ce jour il existe encore peu de systèmes de prothèse cochléaire réalisés de façon totalement numérique. Les moindres algorithmes de filtrage numérique deviennent vite volumineux dans leur réalisation matérielle et gourmands en énergie électrique, caractéristiques difficilement compatibles avec la conception d'un matériel autonome et discret. Pourtant la réalisation numérique câblée ou microprogrammée des traitements évoqués ici ne pose aucun problème théorique.

Nous décrivons ci-dessous une réalisation analogique d'une voie de traitement d'un système de prothèse cochléaire à électrodes multiples procédant d'une analyse spectrale au moyen d'un banc de filtres (figure 2).

Dans ce schéma, les éléments classiques sont le filtre passe-bande, le redressement double alternance, le filtre passe-bas et l'amplificateur logarithmique.

Les fréquences de coupure des filtres passe-bande sont à rechercher dans un bon compromis entre le nombre de voies à prendre en compte et les bandes critiques de l'oreille.

	Nombre d'électrodes intermédiaires										
	0	1	2	3	4	5	6	7	8	9	10
B	42%	35%	35%	33%	36%	17%	14%	24%	22%	18%	19%
Ecart	38%	30%	9%	19%		21%					

Tableau 1. Coefficients de couplage (B) et écart-type (E).

Nous retenons pour la réalisation du filtre passe-bande une structure à 2 cellules SALLEN & KEY passe-bas et passe-haut, non optimale du point de vue des valeurs des composants passifs mais donnant à peu de frais une relation optimale entre fréquence de travail et gain, permettant d'éviter ainsi la préaccentuation de 6 dB/octave habituelle dans tout système de traitement de parole.

Le filtre passe-bas est également de type SALLEN & KEY avec un gain égal à 3 pour compenser l'affaiblissement qui intervient dans l'étage de redressement.

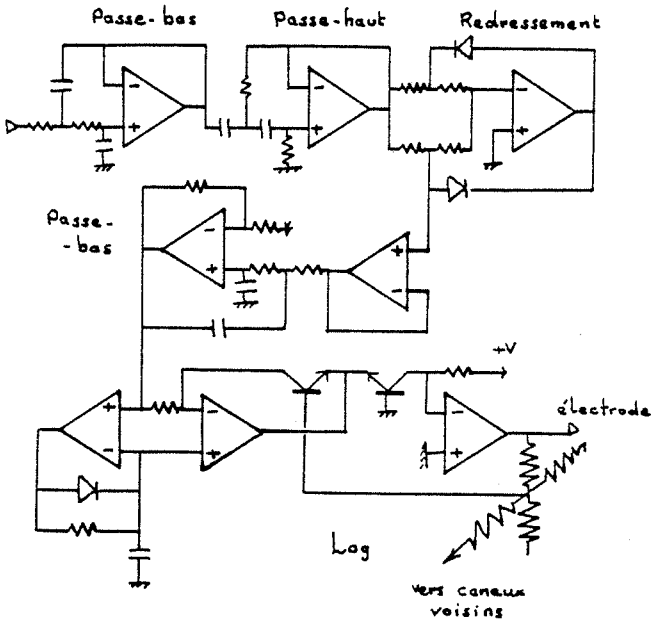


Figure 2. Schéma électrique d'une voie d'analyse.

Nous insérons entre passe-bas et amplificateur logarithmique un étage épieteur permettant de toujours représenter l'énergie du signal par différence par rapport au niveau de bruit supposé correspondre aux niveaux incidents les plus faibles. Cette procédure permet de compenser dans une certaine mesure le peu de dynamique disponible au niveau des électrodes et de ne pas envoyer au patient les niveaux sonores continus qui de toutes façons n'apportent pas d'information utile pour la compréhension de la parole.

Par ailleurs la dynamique du codage logarithmique se voit ainsi augmentée en présence de bruit, réduisant l'effet de masque de celui-ci (renforcement artificiel de l'effet LOMBARD).

L'amplificateur logarithmique est également de structure classique mais sa boucle de réaction terminale contient un réseau de résistances représentant les interactions électriques entre les électrodes, réalisant ainsi automatiquement l'inversion de la matrice X, ... pour autant que les valeurs des résistances puissent être suffisamment représentatives des interactions électriques réelles.

Des études ultérieures devront déterminer si ces composants peuvent être choisis d'une façon moyenne acceptable pour tous les cas qui se présentent ou s'ils doivent être ajustés très précisément pour chaque patient.

REFERENCES.

Y. Cazals & J. F. Rouanet, "Le système d'implant extra-cochléaire à une voie : PRELCO", Innov. et Techn. en Biologie et en Médecine, 1985.

C. H. Chouard, C. Fugain, B. Meyer & H. Lacombe, "Résultats cliniques de l'implant cochléaire à multi-électrodes", Ann. Otolaryngol. 98, pp. 593-612, 1981.

G. M. Clark & Y. C. Tong, "A Multiple-Channel Cochlear Implant. A Summary of Results for Two Patients", Arch. Otolaryngol. 108, pp. 214-217, 1982.

M. Fardeau & P. Orange, "Les implants cochléaires : l'appareillage", Les Cah. d'O.R.L. 14-6, pp. 606-616, 1979.

J. Génin & R. Charachon, "Multi-Channel Cochlear Implants : Psychophysical Experiments", Cochlear Implants in Clinical Use 2, KARGER, pp. 90-97, 1984.

B. C. J. Moore, E. Douek, A. J. Fourcin, S. M. Rosen, J. R. Walliker, D. M. Howard, E. Abberton & S. Frampton, " Extracochlear Electrical Stimulation with Speech Patterns : Experience of the EPI Group (UK)", Cochear Implants in Clinical Use 2, KARGER, pp. 148-162, 1984.

****
* SYNTHESE *
****

SYNTHESE PAR DIPHONES UTILISANT LE CODAGE PREDICTIF MULTIIMPULSIONNEL ET UN VOCODEUR DE PHASE

Michel STELLA et Francis CHARPENTIER

Centre National d'Etudes des Télécommunications
22301 LANNION, FRANCE

RESUME

Le codage prédictif multiimpulsionnel (MPLPC) permet d'obtenir une qualité de parole très naturelle par analyse-synthèse, pour des débits relativement faibles. Cette technique a surtout été étudiée à des fins de transmission ou de stockage de la parole. On montre dans ce papier que l'on peut aussi l'utiliser dans un système de synthèse par diphones. Le point délicat consiste à réaliser une prosodie acceptable pour la parole synthétique. Ce problème est résolu en deux étapes. D'abord, on effectue une simple concaténation des diphones, ce qui fournit un signal synthétique avec un contour mélodique assez plat. On restaure ensuite une prosodie plus appropriée en utilisant une version adéquate du vocodeur de phase.

1. INTRODUCTION

Il est bien connu que la parole codée par prédiction linéaire présente certains défauts comme son timbre métallique et la mauvaise qualité des sons fricatifs voisés. Ces problèmes sont partiellement résolus par l'usage de sources mixtes /2/, mais le timbre s'en trouve généralement dégradé. Malgré cela, les techniques à prédiction linéaire sont largement utilisées en synthèse de parole à partir d'éléments préanalysés (diphones ou demi-syllables), dès lors qu'elles impliquent une décomposition source-filtre bien adaptée à la manipulation des paramètres prosodiques.

A l'heure actuelle, la meilleure solution pour améliorer la qualité de la parole à des débits relativement réduits paraît être la technique MPLPC /1/. Malheureusement, cette technique n'est pas bien adaptée à la synthèse de la parole puisque les paramètres de source (position et amplitude d'un certain nombre d'impulsions) dépendent des coefficients du filtre par l'intermédiaire d'un algorithme d'optimisation. Toute modification de ces paramètres remet en question cette interaction source-filtre et entraîne une dégradation de la qualité.

Cependant, quelques tentatives de modification de la prosodie en utilisant la technique MPLPC ont été décrites dans la littérature /3,4/. Ces expériences portaient seulement sur des modifications limitées. Les techniques utilisées consistaient à modifier directement le schéma d'excitation MPLPC et cela ne paraissait pas toujours conserver le gain de qualité intrinsèque à la méthode .

Dans cet article, nous proposons un algorithme MPLPC pour la synthèse de la parole à partir du texte. Nous avons choisi une approche "a posteriori", c'est-à-dire de restaurer la prosodie désirée sur la sortie du filtre de synthèse. Une telle approche est possible parce que la simple concaténation de diphones codés en MPLPC produit , moyennant certaines précautions au cours de l'analyse, une parole de qualité déjà bonne. La prosodie , un peu artificielle , est ensuite ajustée au moyen d'un vocodeur de phase. On a choisi le vocodeur de phase parce qu'il est capable de réaliser des modifications prosodiques tout en conservant un timbre très naturel /5-7/.

2. LES MODIFICATIONS PROSODIQUES

Afin de restaurer une prosodie naturelle sur un signal de parole possédant une mauvaise prosodie intrinsèque, nous avons repris le système proposé par Seneff /5/, que nous avons adapté de façon à intégrer dans un algorithme unique les modifications de la fréquence fondamentale (F0) et des durées. Le système est dérivé du vocodeur de phase /6/, qui permet une représentation du signal de parole par son spectre à court terme. Un traitement approprié du spectre avant la resynthèse rend possible des modifications indépendantes et par des facteurs relatifs du rythme et du contour mélodique.

Le système est décrit sur la figure 1. Il est fondé sur l'interprétation de la transformée de Fourier à court terme comme un traitement du signal par un banc de filtres /7/. Pour les détails généraux sur le système, le lecteur est renvoyé à /5/. La nouveauté de notre système concerne le troisième étage de l'algorithme, où les modifications de pitch et de durée sont effectuées simultanément. Ces modifications sont fondées sur les remarques suivantes: chaque coefficient DFT est un signal temporel qui peut

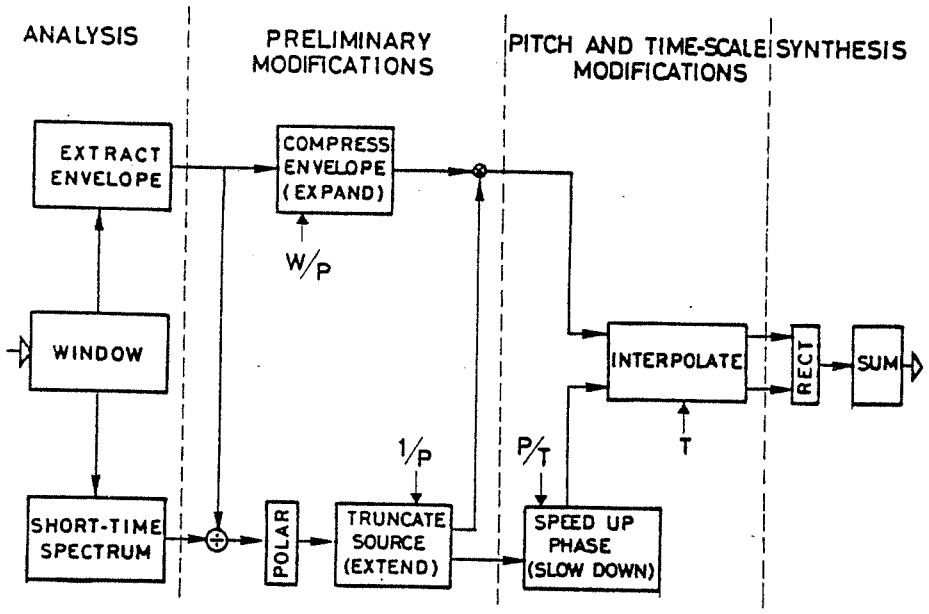


Fig.1 : Implementation de l'algorithme de vocodeur de phase

être interprété comme la sortie d'un filtre passe-bande. L'échelle des fréquences pour chacun de ces filtres DFT peut être redéfinie linéairement par un facteur multiplicatif R. Une telle opération peut être réalisée en effectuant sur chaque signal DFT l'un des deux traitements suivants:
1) une interpolation du signal DFT à une cadence d'échantillonnage différente, R-fois plus rapide; ce traitement équivaut à modifier la vitesse de défilement d'un enregistrement sonore sur bande magnétique. cependant l'échelle des temps est également modifiée par le facteur R;
2) une multiplication par R des variations temporelles de la phase ; ce traitement a l'avantage de conserver l'échelle des temps.

Une modification combinée de l'axe des fréquences par un facteur P et de l'axe des temps par un facteur T est obtenue en utilisant ces deux méthodes en série, la méthode d'interpolation avec un facteur R1=T et la méthode de modification de phase avec un facteur R2=P/T. Grâce à la séparation source-filtre on peut appliquer la modification de l'échelle des fréquences uniquement à la source, et donc modifier F0 par le facteur P. La composante enveloppe peut être modifiée plus avant, et en plus de l'extension nécessaire du facteur 1/P pour compenser la modification de l'échelle des fréquences, une compression-extension linéaire d'un facteur W peut être appliquée. Cette technique est utile pour des modifications de la qualité de la voix, dans des expériences de conversion femmes-hommes, par exemple /5/.

Le traitement décrit dans la Figure 1 est effectué à la fréquence d'échantillonnage de l'entrée. Le suréchantillonnage élevé autorise des techniques d'interpolation simples, comme l'interpolation linéaire de la phase et de l'amplitude. Les modifications prosodiques sont évolutives. Cependant, les modifications spectrales (F0 ou enveloppe) doivent varier relativement lentement de manière à respecter la relation d'incertitude temps-fréquence.

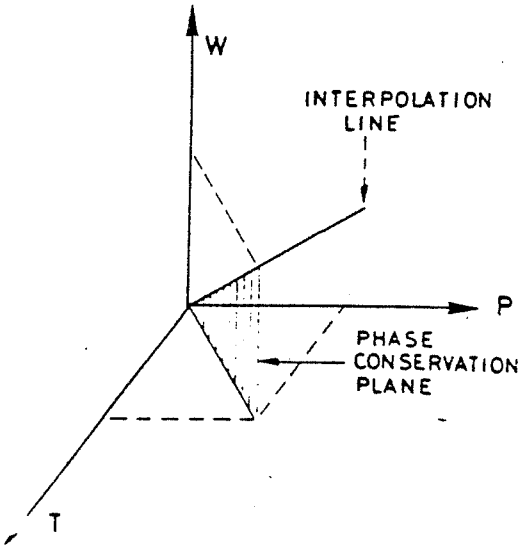


Fig.2 : Représentation des modifications prosodiques dans une espace tridimensionnel

Il est commode de représenter ces modifications dans un espace tridimensionnel (Fig. 2). Les dimensions correspondent aux facteurs de modification de l'échelle des temps (T), du pitch (P) et de l'enveloppe (W). Les modifications d'un facteur constant peuvent être représentées par des points isolés dans cet espace. Les modifications évolutives accomplies par notre système peuvent être décrites comme la trajectoire d'un point mobile. Une simple interpolation du signal correspond à un point sur la ligne médiane, d'équation P=T=W. Le domaine des modifications pouvant être obtenues avec la version originelle du vocodeur de phase /6/ est limité au plan engendré par cette ligne d'interpolation et l'axe des modifications de l'échelle des temps. En utilisant la séparation source-enveloppe introduite par Seneff /5/, notre système est capable d'atteindre n'importe quel point de cet espace . L'axe de modification du pitch est obtenu en inhibant l'opération d'interpolation. Le plan défini par l'équation P=T est obtenu par conservation de la phase . Enfin, les modifications nécessaires au système de synthèse décrit dans le paragraphe suivant peuvent être interprétées comme des trajectoires dans le plan (P,T).

3. LA SYNTHESE PAR DIPHONES

Dans notre système conventionnel de synthèse par diphones LPC du français /9/, l'information sur les diphones est stockée dans un dictionnaire qui spécifie chaque diphone comme une suite de trames LPC. Toute l'information prosodique a été enlevée de ces trames. Cette information (F0 et durée) est fournie par un module prosodique au moment de la synthèse.

Dans notre nouveau système, nous gardons le module prosodique pour spécifier la prosodie désirée. Mais, à cause du codage MPLPC les diphones contiennent une information plus complexe. Les trames MPLPC comprennent les coefficients LPC et le schéma d'excitation. De cette façon, ils contiennent déjà de l'information intrinsèque sur F0. Cependant, pour restaurer la prosodie désirée, nous avons besoin de connaître explicitement F0, pour déterminer les modifications convenables des paramètres de commande du vocodeur de phase.

Pour y parvenir, nous avons construit le nouveau dictionnaire de la façon suivante. Les formes d'onde temporelles des diphones du français, précédemment segmentés pour le dictionnaire de diphones LPC conventionnel, (environ 1200 diphones), ont été échantillonnées à 16 Khz, filtrées passe-bas, et codées en utilisant un convertisseur analogique-numérique à 16 bits.

Ces formes d'onde ont ensuite été visualisées et leurs portions voisées marquées à la main, période par période. Les portions non voisées ont été segmentées en un certain nombre de fenêtres de longueur nominale. Ces formes d'onde ont été codées en MPLPC en utilisant une fenêtre d'analyse synchrone du pitch. Cette technique est efficace pour produire des trains d'impulsions périodiques sur les parties voisées. Cette périodicité du schéma d'excitation MPLPC assure une bonne qualité de synthèse par simple concaténation des trames des diphones. Nous avons utilisé un nombre fixe d'impulsions par fenêtre. Le dictionnaire de diphones MPLPC contient environ 27000 trames.

La procédure de synthèse est décrite dans la figure 3. Les diphones requis sont extraits du dictionnaire MPLPC, et les trames correspondantes sont concaténées (environ 20 trames par diphone). Ensuite, nous synthétisons cette suite de trames, en utilisant la mémoire du filtre LPC aux frontières de diphones. La parole "concaténée" résultante ne contient pas de prosodie "voulue", mais seulement la prosodie intrinsèque des diphones. Cependant, il se trouve que la voix de la locutrice utilisée pour le dictionnaire de diphones est extrêmement régulière, et que les discontinuités de F0 aux frontières des diphones sont raisonnablement petites, si bien que ces phrases obtenues par simple concaténation de formes d'onde sont très intelligibles, et possèdent une très bonne qualité segmentale. On peut même y entendre ue certaine prosodie résiduelle, particulièrement sur les voyelles suivies par des silences, puisque ces voyelles ont été extraites de la dernière syllabe de mots isolés. Le suivi de F0 montre aussi que la microprosodie des consonnes est préservée dans son ensemble, sans doute à cause de la régularité de notre locutrice.

En utilisant les valeurs de F0 et de durée pour la phrase fournies par le module prosodique, et l'information prosodique contenue dans les trames MPLPC nous calculons les paramètres de commande du vocodeur de phase. Les formes d'onde "concaténées" sont ensuite traitées par le vocodeur.

4. RESULTATS ET DISCUSSION

La parole synthétique obtenue par cette technique est très naturelle, et ne souffre plus des défauts habituels de la prédiction linéaire, surtout sur les fricatives voisées. Cependant, la plupart des auditeurs lui trouvent une qualité "rauque", qui n'apparaissait pas dans la parole LPC synthétique habituelle. Cette raucité semble provenir de la difficulté de contrôler l'énergie à la sortie du vocodeur de phase. Ce phénomène apparaît parce que la modification de phase détruit la cohérence de phase des harmoniques voisins. Cette dégradation du signal de parole est à peine perceptible sur la parole naturelle, mais elle devient importante dans le cas de signal concaténé qui contient déjà des imperfections à la frontière des diphones.

Une autre cause possible du problème de raucité se trouve dans la variation des paramètres de modification le long de l'onde de parole. Le marquage à la main des périodes dans le dictionnaire MPLPC peut introduire une certaine "gigue" dans les valeurs de F0 assignées aux diphones. Or, des petites variations de F0 sur l'onde de

sortie peuvent être perçues comme rauques.

En effet, nous avons vérifié qu'en lissant les variations des paramètres de modification, on peut réduire quelque peu la raucité.

A cause du vocodeur de phase, notre système est coûteux en temps de calcul. Bien que les algorithmes aient été implantés sur un processeur vectoriel, ils sont très loin du temps réel. L'algorithme du vocodeur de phase est exécuté à la fréquence d'échantillonnage de l'entrée, et on pourrait donc gagner beaucoup de temps de calcul en utilisant une interpolation de la STFT. De telles améliorations pourraient être implantées si la technique recouvrement-addition de synthèse de STFT /8,10/ peut s'appliquer aux modifications de F0; la difficulté est d'utiliser une technique FFT pour synthétiser la STFT, après une modification arbitraire de l'axe des fréquences.

La taille du stockage nécessaire pour le dictionnaire de diphones est un facteur important, bien que nous n'ayons essayé aucune sorte d'optimisation dans ce domaine. Cependant, puisque ce système de synthèse nécessiterait une machine extrêmement puissante, pour pouvoir tourner en temps réel, cette contrainte de taille de mémoire disparaîtrait d'elle-même.

5. REMARQUES FINALES

On a montré que ,pour des applications de synthèse de parole, une version appropriée de la technique multiimpulsionnelle est l'analyse MPLPC synchrone du pitch. Ceci implique une étape assez lourde de marquage à la main de périodes fondamentales, mais un tel investissement se justifie si l'on veut obtenir de la parole de synthèse de bonne qualité.

La concaténation des diphones après une telle analyse donne une qualité de parole assez naturelle, en ce qui concerne le timbre. Une prosodie plus naturelle peut être restaurée, a posteriori, en utilisant une version du vocodeur de phase capable de traiter des modifications du pitch et de la durée, variables dans le temps et indépendantes.

6. REFERENCES

/1/ D.S. Atal, J. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates", Int. Conf. ASSP, pp.611-614, Paris, 1982

/2/ J. Makhoul, "A mixed source model for speech compression and synthesis", J. Acoust. Soc. Amer., Vol.64, No.6, pp.1577-1581, 1978

/3/ B.E Caspers, B.S. Atal, "Changing pitch and duration in LPC synthesized speech using multipulse excitation", J. Acoust. Soc. Am., Vol.73, S5, Spring 1983

/4/ M.G. Stella, "Modifications des paramètres prosodiques en analyse-synthèse multi-impulsionnelle", 13-emes JEP, Bruxelles, 1984

/5/ S.S. Seneff, " System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction ", IEEE-ASSP, Vol.30, No.4, pp.566-578, 1982

/6/ J.L. Flanagan, R. Golden, " Phase Vocoder", Bell Syst. Tech. J., Vol.45, 1966, pp.1494-1509

/7/ M.R. Portnoff, " Implementation of the digital phase vocoder using the fast Fourier transform ", IEEE Trans. ASSP, Vol.24, No.3, pp.243-248, 1976

/8/ J.B. Allen, R.R. Rabiner, " A unified approach to short-time Fourier analysis and synthesis ", Proc. IEEE, Vol.65, No.11, pp.1558-1564, 1977

/9/ J.L. Courbon, F. Emerard, " SPARTE: a text-to-speech machine using synthesis by diphones ", IEEE Int. Conf. ASSP, Paris, pp.1597-1600, 1982

/10/ R.E. Crochiere, " A weighted overlap-add method of short-time Fourier analysis/synthesis ", IEEE Trans. ASSP, Vol.28, No.1, pp.99-102, 1980

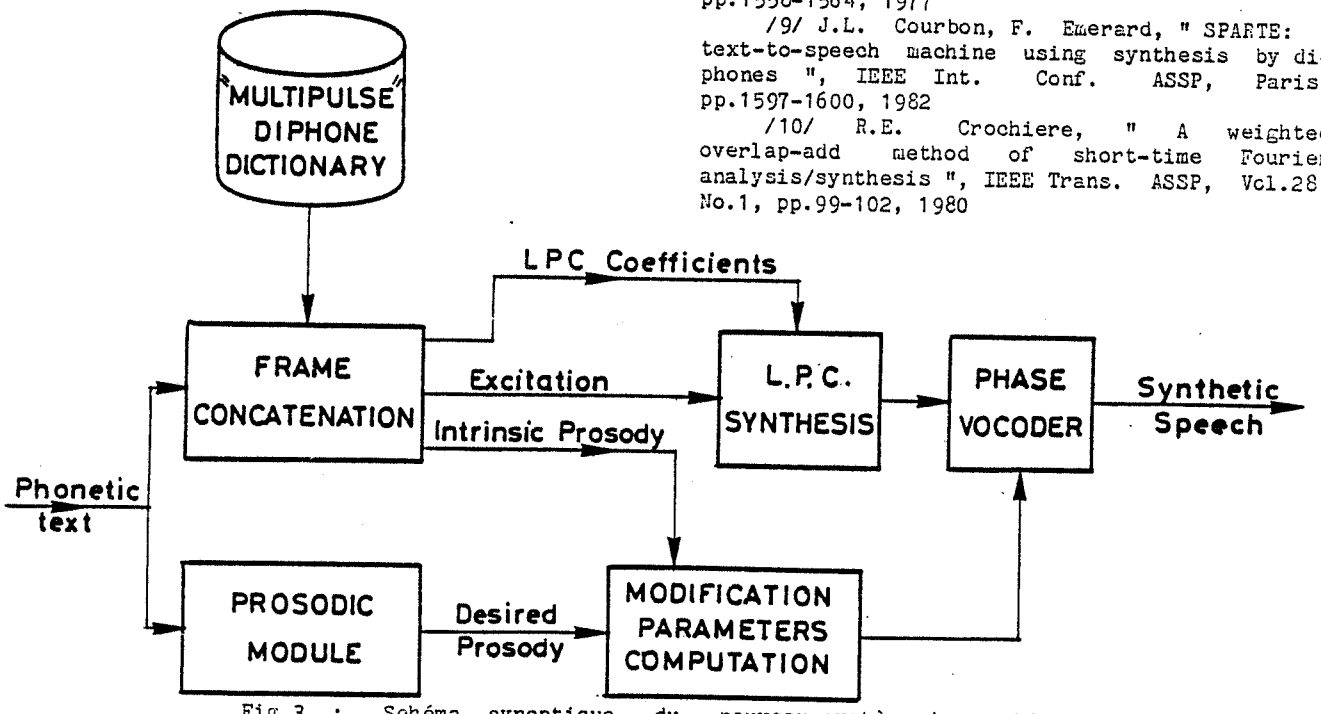


Fig.3 : Schéma synoptique du nouveau système de synthèse

SYNTHESE PAR DIPHONES MULTILANGUE

M. Contini[1], H. Zinglé[2]
V. Aubergé[1], L.J. Boë[1], G. Murillo[1]

[1]. Institut de la Communication Parlée de Grenoble
[2]. Université de Chambéry

Abstract

This paper describes the realization of a multilanguage speech synthesis system using a modular language-independant software.

The achievement of unlimited speech synthesis from text requires the integration of phonetic and linguistic knowledge at all levels :
- corpus elaboration (diphones),
- dictionary segmentation,
- orthophonetic transcription rules,
- intonation rules,
- speech intelligibility and quality evaluation.

This has been done adopting and adapting CNET Lannion's speech department experience and know-how. At each stage procedures have been conceived user-oriented.
The dictionary elaboration and synthesis realization software have been developed on Italian and German.

L'Institut de Phonétique de Grenoble est engagé dans un projet de synthèse multilangues. Les travaux ont débuté par la synthèse de l'allemand et de l'italien (projet SYNTALIT) : ils ont bénéficié du savoir-faire et de l'expérience du CNET pour la synthèse du français au cours de ces dix dernières années (voir notamment [2] et [8]) et de tout l'acquis de l'Institut de la Communication Parlée (voir par exemple LETY [4] et MURILLO [6]). L'élaboration d'une méthodologie commune permettra, à l'avenir, d'entreprendre la synthèse d'autres langues (espagnol et anglais en particulier). L'intérêt d'une telle démarche est de permettre la mise au point d'une carte de synthèse multilangues, les particularités spécifiques de chaque langue étant traitées par logiciel.

Nous exposerons ici trois aspects essentiels de cette démarche, à savoir :
- constitution de la bibliothèque de diphones (1925 pour l'allemand et 1123 pour l'italien),
- conception de la transcription orthographe-phonétique,
- choix d'une stratégie pour le traitement de la prosodie,
- évaluation de la qualité de la parole codée.

1. BIBLIOTHEQUE DE DIPHONES

Les diphones ont été extraits de corpus adaptés à la structure phonique de chaque langue et tenant compte :
- de l'inventaire des unités phonétiques (y compris celles qui apparaissent dans les mots d'origine étrangère),
- de leur distribution,
- des contraintes accentuelles.

Les corpus ont été enregistrés au CNET Lannion selon la procédure mise au point pour la constitution de la base de données des sons du français dans le cadre du GRECO Parole.

A partir de ces enregistrements (fréquence d'échantillonnage de 16 Khz sur 16 bits) ont été réalisés :
- une analyse LPC à 18 coefficients,
- des tracés de type sonagraphique ([5]) ("monnéogrammes") comportant un repère pour la segmentation avant le début de chaque mot.

Pour la création du dictionnaire ont été respectées les normes du logiciel DICOPHONE développé par le CNET [8]. Toutefois la segmentation des diphones a été réalisée non pas à partir de la visualisation sur écran des paramètres LPC mais directement à partir de "monnéogrammes" à l'aide d'une tablette d'acquisition graphique.

Cette procédure a permis d'utiliser au mieux les compétences des phonéticiens spécialistes des langues traitées, la segmentation étant indépendante de la création proprement dite du dictionnaire.

Le logiciel de synthèse commun s'inspire du logiciel MONIT du CNET : seules les spécifications prosodiques et la bibliothèque de diphones sont différentes d'une langue à l'autre.

2. LA TRANSCRIPTION ORTHOGRAPHE-PHONETIQUE

Elle a été conçue dans une perspective identique : pouvoir disposer d'un logiciel commun et ne traiter les particularités de chaque langue que par le seul biais des règles de transcription. Pour l'écriture des règles elles-mêmes a été adoptée une même formalisation. Un ensemble de règles contex-

tuelles permet de définir la valeur phonétique exacte de chaque graphème ou de chaque groupe de graphèmes [1]. Le nombre de règles varie d'une langue à l'autre : 478 ont été retenues pour l'allemand alors que 108 suffisent pour l'italien. Dans les deux cas elles tiennent compte des contraintes accentuelles. En allemand le problème des durées vocalique et consonantique, lié à la place de l'accent, a été traité à part dans un module ad hoc. En italien on doit tenir compte des possibilités d'opposition e/ɛ, o/ɔ en syllabe accentuée et de la neutralisation à l'avantage de e,o en syllabe inaccentuée.

3. PROSODIE

Pour chaque langue les caractéristiques prosodiques ont été étudiées séparément. Des corpus ont été élaborés pour rendre compte des structures syntaxiques les plus courantes dans des phrases de complexité variable.

Leur enregistrement, réalisé à l'Institut de Phonétique de Grenoble a permis l'obtention de tracés mingographiques (signal, intensité et Fo). Durée et Fo ont été retenus comme paramètres prioritaires. La segmentation des énoncés a été faite sur les tracés. Ont été mesurées :

- les durées de tous les segments vocaliques et consonantiques en tenant compte de leur place par rapport à l'accent,
- les variations de Fo pour chaque voyelle et pour chaque consonne voisée.

Une procédure d'acquisition à partir de la tablette d'entrée graphique a été également mise au point pour faciliter l'acquisition des paramètres étudiés.

Nous avons retenu les procédés développés par F. EMERARD pour la prosodie.

4. EVALUATION

La dernière étape du projet SYNTALIT est l'évaluation de la qualité de la parole codée. Elle est réalisée par :

- des tests d'intelligibilité
  Sont utilisés notamment les tests DRT (Diagnostic Rhyme Test), DAT (Diagnostic Alliteration Test) et MRT (Modified Rhyme Test) développés par VOIERS aux Etats-Unis (1967), adaptés pour le français par PECKELS et ROSSI [7].
- des tests d'agrément
  La méthode employée, déjà appliquée aux Etats-Unis et en France, comprend le jugement par catégories, le DAM (Diagnostic Acceptability Measure) et l'essai d'isopréférence [9].

Ces derniers tests impliquent les faits segmentaux et suprasegmentaux et font appel à la notion de qualite.

CONCLUSION

Ce projet concrétise le franchissement d'une étape importante dans les recherches sur la parole : associations étroites de chercheurs de formations différentes, collaborations entre Centres de Recherches et développement en vue de transferts vers l'industrie.

Plus qu'une étude orientée vers les applications, il s'agit en fait d'un outil de recherche qui par sa conception modulaire doit permettre l'intégration de nouvelles langues, l'amélioration des règles et l'interchangeabilité des modules.

Remerciements

Nous tenons à remercier l'équipe de départ RCP et tout particulièrement F. EMERARD pour sa collaboration constante au cours de ce travail.

BIBLIOGRAPHIE

[1] V. Aubergé, "Passage automatique du texte orthographique vers le texte phonétique", 14èmes JEP GCP-GALF, 1985.

[2] F. Emerard, "Synthèse par diphones et traitement de la prosodie", Thèse de 3ème Cycle de Phonétique, Université de Grenoble III, 1977.

[3] F. Ferrero (Ed.), "Atti delle tavola rotonda sulla sintesi della parola", Padova, Centro di Studi per le Ricerche di Fonetica del CNR, 1975.

[4] M. Lety, "Transcription orthographique-phonétique : un système interpréteur", Thèse Doct.-Ing. Informatique, Université Grenoble I, 1980.

[5] J. Monne, "Programme de calcul de spectrogramme numérique", Séminaire "Traitement du signal", GCP-GALF, 1983.

[6] G. Murillo, "Application de l'analyse par la synthèse à la génération des consonnes occlusives voisées du français. Expériences de perception", Thèse Doct.-Ing., I.N.P. Grenoble, 1984.

[7] J.P. Peckels & M. Rossi, "Le test de diagnostic par paires minimales" 2èmes JEP, GCP-GALF Bd Be Bf Bg, 1971.

[8] M. Stella, "Fabrication semi-automatique de dictionnaires de diphones", Recherches/Acoustique CNET VIII, 1983.

[9] W.D. Voiers, "Performance evaluation of speech processing devices. III : Diagnostic evaluation of speech intelligibility", AFCRL Final Report, Contract AF 19 (628)-4987, AFCRL-67-0101. AD 650 158, 1967.

[10] H. Zinglé, "Traitement de la prosodie allemande dans un système de synthèse de la parole", Thèse d'Etat, Strasbourg II, 1982.

SEGMENTATION D'UNE BASE DE DONNEES DE "POLYSONS", APPLICATION A LA SYNTHESE DE LA PAROLE

F. Laferrière(**), G. Chollet(*), L. Miclet(*), J.P. Tubach(*)

(*) ENST, Département SYC (CNRS, UA 820), 46 Rue Barrault, 75013 Paris.
(**) INRS-Télécommunications, Ile des Soeurs, Verdun, Québec, Canada.

RESUME

Cet article présente une méthode de synthèse de parole, utilisant un dictionnaire constitué de segments de parole plus longs que des diphones, de type demi-syllabique; ces segments ont été baptisés "polysons".
La première étape est la création de ce dictionnaire, à partir d'une base de données vocale constituée de l'ensemble des polysons: il faut les extraire du contexte phonologiquement neutre dans lequel ils ont été prononcés.
La seconde étape est le calcul sous forme appropriée à la synthèse LPC des paramètres représentatifs.
Enfin, il faut concaténer ces polysons pour synthétiser des phrases.

LA BASE DE DONNEES "POLYSONS"

L'ensemble des polysons provisoirement jugés utiles pour la synthèse (et éventuellement pour des expériences de reconnaissance) a été enregistré en Juillet 1984, au CNET Lannion par L. MICLET [1]. Cet ensemble a été constitué afin d'améliorer la synthèse segmentale. En effet, les segments consitués d'une consonne vocalique (/w/, /j/, /ɥ/, /l/, /R/) et d'une autre consonne posent des problèmes de segmentation et de concaténation. Ces consonnes vocaliques sont donc stockées avec leur contexte sous l'appellation de "polysons".

L'enregistrement a été effectué selon les normes de la base de données des sons du Français du GRECO Communication parlée.

Chaque polyson est enregistré dans un contexte phonologique neutre, le rendant plausible dans un contexte général ([2], [3]), et prévu pour ne pas provoquer trop de difficultés de segmentation. Plus précisément, un polyson est nécessairement précédé d'un des éléments de la liste suivante:

silence : si le polyson est du type silence - phonème.
/a/ : si un polyson long commence par une consonne qui n'est pas une plosive sourde
/a/+silence : si un polyson long commence par par une plosive sourde
/a t/ : si un polyson long commence par une voyelle
/a t a/ : si un polyson court commence par une consonne qui n'est pas une plosive sourde
/a t a t/ : si un polyson court commence par une voyelle

et nécessairement suivi d'un élément de la liste suivante:

silence : si un polyson est du type phonème - silence
/t a/ : si un polyson long finit par une voyelle
/a t a/ : si un polyson long finit par une consonne
/t a t a/ : si un polyson court finit par une voyelle

Exemples de polysons et de leur contexte: /plo/ dans /aplota/; /dli/ dans /adlita/; /elv/ dans /atelvata/.
Il y a 2630 polysons différents dans la base de données.

Les fichiers originaux, digitalisés à 16000Hz, ont été filtrés passe bas et sous échantillonnés à 8000 Hz. On a également pratiqué un filtrage passe-haut à 100Hz, pour éliminer des ondes de basse fréquence présentes dans le signal original et vraisemblablement dues à une onde de compression atteignant un microphone nu.

SEGMENTATION

Comme on connait exactement le contexte de chaque polyson, la méthode adoptée consiste à reconnaitre ce contexte avec précision, pour mieux le faire disparaitre.

Pour effectuer l'analyse nécessaire à la segmentation, chaque phrase a été découpée en fenêtres (rectangulaires) de 128 points, ne se recouvrant pas. Sur chacune de ces fenêtres, on a calculé:
- l'énergie, en dB.
- le nombre de passages par zéro de la dérivée du signal

- un ensemble de 10 coefficients de prédiction linéaire, a_i
- la distance d'Itakura séparant la trame considérée d'un /a/ typique du locuteur considéré. Cette distance est donnée par:

$$D(f1,f2) = \frac{A_2^T \, Rp(f1) \, A_2}{A_1^T \, Rp(f1) \, A_1} - 1$$

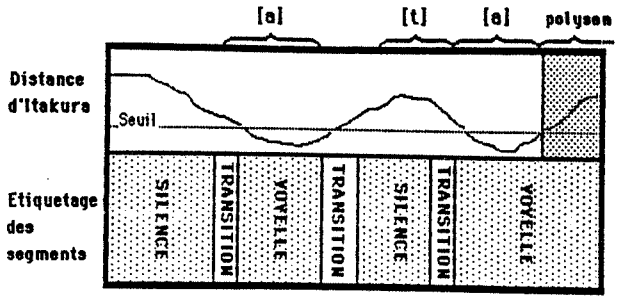
où f1 et f2 représentent les deux signaux à comparer, $A_i = (1, a_1, a_2, \ldots a_n)$ $(i = 1,2)$ le vecteur contenant les n prédicteurs et 1 comme première entrée, et Rp(f1) la matrice d'autocorrélation du signal f1 (référence [4]).

On fournit l'énergie et le nombre de passages par zéro à une procédure de segmentation et d'étiquetage, qui détermine les frontières et classe les segments en quatre types: Silence, Vocalique, Fricatif, et Indéterminé. (on utilise une version simplifiée d'un algorithme proposé dans [5]). C'est surtout la détermination de la position des silences qui s'est avérée utile, mais les autres étiquettes sont servent à l'occasion, et surtout sont précieuse pour vérifier et reprendre à la main les segmentations manquées ou douteuses.

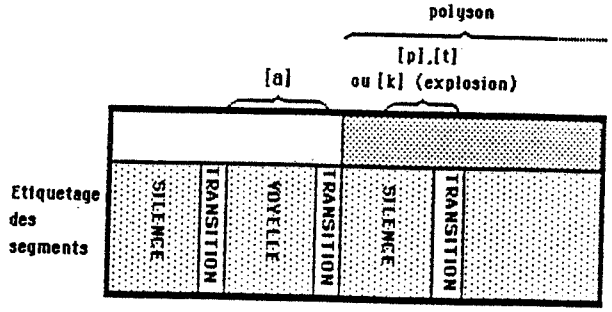
C'est à partir de l'étiquetage et de la distance d'Itakura que le choix des frontières est effectué. En effet, si on examine la liste des contextes possibles, on se rend compte que la segmentation se fait soit à la fin ou au début d'un silence, soit à la fin ou au début d'un /a/.

La distance d'Itakura s'est révélée être d'une remarquable fiabilité pour la détection de /a/, une fois bien sûr que l'on a choisi judicieusement une trame bien caractéristique d'un /a/ typique du locuteur, et une valeur convenable pour le seuil.

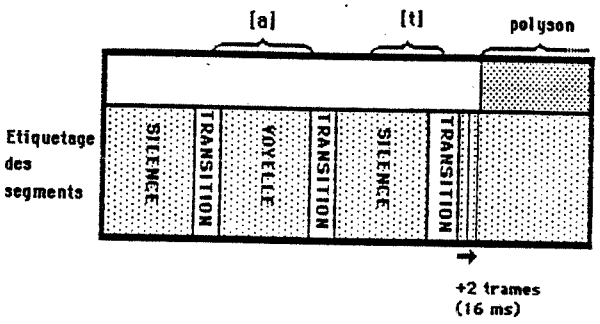
Les différents cas sont les suivants:
- si on segmente sur une frontière du type /a/ - polyson, on utilise la distance d'Itakura pour déterminer avec précision la fin du /a/ après avoir déterminé approximativement son début en comptant les silences.



- si on segmente sur une frontière silence - polyson, on n'a qu'à compter les silences (on récupèrera le silence précédant l'explosion de la plosive comme faisant partie du polyson).



- si on segmente sur une frontière /a t/ - polyson, on cherche la fin de la transition qui suit l'explosion,et on laisse tomber encore deux trames (32ms) pour s'assurer de perdre toute influence du /t/.



Pour la recherche des frontières de fin de polyson,on procède sensiblement de même de droite à gauche.

Cette procédure a été menée de façon automatique. A l'usage, nous avons ensuite été amenés à modifier manuellement un petit nombre de décisions de segmentation, après examen à la console graphique et à l'écoute.

CREATION ET STOCKAGE DE L'INFORMATION LPC

Pour faire de la synthèse LPC classique ([6]), ce à quoi nous nous sommes limités pour l'instant, il faut conserver, pour chaque trame que l'on voudra reconstruire:
- les coefficients LPC (pour nous, 14 coefficients de réflexion Ki)
- l'énergie résiduelle (permettant de calculer l'amplitude d'entrée du filtre de synthèse)
- un sémaphore de voisement permettant de savoir si on doit injecter du bruit ou un train d'impulsions à l'entrée du filtre de synthèse.

Nous avons de plus retenu la valeur du fondamental dans les parties voisées, de façon à reproduire à la synthèse la structure micro - prosodique (variations sur une courte échelle de temps autour du fondamental moyen).

L'analyse est faite sur des fenêtres de Hamming de 128 points, séparées entre elles de 64 points (soit 8ms). Le signal est préaccentué (facteur 0.90). Le calcul des Ki est fait, par la méthode de Levinson, et la détection de pitch par

une méthode cepstrale

LOGICIEL DE SYNTHESE

On part d'une séquence de polysons représentant la phrase à synthétiser. (il est clair que cette séquence devrait et pourrait être obtenue à partir du texte, ([7], [8]), mais nous n'avons pas encore développé cette partie).

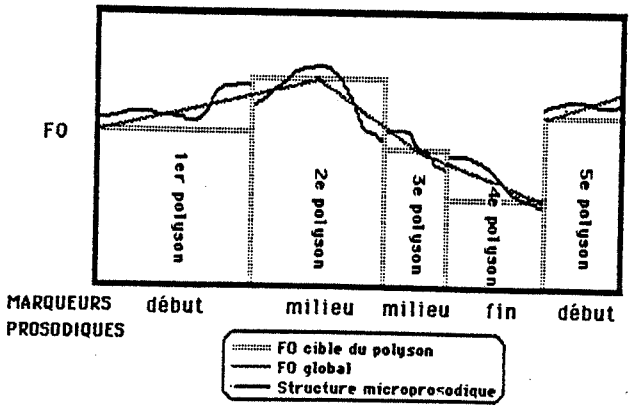
Le programme ASSPOLY fait l'assemblage des polysons de cette liste, et génère pour chaque trame l'information utile pour la synthèse, à savoir: le fondamental (s'il y a lieu), l'amplitude résiduelle, le nombre de points de la trame (pris pour l'instant égal à 64 points, c'est à dire égal à la distance entre les trames d'analyse qui ont permis de générer les Ki), et enfin les 14 Ki.

Ces informations sont passées au programme SYNTLPC, qui fait une synthèse LPC classique, à l'aide d'un filtre en treillis alimenté par une source alternée voisée / bruitée.

Plus précisément, les informations exploitées par ASSPOLY sont les suivantes:
- numéro du polyson (permettant de retrouver ses paramètres)
- valeur du fondamental moyen
- position (début, milieu ou fin) dans le segment prosodique.

Raccords prosodiques

Si on utilisait brutalement le f0 associé à chaque polyson, on obtiendrait une mélodie en créneaux du plus mauvais effet. Pour obtenir une mélodie plus naturelle, on lisse linéairement le fondamental entre les polysons se trouvant dans le même segment prosodique. La technique est illustrée par la figure ci-dessous.



On tend les segments de ce lissage entre des points d'ancrage se trouvant au milieu du polyson si il est marqué "milieu", à sa fin si il est marqué "fin, à son début s'il est marqué "début".

Pour ajouter du naturel à cette mélodie linéaire, on lui surimpose la structure micro-mélodique extraite de l'original. Cette microprosodie est récupérée en prenant, pour chaque trame de l'original, l'écart de f0 à la moyenne pour toutes les trames voisées du polyson.

De plus on ajoute du silence entre les segments prosodiques, et enfin, pour éviter une variation brusque et désagréable de l'énergie au voisinage d'un silence, on multiplie les entrées (les amplitudes résiduelles) des premières trames suivant un silence par une rampe montante (respectivement descendante, en fin de segment).

Raccord acoustique

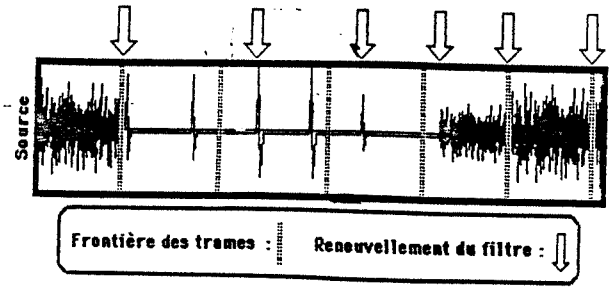
Un des principaux problèmes associés à la concaténation de segments de parole codés en LPC provient des variations brusques des coefficients de réflexion Ki, et en conséquence des caractéristiques acoustiques, autour d'une frontière. Pour y remédier, il faut naturellement procéder à un lissage. L'expérience a montré que l'effectuer directement sur les Ki n'est pas satisfaisant, en raison de leur manque de signification physique directe. Par contre, la théorie de la prédiction linéaire établit un lien entre les Ki et les rapports d'aires des sections du tube du modèle acoustique équivalent au filtre LPC ([6]):

$$\frac{Ri}{Ri+1} = \frac{1 - Ki}{1 + Ki}$$

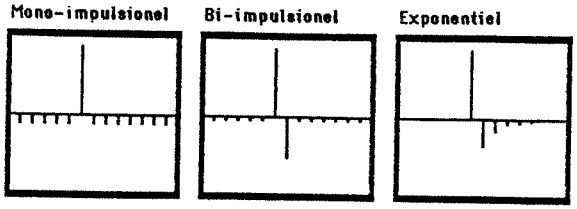
On pratiquera le lissage sur les logarithmes de ces rapports d'aires. On le fait sur quatre trames, de part et d'autre des frontières entre segments voisés

Synthèse

Le signal est généré par un filtre en treillis alimenté par une source alternative: bruit blanc, ou train d'impulsions, selon que le son est respectivement non voisé ou voisé. A la sortie de ce treillis le signal est désaccuentué d'un facteur ajustable. Le renouvellement des coefficients du filtre ne se fait pas nécéssairement au début de chaque trame: en effet, si le son est voisé, on attend le moment où l'on doit générer la prochaine impulsion pour renouveler le filtre. Par contre si la source est bruitée, on le renouvelle au début de chaque trame. Cette règle est illustrée ci dessous:



Le programme SYNTLPC propose un certain nombre de formes possibles pour les impulsions: mono-impulsionnel, bi-impulsionnel, et bi-impulsionnel avec retour exponentiel à zéro.



CONCLUSION

Les résultats obtenus sont encourageants, et mettent bien en évidence les avantages des polysons dans les cas difficiles en synthèse par diphones (exemple: "les flancs blancs...."). Dans un nombre considérable de cas, les raccords difficiles sont évités grâce aux polysons.

Par contre l'actuel dictionnaire de ces éléments comporte quelques lacunes, par exemple du type voyelle - liquide - voyelle, qui éviteraient certains raccords problématiques, surtout pour la liquide /l/. De même des segments du type plosive - plosive - voyelle seraient les bienvenus. Ceci revient à élargir l'ensemble des ploysons, jusqu'ici limité à une partie des demi-syllabes, à quelques syllabes complètes particulièrement difficiles à reconstituer.

D'autre part, certaines suites bien improbables ont peu de chances d'être utilisées, et on pourrait tenir compte pour l'établissement d'un nouveau dictionnaire de polysons, des statistiques disponibles sur les fréquences des suites de phonèmes ([9]).

BIBLIOGRAPHIE

[1] L. MICLET: Enregistrement d'une base de données vocales. Rapport de mission au CNET LLL/TSS/RCP, Juillet 1984.

[2] M. STELLA: Fabrication semi-automatique d'un dictionnaire de diphones. Recherches acoustiques, CNET, vol VII

[3] M. STELLA: La synthèse de la parole. L'Echo des Recherches, mars 1984

[4] Y. GRENIER: Modélisation et reconnaissance de la parole. in "Outils et modèles mathématiques pour l'automatique, l'analyse des systèmes et le traitement du signal", tome 2, I.AD. LANDAU éditeur, Editions su CNRS, Paris, 1982.

[5] M.D. DI BENEDETTO, J.P. TUBACH: Two cooperative methods for the segmentation of running speech. Fortschritte der Akustik, FASE/DAGA'82, Goettingen, Septembre 1982.

[6] J.D. MARKEL, A.H. GRAY: Linear prediction of speech. Springer Verlag, Berlin, Heidelberg, 1976

[7] J. LEROUX, L. MICLET: Transcription orthographique - phonétique et synthèse en temps réel de la parole par prédiction linéaire. 10èmes Journées d'études sur la parole, Grenoble, 1979

[8] B. PROUTS: Contribution à la synthèse de la parole à partir du texte; transcription graphème phonème en temps réel sur microprocesseur. Thèse de Docteur Ingénieur, Université Paris Sud (Orsay) Novembre 1980.

[9] J.P. TUBACH, L.J. BOE: Un corpus de transcriptions de 300000 phones: constitution et exploitation statistique. Document ENST, 85 D001, Paris, Avril 1985.

SYNTHESE DE PAROLE
A L'AIDE DES FONCTIONS D'AIRE LOGARITHMIQUES EVOLUTIVES.

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

ENST, Dept. SYC, CNRS UA-820
46 rue Barrault, 75634 PARIS CEDEX 13, FRANCE.

RESUME.

Les modèles évolutifs permettent de prendre en compte le caractère non-stationnaire du signal de parole: les coefficients sont décomposés linéairement sur une base de fonctions connues à priori. Si de tels modèles s'avèrent utiles en pratique, ils présentent toutefois un manque de stabilité qui est gênant pour cette application.

Dans ce papier, nous décrivons une classe de modèles évolutifs autorégressifs qui, en dépit de leur non-stationnarité, restent stables à chaque instant: ceci est réalisé par l'intermédiaire des fonctions d'aire logarithmiques. Les résultats issus d'expériences de restitution montrent l'intérêt de la technique évolutive en analyse de la parole: l'estimation d'un modèle s'effectue sur des entités de type syllabique, la restitution étant ensuite réalisée en excitant le modèle ainsi déterminé au moyen du résidu préalablement codé.

1. INTRODUCTION.

La qualité de la synthèse de parole par les méthodes paramétriques de traitement du signal dépend du modèle estimé. Or, si le signal de parole est caractérisé par sa non-stationnarité, seuls quelques modèles permettent de le représenter de manière correcte. Un de ces modèles est le modèle ARMA à coefficients dépendant du temps, modèle pour lequel chacun des coefficients est supposé approché par une combinaison linéaire sur une base de fonctions connues, mais dont les pondérations, invariantes, sont à déterminer. Un tel modèle est souvent appelé modèle 'évolutif' pour le distinguer d'autres modèles eux-aussi à coefficients dépendant du temps (modèles adaptatifs, modèles à coefficients aléatoires, ...).

Les modèles évolutifs se sont révélés utiles en reconnaissance et en synthèse de parole mais leur emploi est limité par certaines difficultés, la principale étant que la technique en question n'assure pas la stabilité du modèle estimé.

Le propos de ce papier est de décrire une classe de modèles évolutifs autorégressifs qui, en dépit de leur non-stationnarité, restent stables à chaque instant. Ces modèles sont représentés par des filtres en treillis dont les coefficients de réflexion sont non pas décomposés linéairement sur une base de fonctions mais subissent une transformation non-linéaire qui garantit la stabilité du modèle.

Les résultats issus d'expériences de restitution montrent l'intérêt de la technique évolutive en analyse de la parole: l'estimation d'un modèle s'effectue sur des entités de type syllabique, la restitution étant ensuite réalisée en excitant le modèle ainsi déterminé au moyen du résidu préalablement codé. Des exemples de signaux ainsi synthétisés permettront de juger la qualité de la synthèse obtenue.

2. MODELES EVOLUTIFS.

Dans cette approche, le signal non-stationnaire est représenté par la sortie d'un filtre autorégressif linéaire dont les coefficients sont décomposés sur une base de fonctions. L'idée, émise par Rao /1/ et Mendel /2/ au début des années 1970, a été développée par Liporace /3/, Hall et al. /4/ qui ont essentiellement étudié l'estimation des paramètres d'un modèle autorégressif à structure transverse. L'un des auteurs de ce papier a généralisé cette technique aux filtres en treillis ainsi qu'aux modèles MA (modèle à Moyenne Ajustée) et ARMA (AutoRégressif à Moyenne Ajustée) /5/.

2.1. Structure transverse.

Supposons le signal y_t observé sur l'intervalle de temps $[0,T]$. La relation définissant le modèle AR à coefficients dépendant du temps est donnée par (1) en fonction de l'innovation ϵ_t:

$$y_t + a_1(t-1)y_{t-1} + \ldots + a_p(t-p)y_{t-p} = \epsilon_t \qquad (1)$$

L'hypothèse fondamentale est la suivante: chacun des coefficients $a_i(t)$ peut être approché par une combinaison linéaire sur une base de fonctions $[f_0(t),\ldots,f_m(t)]$, choisies à priori et définies sur le même intervalle de temps $[0,T]$:

$$a_i(t) = \sum_{j=0}^{m} a_{ij} f_j(t) = [a_{i0} \ldots a_{im}]F(t) \qquad (2)$$

Sous cette hypothèse, le problème est réduit à l'estimation des $p(m+1)$ scalaires $(a_{ij}, 1\leq i\leq p, 0\leq j\leq m)$.

L'innovation ϵ_t apparaissant dans (1) est à variance, soit constante $E(\epsilon_t^2)=\sigma^2$, soit fonction du temps $E(\epsilon_t^2)=\sigma^2(t)$. Dans ce dernier cas, le paramètre $\sigma(t)$ peut être décomposé, de la même manière que les coefficients de prédiction (2), sur la base de fonctions:

$$\sigma(t) = \sum_{j=0}^{m} \sigma_j f_j(t) \qquad (3)$$

Montrons comment estimer, de manière simple, les coefficients $\{a_{ij}\}$. Pour cela introduisons le vecteur Y_t des projections du signal $\{y_t\}$ sur la base de fonctions $f_j(t)$, Y_t est défini par (4):

$$Y_t = F(t)y_t \qquad (4)$$

Avec une telle notation et en utilisant (2), l'équation (1) peut se réécrire ainsi:

$$y_t + [Y_{t-1}^T \cdots Y_{t-p}^T]\theta = \epsilon_t \qquad (5)$$

où θ est le vecteur contenant les paramètres inconnus $[a_{10}\cdots a_{1m}\ a_{20}\cdots a_{pm}]$:

$$\theta = [a_{10}\cdots a_{1m}\ a_{20}\cdots a_{pm}]^T \qquad (6)$$

(le signe 'T' indique que le vecteur doit être transposé).

(5) montre que la modélisation non-stationnaire d'un signal scalaire y_t a été réduite à la modélisation d'un signal vectoriel donc plus complexe mais pour lequel le modèle à estimer est stationnaire. Si la variance de l'innovation ϵ_t est constante, le vecteur θ des paramètres à identifier peut facilement être déterminé en maximisant la vraisemblance $p(y_0\cdots y_t|\theta)$; une telle démarche conduit aux équations de Yule-Walker (7):

$$\sum_{t=p}^{T} \begin{bmatrix} Y_{t-1} \\ \cdot \\ \cdot \\ Y_{t-p} \end{bmatrix} [Y_{t-1}^T \cdots Y_{t-p}^T]\theta = -\sum_{t=p}^{T} \begin{bmatrix} Y_{t-1} \\ \cdot \\ \cdot \\ Y_{t-p} \end{bmatrix} Y_t \qquad (7)$$

2.2. Structure en treillis.

Dans le cas stationnaire, la structure en treillis est souvent utilisée à la place de la structure transverse. Dans le cadre de la modélisation non-stationnaire, la structure globale du filtre n'est pas modifiée . Chaque cellule a la même structure interne (figure 1) définie par (8) mais qui diffère du cas stationnaire par la présence de deux opérateurs retard au lieu d'un seul et de deux coefficients de réflexion $k_i^+(t)$ et $k_i^-(t)$.

$\epsilon_i^+(t)$ représente l'erreur de prédiction directe obtenue quand y_t est prédit à partir de $\{y_{t-1},\ldots,y_{t-i}\}$ tandis que $\epsilon_i^-(t)$ désigne l'erreur de prédiction rétrograde obtenue par prédiction de $\{y_{t-i}\}$ à partir de $\{y_t\cdots y_{t-i+1}\}$.

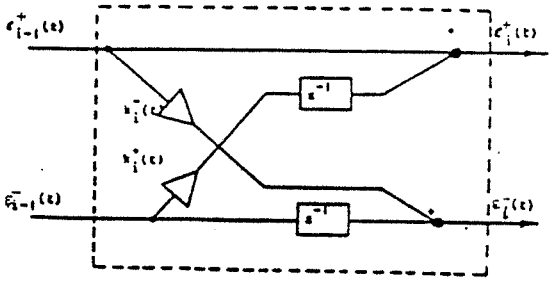


figure 1: structure d'une cellule à deux retards.

$$\begin{bmatrix} \epsilon_i^+(t) \\ \epsilon_i^-(t) \end{bmatrix} = \begin{bmatrix} 1 & k_i^+(t-1) \\ k_i^-(t) & 1 \end{bmatrix} \begin{bmatrix} \epsilon_{i-1}^+(t) \\ \epsilon_{i-1}^-(t-1) \end{bmatrix} \qquad (8)$$

L'identification du modèle (8) est calquée sur les méthodes qui existent dans le cas stationnaire, en postulant que chaque coefficient de réflexion s'exprime sur la base de fonctions:

$$k_i^+(t) = \sum_{j=0}^{m} k_{ij}^+ f_j(t) \text{ et } k_i^-(t) = \sum_{j=0}^{m} k_{ij}^- f_j(t) \qquad (9)$$

De la même manière que dans la méthode de Burg, on cherche à optimiser les cellules l'une après l'autre, en initialisant le filtre avec $\epsilon_0^+=\epsilon_0^-=y_t$. Le calcul de la $i^{ème}$ cellule s'effectue en minimisant l'énergie des erreurs de prédiction $\epsilon_i^+(t)$ et $\epsilon_i^-(t)$.

L'utilisation de ces modèles évolutifs en restitution et à plus long terme en synthèse de parole est-elle valable? Une difficulté de première évidence est le manque de stabilité des modèles estimés.

Prenons l'exemple de la structure transverse et désignons par $A_t(z)$ la transformée en z de la séquence $\{1,a_1(t),\ldots,a_p(t)\}$, les coefficients de cette séquence étant fixés à l'instant t. Dans le cadre de la restitution, nous nous intéressons à la stabilité locale du modèle autorégressif défini par (1). En effet, une telle stabilité assure qu'à tout instant les zéros de $A_t(z)$ restent intérieurs au cercle unité, ceci pour éviter les sauts d'énergie qu'induisent les excursions des zéros hors du cercle unité, et qui ont pour conséquence de dégrader la parole synthétisée.

Mais aucun fondement théorique n'assurant cette condition, une solution consiste à renvoyer à l'intérieur du cercle unité tous les zéros dont le module est plus grand que un, en utilisant lors de la phase de synthèse du signal une procédure telle que la factorisation de Schur: $A_t(z)$ est remplacé par un polynôme $\tilde{A}_t(z)$ de même gain mais à phase minimale et tel que $\tilde{A}_t(z)\tilde{A}_t(z^{-1}) = A_t(z)A_t(z^{-1})$. L'algorithme utilisé est le suivant:

a) Détection des instabilités de $A_t(z)$ à l'aide du critère de Schur-Cohn.

b) Calcul de la corrélation $R_t(z)$ de la séquence $\{1,a_1(t),\ldots,a_p(t)\}$:

$$R_t(z) = A_t(z)A_t(z^{-1})$$

c) Factorisation de $R_t(z)$ en un produit de deux facteurs (utilisation de la paramétrisation de Schur de $A_t(z)$): l'un contient la partie stable $\tilde{A}_t(z)$, l'autre la partie instable $\tilde{A}_t(z^{-1})$.

Mais une telle procédure nécessite de nombreux calculs, ce qui a pour conséquence d'augmenter de manière importante le temps de restitution du signal.

Une autre possibilité est d'imposer la stabilité du modèle, non pas lors de la restitution, mais dès l'analyse du signal. Mais avant de présenter la solution que nous avons mise en oeuvre, remarquons que l'estimation d'un filtre en treillis présente le même inconvénient que précédemment: même si les coefficients de réflexion $k_i(t)$ sont de module inférieur à 1, leur approximation par une décomposition linéaire sur une base de fonctions n'est pas contrainte à appartenir à l'intervalle $]-1,+1[$. Une solution /8/ consiste à forcer la stabilité du modèle au moyen d'une transformation non-linéaire des coefficients de réflexion, de sorte que l'intervalle de stabilité ne soit plus défini par $]-1,+1[$ mais soit étendu à $]-\infty,+\infty[$, le nouveau jeu de coefficients étant à estimer sur la même base de fonctions. La fonction retenue transforme les $k_i(t)$, $(1\leq i\leq p)$ en Log Area Ratios ou fonctions d'aire logarithmiques définies par (10):

$$\gamma_i(t)=\mathrm{Log}\left(\frac{1+k_i(t)}{1-k_i(t)}\right) \text{ avec } \gamma_i(t)=\sum_{j=0}^{m}\gamma_{ij}f_j(t) \quad (10)$$

L'intérêt de déterminer les $\{\gamma_{ij}\}$ est d'assurer, en pratique, la stabilité du modèle estimé et donc de supprimer toute procédure de stabilisation.

2.3. LAR par l'intermédiaire d'un treillis.

L'estimation directe des $\{\gamma_{ij}\}$ consisterait à minimiser, à la sortie de la $i^{ème}$ cellule $(1\leq i\leq p)$, la somme du carré des erreurs. Mais une telle procédure conduit à un problème non-linéaire et coûteux en calculs. Une autre solution est de résoudre le problème en deux étapes, chacune d'entre elles étant linéaire: dans un premier temps, le filtre en treillis est déterminé par les projections $\{k_{ij}\}$ de ses coefficients de réflexion sur la base de fonctions; les trajectoires des LAR sont ensuite approchées par celles des $k_i(t)$:

$$\sum_{t\in T} F(t)F^T(t)\begin{bmatrix}\gamma_{i0}\\ .\\ .\\ .\\ \gamma_{im}\end{bmatrix} = \sum_{t\in T} F(t)\,\log\left(\frac{1+k_i(t)}{1-k_i(t)}\right)$$

Bien qu'une telle méthode soit approchée, elle permet toutefois d'obtenir des résultats satisfaisants; le coût de calculs lors de la phase d'analyse est cependant plus important que pour l'estimation d'un modèle autorégressif tel que celui à structure transverse: il faut tout d'abord estimer un filtre en treillis et ensuite calculer les LAR à partir des coefficients de réflexion obtenus; mais cet inconvénient est contrebalancé lors de la restitution du signal puisqu'il n'est plus nécessaire d'introduire une méthode de stabilisation.

3. EXPERIENCES EN RESTITUTION DE PAROLE.

La technique de modélisation évolutive décrite au paragraphe précédent est actuellement appliquée à des expériences d'analyse et de restitution de parole. Le propos de ce paragraphe est de montrer la validité de la méthode et de définir des critères pour déterminer les différents paramètres caractérisant le modèle non-stationnaire, à savoir le type de la base de fonctions, son degre m et l'ordre AR p. Chaque expérience comporte les mêmes phases que nous allons décrire dans les lignes qui suivent.

3.1. Segmentation.

Le signal est segmenté, un modèle devant être estimé pour chacun des segments obtenus: l'analyse s'effectue sur chaque segment débutant et finissant à l'un des points de segmentation.

Les modèles évolutifs permettant de représenter une transition entre contenus spectraux différents. Le but d'une première série d'expériences a été de réaliser une synthèse par diphones (segmentation aux maxima de la stabilité spectrale): ceci implique la concaténation de modèles sur des maxima d'énergie et la nécessité d'assurer une continuité des modèles à ces instants. La difficulté à réaliser cette opération nous a conduits à nous orienter vers une synthèse utilisant des entités de type syllabique (entités plus longues que les diphones). Dans ce cas la segmentation, c'est à dire la détection des limites entre syllabes, se fait en coupant le signal aux instants où sa courbe d'énergie calculée sous une fenêtre glissante passe par un minimum. Ainsi les problèmes liés à la concaténation se produiront dans les zônes les moins audibles.

3.2. Analyse.

Cette étape consiste à obtenir, pour chaque segment, un modèle autorégressif évolutif. C'est ici que se situe la variable essentielle de l'expérience, caractérisée par le choix des différents paramètres définissant la structure du modèle.

3.2.1. Choix d'une base de fonctions.

Contraindre les fonctions $(f_i(t))$ à être orthogonales n'est pas nécessaire pour l'estimation du modèle; toutefois cela permet d'améliorer l'estimation par un meilleur conditionnement des systèmes à résoudre. Trois types de base (base de Legendre, série de Fourier et fonctions sphéroidales aplaties) ont été étudiés.

a) l'introduction de la base des polynômes de Legendre permet d'annuler les dérivées $n^{èmes}$ des $a_i(t)$: ceci a pour effet d'accentuer le lissage.

b) une limitation aux premiers coefficients de la série de Fourier accentue l'absence des hautes fréquences dans l'évolution des $a_i(t)$.

c) les fonctions sphéroidales aplaties sont

celles qui approchent au mieux les fonctions à bande limitée, sans pour autant introduire une hypothèse telle que la périodicité (cas de la base de Fourier).

Les expériences d'analyse et de synthèse réalisées avec ces trois types de fonctions montrent que les résultats obtenus sont équivalents.

3.2.2. Sélection de l'ordre.

Pour un jeu de fonctions données, la mise en pratique des modèles évolutifs nécessite l'estimation du nombre m de fonctions ainsi que de l'ordre autorégressif p. Seuls Kozin, Nakajima /7/ ont considéré ce point: ils montrent que la consistence et la normalité des paramètres, estimés au sens du maximum de vraisemblance, assurent la validité du critère d'Akaike /6/ qui, adapté au cas des modèles évolutifs, est défini comme suit:

$$C(p,m) = N\log(\sigma^2_{p,m}) + 2p(m+1)$$

où N est le nombre d'échantillons dans le segment de parole et $\sigma^2_{p,m}$ est une estimation de la variance de l'innovation pour p et m fixés.

Les valeurs de p et m sont choisies de sorte à minimiser C(p,m). La figure 2 représente le critère pour des valeurs de p comprises entre 6 et 16 et de m entre 3 et 6. Dans cet exemple, les valeurs 'optimales' de p et m sont respectivement 10 et 5.

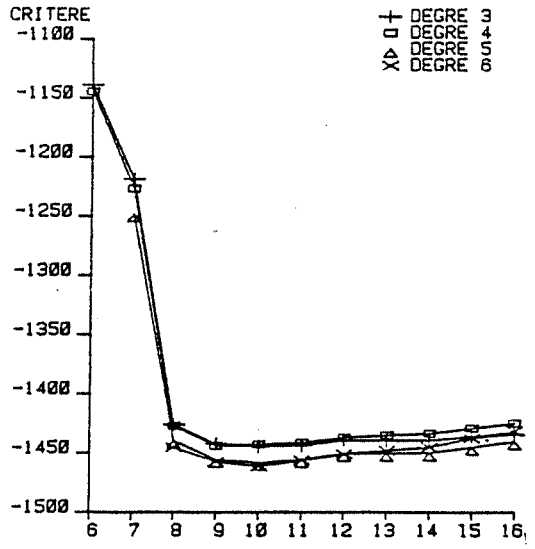


figure 2: tracé du critère d'Akaike (ordonnée) pour différentes valeurs de l'ordre (abscisse) et du degré (paramètre), les modèles étant estimés sur un segment de parole de durée 100 ms. Ordre et degré optimaux sont respectivement 10 et 5.

La validité de ce critère a également été vérifiée expérimentalement: un seul paramètre (ordre ou degré) varie à chaque expérience, la qualité du signal synthétisé étant évaluée de manière subjective. Au delà d'un ordre 12 et d'un degré 5 (pour une fréquence d'échantillonnage de 8 kHz), aucune amélioration sensible de la qualité de la synthèse n'apparait.

3.2.3. Choix du type de coefficients.

Nous avons testé expérimentalement chacun des trois jeux de coefficients: coefficients autorégressifs ($a_i(t)$), coefficients de réflexion ($k_i(t)$) et Log Area Ratios ($\gamma_i(t)$). La qualité de la synthèse obtenue au moyen des ($k_i(t)$) et des ($\gamma_i(t)$) est légèrement meilleure que celle issue d'un modèle à structure transverse. Toutefois le coût de calcul, lors de la phase d'analyse, est moins important dans ce dernier cas que dans les deux premiers: en pratique, l'estimation des Log Area Ratios sur une base de degré 5 (m=5) nécessite un nombre d'opérations trois fois plus élevé que l'estimation des coefficients autorégressifs sur la même base (rappelons qu'il faut tout d'abord estimer un filtre en treillis et ensuite calculer les LAR à partir des coefficients de réflexion obtenus). Mais cet inconvénient est contrebalancé lors de la restitution du signal puisqu'il n'est plus nécessaire d'introduire une méthode de stabilisation.

3.3. Restitution.

L'entrée idéale pour exciter le modèle AR serait l'innovation ϵ_t calculée à partir de (1): ceci permettrait d'obtenir une qualité de synthèse analogue à celle du signal de parole original, mais le débit associé atteindrait une valeur inacceptable! Différentes techniques permettent de coder l'innovation ou résidu ϵ_t. Auparavant, notons que dans le cas d'un modèle non-stationnaire ce résidu doit être normalisé, ce qui nécessite l'estimation de sa variance $E\{\sigma^2(t)\}$: $\epsilon_t = \sigma(t)e_t$, e_t étant l'entrée normalisée qui sera codée.

e_t étant gaussien avec une variance unité, la détermination de $\sigma(t)$ peut être réduite à un problème linéaire en supposant que ϵ_t est gaussien et en considérant le signal $|\epsilon_t|$. La décomposition de $\sigma(t)$ sur la base de fonctions (3), permet d'estimer les coefficients de pondération σ_i qui minimisent au sens des moindres carrés l'erreur entre $\sigma(t)$ et $\sqrt{\frac{\pi}{2}}|\epsilon_t|$.

L'entrée normalisée, e_t, peut être codée en utilisant différentes techniques: codage en bande de base, technique multi-impulsionnelle, quantification sphérique vectorielle, détection de la fréquence fondamentale ... La technique multi-impulsionnelle proposée par Atal /9/ donne d'excellents résultats dans le cas de la modélisation stationnaire; c'est pourquoi, cette méthode a été étendue à la modélisation évolutive /10/ bien que la complexité de calculs soit plus importante que précédemment, la réponse impulsionnelle non-stationnaire dépendant de la position de l'impulsion.

La figure 3 montre un exemple de signal de parole original (a), le résidu obtenu par filtrage inverse de ce signal à travers le modèle évolutif (b), l'excitation multi-impulsionnelle (c) et le signal synthétique correspondant (d). La durée du signal étudié étant de 200 ms, 200 impulsions ont été positionnées afin d'obtenir, en moyenne, une impulsion toutes les millisecondes.

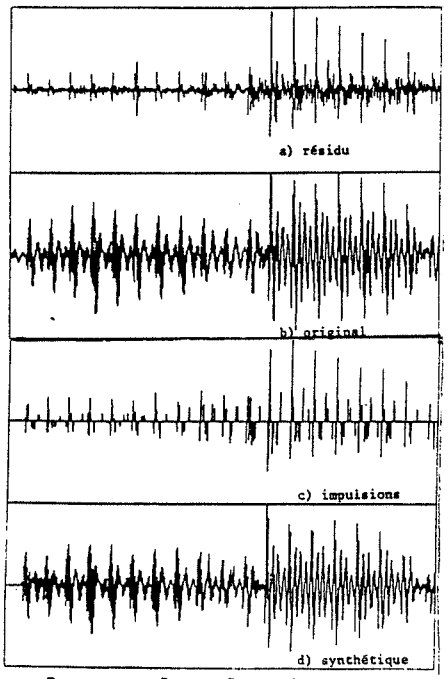


figure 3: exemples de signaux obtenus par analyse évolutive et codage multi-impulsionnel -durée du signal: 200 ms-

CONCLUSION.

Les résultats obtenus valident l'intérêt des modèles évolutifs en traitement de la parole, une telle méthode permettant de paramétrer de manière globale l'évolution du signal sur le segment analysé.

Nous avons présenté le principe d'un algorithme pour estimer des fonctions d'aire logarithmiques à partir d'un treillis. Bien que, lors de la phase d'analyse, cette technique soit plus coûteuse en calculs, elle permet toutefois d'améliorer le temps de restitution du signal. L'ordre du modèle ainsi que le degré de la base de fonctions peuvent être déterminés à l'aide du critère d'Akaike, le choix de l'une des trois bases donnant des résultats équivalents.

A l'issue de précédentes expériences, nous avons trouvé que les coefficients du modèle doivent être codés à raison d'environ 250 par seconde. Ceci est à comparer aux 500 coefficients par seconde que représente la LPC-10 (prédiction linéaire à l'ordre 10, un modèle toutes les 20 ms). Bien que le problème du codage de l'ensemble des paramètres du modèle évolutif reste encore à étudier, on peut espérer un gain d'un facteur 2 sur la LPC-10 avec une qualité analogue, si ce n'est meilleure...

REFERENCES.

/1/ T.S. Rao, 'The fitting of non-stationary time-series models with time-dependent parameters', J. of the Royal Statist. Soc., Series B, vol.32, n°2, pp 312-322, 1970.

/2/ J.M. Mendel, 'A priori and a posteriori identification of time-varying parameters', 2nd Hawai Int. Conf. on Syst. Sciences, pp 207-210, 1969.

/3/ L.A. Liporace, 'Linear estimation of non-stationary signals', J. Acoust. Soc. Amer., Vol.58, n°6, pp 1218-1295,1975.

/4/ M. Hall, A.V. Oppenheim, A. Willsky, 'Time-varying parametric modelling of speech', Signal Processing, Vol.5, n°3, pp 267-285, 1983.

/5/ Y. Grenier, 'Time-dependent ARMA modeling of non-stationary signals', IEEE Trans. on ASSP, Vol.31, n°4, pp 899-911, 1983.

/6/ H. Akaike, 'A new-look at the statistical model identification', IEEE trans. Aut. Contr., Vol.19, n°6, pp716-723, 1974.

/7/ F. Kozin, F.Nakajima, 'The order determination problem for linear time-varying AR models', IEEE Trans. on AC, Vol.25, n°2, pp250-257.

/8/ M.C. Chevalier, Y. Grenier, 'Autoregressive models with time-varying Log Area Ratio', IEEE ICASSP-85.

/9/ B.S. Atal and J.R. Remde, 'A new model of LPC excitation for producing natural-sounding speech at low bit rates', IEEE ICASSP-82, pp614-617, 1982.

/10/M.C. Omnes-Chevalier, Y. Grenier, G. Chollet, 'Codage multi-impulsionnel pour la restitution de parole par modèles évolutifs', GRETSI-85.

/11/D. Slepian, 'Prolate spheroidal wavefunctions, Fourier analysis and uncertainty-V: the discrete case', Bell. Syst. Tech. J., vol.57, pp.1371-1430, 1978.

PASSAGE AUTOMATIQUE DU TEXTE ORTHOGRAPHIQUE VERS LE TEXTE PHONETIQUE

V. Aubergé

(I.M.S.S. / C.R.I.S.S.)
Institut de la Communication Parlée Grenoble

Le passage d'une chaîne orthographique vers la chaîne de sons correspondante utilise plusieurs niveaux de connaissances définis par leur unité linguistique significative (lettre, constituant du mot, mot dans son contexte énonciatif).

Le domaine de la lettre est décrit par un ensemble de règles de transduction contextuelles.

Nous avons mis au point un formalisme et une méthodologie algorithmique s'appuyant sur les expériences antérieures (DIVAY & GUYOMARD, 1977; LETY, 1980; PROUTS, 1980).

Nous avons envisagé la synthèse d'une langue écrite comme étant une succession chaînée d'étapes. Chacun de ces processus étant construit pour être adaptable à chacune des langues considérées.

Actuellement le noyau de ces langues est constitué de l'Italien et de l'Allemand. Les connaissances accumulées sur le Français et l'Anglais nous permettent de savoir que nous pourrons les intégrer prochainement.

Le processus auquel nous nous sommes intéressée ici est le passage de la chaîne linguistique écrite vers la chaîne des symboles des sons correspondante. Nous n'entrerons pas dans les détails linguistiques dans ce propos, notre intention étant d'y décrire l'outil informatique issu des connaissances réunies par nos chercheurs.

Le logiciel élaboré jusqu'à présent s'intéresse aux données linguistiques qui examinent le texte écrit comme une succession de lettres (le qualificatif graphème étant encore mal défini).

En n'utilisant pas d'unités linguistiques supérieures, on obtient cependant une traduction phonétique acceptable.

Nous avons donc essayé, sur la base des travaux déjà effectués en Français et sur l'expérience qui en a été retirée, d'établir un formalisme d'expression de ces règles, le plus souple et agréable possible pour un linguiste.

On montre qu'on obtient ainsi un transducteur fini de type contextuel dont voici la description syntaxique du langage :

- en majuscule les éléments terminaux;
- '[' et ']' désignent les éléments facultatifs.

```
grammaire     ::=  règle*

règle         ::=  partie_g = partie_d

partie_d      ::=  (SON*)

partie_g      ::=  [cxt] + ch_à_trans + [cxt]

cxt           ::=   élt_cxt [+élt_cxt]*

élt_cxt       ::=   unité_cxt [,unité_cxt]*

unité_cxt     ::=   [/]  "CLASSE"
                  ::=   [/]   LETTRE
                  ::=   [/]   FIN_DE_MOT
                  ::=    &

ch_à_trans    ::=    LETTRE+
```

N.A.B.N.

/ est l'opérateur "sauf"

LETTRE = un élément quelconque du Vocabulaire d'entrée : Ve

SON = un élément du Vocabulaire de sortie : Vs

"CLASSE" est un sous-ensemble de Ve

& suit un élément de CLASSE et le répète (géminée)

Exemples : (sans réalités linguistique !!)

si "VOY" est une CLASSE :

```
(d,s + "VOY") + x + (FIN_DE_MOT)  =  (S)
              + eau +             =  (O)
```

L'ordre induit au transducteur (selon le ruban d'entrée de gauche à droite), est reproduit naturellement par une partition sur les règles selon le premier caractère à transcrire pour chaque règle. On introduit de plus un ordre local à chaque partition, ordre d'application sur les règles défini soit par une fonction d'ordre, soit par point par le linguiste (par exemple : ordre d'écriture des règles); ceci a l'avantage de simplifier considérablement l'écriture des règles, l'explicitation de certains contextes étant rendu implicite par la relation d'ordre.

La définition complète et non ambiguë de ce langage nous permet, pour chaque Ve et Vs (selon la langue), de bénéficier d'aide à l'écriture de ces grammaires : analyse syntaxique (LL1);
contrôle logique ...
et de réaliser indépendamment interpréteur ou compilateur.

BIBLIOGRAPHIE

[1] M. Divay & M. Guyomard, "Conception et réalisation sur ordinateur de transcription graphémo-phonétique du Français", Thèse de 3ème cycle, Informatique, Univ. Rennes, 1977.

[2] H. Fervers, J. Le Roux & L. Miclet, "Programme de transcription orthographique phonémique en langue française", Document E.N.S.T. - D - 76003, 1976.

[3] S. Hertz, "Interactive Speech Synthesis with Linguistic Rules", Cornell Working Papers in Linguistics, Cornell University, 1980.

[4] M. Lety, "Transcription orthographique-phonétique", Thèse de 3ème cycle, Grenoble, Juin 1980.

[5] B. Pratt & G. Silva, "Phontrns", Published by Monash University, Aôut 1967.

[6] B. Prouts, "Contributions à la synthèse de la parole à partir du texte; transcription graphème-phonème en temps rêel sur micro-processeur", Thèse de Docteur-Ingénieur, Univ. Paris Sud. Centre d'Ursay, Novembre I980.

SYNTHESE DE MESSAGES ORAUX A PARTIR D'UNE REPRESENTATION SEMANTIQUE

L. Danlos, F. Emerard, C. Sorin

Laboratoire d'Automatique Documentaire et Linguistique, CNRS,
2, Place Jussieu, 75005 Paris
Centre National d'Etudes des Telecommunications,
22301 Lannion

SUMMARY

In working toward spoken man-machine dialogue, an attractive solution is to base speech synthesis on semantic representation. In order to achieve this, a text translating the semantic representation must first be generated and then transmitted to a speech synthesis system. The aim of this article is to describe the step linking two existing systems carried out independently : automatic representation of written texts and speech synthesis. Thanks to this step, satisfactory automatic prosodic processing is made possible.

INTRODUCTION

Cet article présente un logiciel qui engendre un message oral à partir d'une représentation sémantique des informations qui doivent être véhiculées. Il intègre deux systèmes : un système de génération de textes écrits traduisant une représentation sémantique /1/ et un système de synthèse de la parole /2/. L'articulation entre le générateur et le synthétiseur de parole est constituée de textes écrits annotés d'informations prosodiques sous forme de marqueurs. En effet, le synthétiseur de parole ne peut produire de textes "naturels" qu'à condition qu'on lui fournisse des informations prosodiques qui jusqu'alors étaient introduites manuellement ; il a donc fallu d'une part formaliser le positionnement de ces marqueurs en dégageant une représentation syntaxico-prosodique des phrases compatible avec la représentation syntaxique utilisée dans le générateur et d'autre part développer un algorithme qui engendre automatiquement ces marqueurs.

I- PRESENTATION DU GENERATEUR ET DES MARQUEURS PROSODIQUES.

I-1- LE GENERATEUR.

Le générateur est constitué de deux modules : le composant stratégique et le composant syntaxique. A partir d'une représentation sémantique, le composant stratégique prend les décisions conceptuelles (ordre d'apparition des informations, détermination des informations qui doivent être exprimées explicitement et de celles qui peuvent être laissées implicites) et les décisions linguistiques (choix des formes verbales et des constructions syntaxiques, découpage du texte en phrases). Ce composant fournit un "schéma de phrases" indiquant la linéarisation du message et les formes verbales. Ce schéma de phrase est transmis au composant syntaxique qui le développe en phrases. Pour cela, un schéma de phrase est annoté de marqueurs qui indiquent les catégories syntaxiques de ses éléments. Ces catégories, qui sont hiérarchisées, permettent d'appliquer les règles de grammaire (règles d'accord, placement de la négation, forme des pronoms, etc...). Ce sont elles qui sont à la base du positionnement des marqueurs prosodiques : elles ont été adaptées afin d'obtenir des catégories syntaxico-prosodiques pertinentes pour le traitement prosodique.

I-2- LES MARQUEURS PROSODIQUES.

Le comportement prosodique d'un locuteur a été étudié, une simplification acceptable de ce comportement a été opérée. Cette schématisation, qui donne des solutions prosodiques toujours plausibles, constitue le traitement prosodique du système de synthèse /3/.

Les 1200 diphones du dictionnaire ont été stockés avec une durée spécifique et figée, déterminée par l'opération de segmentation ; puis les valeurs de Fo attachées à chaque trame de chacun des segments ont été éliminées. En revanche, chaque diphone du dictionnaire a été enrichi d'informations de type phonémique permettant de connaître les frontières des segments de parole. Pour tout énoncé de nature énonciative, impérative, interrogative, une quinzaine de points prosodiques a été retenue parce que considérée comme suffisante pour assurer une prosodie de phrase satisfaisante à la synthèse. Ces points sont concrètement matérialisés par des "marqueurs". Le traitement exploite à la fois des "marqueurs naturels" (ceux qui apparaissent sur tout message écrit : l'espace entre les mots, le point en fin de phrase et la virgule par exemple) et des "marqueurs artificiels" (# $ *) qui peuvent se substituer aux marqueurs naturels ; les uns comme les autres exécutent une

action mélodique et rythmique spécifique sur chacune des syllabes du mot qui les précède.

Les marqueurs artificiels se divisent en deux classes:

- Classe 1:

- pas de rupture dans la concaténation des diphones,
- pas de pause,
- faible amplitude de variation de Fo sur la dernière voyelle de mot, le schéma de Fo peut être descendant ou montant,
- faible allongement de durée sur la dernière voyelle de mot.

Cette classe ne comporte que deux marqueurs qui se différencient par le sens de la pente de Fo.

- Classe 2:

- rupture dans la concaténation des diphones,
- introduction d'une pause,
- grande amplitude de variation de Fo sur la dernière voyelle, le schéma de Fo peut être descendant ou montant,
- allongement de durée important sur la voyelle finale.

Ce sont ces marqueurs artificiels -jusque là introduits manuellement sur la chaîne orthographique- appartenant à ces deux classes que l'on souhaite positionner automatiquement à partir de la structure syntaxico-prosodique.

II- PRINCIPES DE DERIVATION DES MARQUEURS PROSODIQUES.

II-1- LA REPRESENTATION PROSODIQUE

On appellera "intervalle prosodique" une suite de mots suivie d'un marqueur prosodique. Un intervalle prosodique est noté "A x" où A est la suite de mots, x le marqueur prosodique. La phrase -suite de mots suivie par un point, point_virgule ou point d' interrogation- est un intervalle prosodique : son signe de ponctuation final correspond à un marqueur prosodique. La phrase est à la fois l'unité syntaxique maximale et l'unité prosodique maximale : c'est à l'intérieur de cette unité que s'appliquent les règles de grammaire et le positionnement des marqueurs prosodiques. La structure syntaxique munit la phrase d'une représentation arborescente. Le traitement prosodique de la phrase requiert une structure arborescente.

En effet, d'une part les marqueurs prosodiques ne doivent pas être introduits simultanément, mais progressivement, d'autre part les marqueurs introduits à une étape ne sont pas altérés à une étape ultérieure. Le positionnement des marqueurs dans une phrase revient donc à construire un arbre dont la racine est l'intervalle prosodique constitué par la phrase, dont les noeuds sont des intervalles prosodiques, dont les fils d'un noeud I=A2 x2 sont les intervalles prosodiques I1=A1 x1, I2=A2 x2, ...In=An x où x1 x2..xn-1 sont les marqueurs prosodiques introduits en une fois dans la séquence A=A1 A2..An. Les feuilles de l'arbre sont des intervalles prosodiques non segmentables, c'est_à_dire des intervalles prosodiques dans lesquels on ne peut plus introduire de marqueurs. La concaténation de ces feuilles représente la phrase enrichie de ses marqueurs prosodiques. Cette structure prosodique est obtenue par l'application récursive d'une opération de base : la segmentation d'un intervalle I=A x en sous_intervalles I1=A1 x1, I1=A2 x2 ..In=An x.

II-2- LA SEGMENTATION D'UN INTERVALLE PROSODIQUE.

Cette opération nécessite deux types de connaissance :
- l'emplacement des marqueurs, c'est à dire la détermination des sous séquences A1 A2..An de A, les marqueurs étant placés entre Ai et Ai+1,
- la valeur des marqueurs, c'est à dire la détermination de x1, x2,..xn-1.

II-2-1- EMPLACEMENT DES MARQUEURS.

L'emplacement des marqueurs est calculé à l'aide d'une structure qui, d'une part est dérivable de la structure syntaxique, d'autre part est superposable à la structure prosodique. On l'appellera donc structure "syntaxico-prosodique". Cette structure est définie par un ensemble hiérarchisé de catégories syntaxico-prosodiques. Une catégorie syntaxico-prosodique est soit une catégorie syntaxique, soit un regroupement de catégories syntaxiques, soit une séquence de catégories syntaxiques.

Ainsi les catégories syntaxico-prosodiques "phrase" (P) et "forme verbale" (FV) correspondent aux catégories syntaxiques phrase et forme verbale ; la catégorie syntaxico-prosodique "complément de verbe" (CV) regroupe les catégories syntaxiques sujet, objet, prép-objet ; la catégorie syntaxico-prosodique "séquence post-verbale" (SPOV) est constituée de la séquence des catégories syntaxiques apparaissant après la forme verbale.

Elle est superposable à la structure prosodique dans le sens où la suite de mots A composant un intervalle prosodique correspond toujours à une catégorie syntaxico-prosodique. La décomposition de A en sous_séquences A1 A2..An correspond à l'analyse de la catégorie syntaxico-prosodique de A en sous_catégories syntaxico-prosodiques. Autrement dit, les arbres prosodiques et syntaxico-prosodiques sont superposables.

II-2-2- VALEUR DES MARQUEURS.

La valeur des marqueurs est calculée en deux temps :

-dans un premier temps, on détermine si un marqueur est de la Classe 1 ou de la Classe 2,
-dans un deuxième temps, on calcule l'élément de la Classe.

II-2-2-1- LA CLASSE DES MARQUEURS.

La détermination de la classe des marqueurs s'effectue à l'aide des notions de "relation de dépendance" et "relation d'indépendance" qui viennent se greffer sur la structure syntaxico-prosodique. Lorsqu'une catégorie syntaxico-prosodique s'analyse en sous_catégories syntaxico-prosodiques, on dira que celles-ci sont en relation de dépendance (resp. d'indépendance) si elles doivent être séparées par un marqueur de la Classe 1 (resp. de la Classe 2). Ainsi la séquence post-verbale "le frère de Max avec attention" dans la phrase "Marie a regardé le frère de Max avec attention" s'analyse en

((le frère de Max)(avec attention))
SPOV CV CP

avec une relation d'indépendance entre les catégories CV et CP. Par contre, le complément de verbe "le frère de Max" s'analyse en

((le frère)(de Max))
CV N PN

avec une relation de dépendance entre les catégories N et PN.

II-2-2-2- L'ELEMENT DE LA CLASSE.

Le calcul de la valeur d'un marqueur, connaissant sa classe, s'effectue mécaniquement à partir des règles suivantes :

- Un marqueur de la classe 2 introduit à l'intérieur d'un intervalle prosodique A x est une fonction de x, il est de pente opposée à x (cependant, il arrive parfois qu'un marqueur de la classe 2 soit aussi fonction de la catégorie syntaxico-prosodique de A).

- Un marqueur de la classe 1 introduit dans un intervalle prosodique n'est pas déterminé ; il le sera à la fin des opérations de segmentation. Provisoirement, on lui affecte comme valeur □, □ étant un marqueur postiche. A la fin des opérations de segmentation, c'est_à_dire quand on a obtenu une suite d'intervalles prosodiques non segmentables, la réécriture de □ est une opération qui s'effectue de droite à gauche en obéissant à un principe d'alternance de montées et de descentes de Fo ; plus précisément si le marqueur prosodique x situé à droite de □ monte (resp. descend), □ a pour valeur le marqueur de dépendance qui descend (resp. monte).

En résumé, la structure prosodique d'une phrase est obtenue à partir de sa structure syntaxico-prosodique et des relations de dépendance et d'indépendance établies entre catégories syntaxico-prosodiques. Cette connaissance détermine l'emplacement des marqueurs et leur classe. Quant à leur valeur, elle relève de la philosophie même du traitement prosodique qui prend en compte l'influence du contexte droit.

III- DONNEES NECESSAIRES POUR LA DERIVATION.

La catégorie syntaxico-prosodique "phrase" (P) se décompose en "phrases syntaxiques" (PS) avec des relations d'indépendance entre phrases syntaxiques. Ainsi la phrase

((Max a donné un livre à Marie)
P PS
(parce qu'il voulait lui faire plaisir)).
PS

est composée de deux PS. La phrase

((Max a donné ce livre à Marie)).
P PS

est réduite à une seule phrase syntaxique. La catégorie PS se décompose en
- séquence pré-verbale (SPRV),
- forme verbale (FV),
- séquence post-verbale (SPOV),

((Max)(a donné)(ce livre à Marie avec plaisir)).
PS SPRV FV SPOV

avec une relation d'indépendance entre la séquence pré-verbale et la forme verbale, et une relation de dépendance entre la forme verbale et la séquence post-verbale. La catégorie "forme verbale" ne s'analyse pas. Les séquences pré- et post-verbales s'analysent en " compléments de phrase" (CP) et "suite de compléments de verbe" (CV*) avec des relations d'indépendance entre ces catégories (pour la distinction "compléments de verbe" "compléments de phrase", voir /4/, /5/) :

((ce livre à Marie)(avec plaisir)).
SPOV CV* CP

Une suite de compléments de verbe s'analyse en compléments de verbe avec des relations d'indépendance entre ces compléments :

((ce livre)(à Marie)).
CV* CV CV

Un complément de verbe peut être constitué d'une complétive ou d'une infinitive réduite d'une complétive ; si tel est le cas, il s'analyse comme une phrase syntaxique :

Luc souhaite ((que Marie parte au Sud)).
CV PS
Luc souhaite ((partir au Sud)).
CV PS

Sinon, l'analyse des compléments de verbe est identique à celle des compléments de phrase : elle dépend des modifieurs qui apparaissent :

- les relatives : les relatives restrictives sont séparées de leur antécédent par une relation de dépendance, les relatives explicatives par une relation d'indépendance ;
- les adjectifs ou participes passés : ceux qui ne sont pas porteurs de compléments ne sont pas

séparés du nom qu'ils modifient ; ceux qui sont porteurs de compléments sont considérés comme des réductions de relative par effacement de "qui être" ; on effectue donc une analyse analogue à celle décrite pour les relatives.

- les compléments prépositionnels : grosso modo, les compléments prépositionnels situés à droite du nom tete du CN sont séparés du nom tête par une relation de dépendance, les autres par des relations d'indépendance.

CONCLUSION

Ces travaux ont conduit à l'élaboration d'un prototype en cours de réalisation et d'implantation sur un micro -calculateur équipé d'une carte de synthèse ; ce prototype concerne de courts récits d'attentats conçus dans un style journalistique ; les informations relatives aux attentats peuvent être entrées en machine à partir d'un programme de menu qui pose des questions à l'usager et construit automatiquement une représentation sémantique à partir de ses réponses. Le jeu de marqueurs prosodiques qui a été utilisé est adapté à un style d'élocution que l'on peut qualifier de "neutre" ; la question se pose de savoir si la méthode utilisée s'applique à d'autres styles d'élocution ; pour cela, l'étude actuelle s'élargit au positionnement automatique de marqueurs adaptés à la production de messages publicitaires /6/ ; il semble déjà que la structure syntaxico-prosodique et les relations de dépendance et d'indépendance dégagées pour une élocution "neutre" soient aussi utilisables pour cet autre jeu de marqueurs prosodiques. Si ce fait se confirme, il indiquerait qu'on peut engendrer des messages oraux simulant différents styles de voix à partir d'une même structuration syntaxico-prosodique.

BIBLIOGRAPHIE

/1/ DANLOS, L., Génération automatique de textes en langues naturelles ; Editions Masson, Paris, 1985.

/2/ COURBON, J.L, et EMERARD, F., Sparte : A text-to-speech machine using synthesis by diphones, IEEE-ICASSP Paris, vol.3, 1597-1600, 1982.

/3/ EMERARD, F., Synthèse par diphones et traitement de la prosodie, Thèse de 3ème cycle ; Université de Grenoble III.

/4/ BOONS, J.P., GUILLET, A., LECLERE, C.H., La structure des phrases simples en français. Constructions intransitives ; Editions Droz, Genève, 1976.

/5/ GROSS, M., Grammaire transformationnelle du français : syntaxe de l'adverbe ; Editions Hachette, Paris, 1986 (à paraître).

/6/ SORIN, C., STELLA, M., AGGOUN, A., Règles prosodiques et synthèse de la parole ; Symposium Franco-Soviétique sur le dialogue Homme-Machine, Pouchino, 1984.

VALIDATION DES INDICES ACOUSTIQUES DANS LE PROJET A.R.I.A.L II (*) PAR LA SYNTHESE

M.H SERRA - J. CAELEN

CERFIA - TAPT - UA 824 DU CNRS
118, route de Narbonne, 31062 TOULOUSE CEDEX

ABSTRACT :

In this paper we present an analysis/synthesis method for the validation of a set of acoustic cues which are used at the decoding level in the ARIAL Project. The problem is to control a speech-synthesizer with five spectral parameters which are issued from a reduced-redundancy coding of the spectral envelope. The model of production is a channel vocoder in which the filters width is in accordance with the speaker average fundamental frequency. We choose a linear programming method in order to restore the spectral redundancy and to compute the filters gains. The control parameters are adjusted by supplying the linear system with spectral continuity constraints based on flag-marks placed on the acoustical signal.

(*) Analyse et Reconnaissance des Informations Acoustiques et Linguistiques.

1. INTRODUCTION

Il semble admis aujourd'hui que les taux de reconnaissance de la parole continue multilocuteur dépendent largement de la qualité du décodage acoustique/phonétique.

Dans le projet ARIAL II, les relations existant entre le signal acoustique continu et les unités discrètes du langage sont fondées sur les traits classificatoires définis à partir des indices, qui sont eux-mêmes fonction des paramètres.

Il parait donc essentiel de valider les informations retenues au niveau de l'analyse paramétrique et indicielle avant toute reconnaissance. Une approche possible est de s'aider de la synthèse : c'est ce que tente de décrire cet article.

Après avoir placé notre étude dans un contexte plus général nous posons les principes de notre méthode. Nous décrivons dans une première partie la simulation d'un vocodeur à canaux numérique et dans une seconde partie la commande de ce vocodeur par cinq indices.

2. VALIDATION DES PARAMETRES CARACTERISTIQUES DE LA PAROLE

Les indices acoustiques doivent répondre à différentes exigences pour l'entrée et la sortie des informations parlées. En reconnaissance ils doivent participer aux 3 fonctions suivantes :

- segmentation du signal en unités pseudo-phonétiques.
- réduction de la variabilité inter et intralocuteur et corrélativement de la réduction de la redondance acoustique.
- classification et/ou discrimination des sons.

En synthèse il suffit généralement de reproduire le message d'un locuteur standard à partir d'un système d'indices Par conséquent le problème du filtrage des caractéristiques individuelles au niveau phonétique et linguistique (règles de production) ne se pose pas en termes aussi aigus.

Il convient donc de différencier et de valider de façon spécifique les informations sélectionnées pour l'encodage et le décodage de la parole.

- Pour l'encodage on a traditionnellement recours à l'analyse/synthèse. Depuis les travaux de Delattre [1] sur le Pattern Play Back, cette méthode est apparue comme un outil efficace pour la recherche et la vérification des paramètres significatifs du message parlé [2] [3] [4]. Le principal inconvénient de l'analyse/synthèse pour l'évaluation de la pertinence des indices, vient du fait qu'elle se fonde sur une base perceptuelle, entièrement subjective.
- Les analyses de données sont généralement appliquées aux informations utilisées pour le décodage. Elles permettent de s'affranchir de la dépendance et de la redondance des paramètres et permettent de dégager leurs rôles hiérarchiques dans la classification des données.

Ces deux types de validation sont complémentaires. En effet contrairement à l'analyse/synthèse les analyses de données sont purement objectives. Mais en contrepartie elles négligent certains facteurs perceptuellement importants, comme par exemple la redondance des informations et l'intégration asynchrone des indices dans le temps.

Un de nos objectifs est de compléter par l'analyse/synthèse les résultats des analyses statistiques (analyse discriminante et analyse en composantes principales) réalisées au CERFIA sur les indices utilisés dans le module de décodage acoustique/phonétique [5].

3. POSITION DU PROBLEME

3.1 Les indices acoustiques et phonétiques dans A.R.I.A.L. II

Les premières étapes de l'analyse du signal vocal dans le module de décodage sont effectuées par 2 fonctions séparées :

- mesure et codage de l'enveloppe spectrale (cadence d'échantillonnage de 8 ms).
- codage du signal d'excitation (détection et mesure du fondamental) [6].

L'enveloppe spectrale est codée dans un premier temps sur les 24 canaux d'un modèle d'oreille, puis projetée sur 6 indices. Les effets de distorsion non linéaire des processus auditifs périphériques sont rendus par le couplage entre les canaux. Il en résulte un lissage et une squelettisation de la courbe des énergies (fig. 2)

Les 5 premiers indices sont calculés à partir de combinaisons linéaires des énergies des 24 canaux. Ils utilisent la

quantité d'énergie dans cinq bandes particulières du spectre (bande du fondamental, des 3 premiers formants et du formant de bruit)[7]. Ce calcul fait intervenir 11 sous-bandes. L'énergie dans ces sous-bandes est notée (f1, f2) , f1 et f2 étant les bornes fréquentielles exprimées en Hz. Soit N(i) et N(i+1) les numéros des canaux correspondant à ces bornes avec $1 \leq i \leq 22$
L'énergie dans une bande, notée W(N(i)), est normalisée par rapport au nombre de canaux pris en compte.

Aigu/Grave : AG = (180, 380) - (3550, 6400)
Fermé/Ouvert : FO = (650, 760) - (260, 320)
Bémolisé/Diésé : BD = (2350,3100) - (1000, 1340)
Ecarté/Compact : EC = (1340, 2350) + (320, 650) - (880, 1000)
Doux/Strident : DS = (4000, 5000) - (2060, 2700)
avec : (180, 380) = (W(N(1))+..+W(N(2)))/(N(2) - N(1)+1)
(3550, 6400) = (W(N(3))+..+W(N(4)))/(N(4) - N(3)+1) etc..

Le sixième indice correspond à la dérivée spectrale à énergie moyenne constante.

Les indices, définis sur le plan acoustique, sont en relation avec les paramètres articulatoires et perceptifs : la référence à une représentation fréquentielle auditive permet de lisser les informations de façon pertinente. Les indices intègrent les variations des paramètres formantiques caractéristiques de l'appareil phonatoire [7]. Ces 2 aspects font que la variabilité des indices est moins grande que celle des formants. Par rapport aux canaux les 5 indices (AG, FO, BD, EC, DS) ont une redondance moindre. L'intérêt du codage vectoriel sur une telle base est de diminuer le débit d'information en appliquant des règles déduites des propriétés acoustiques et phonétiques de la parole. Ceci s'oppose aux méthodes purement mathématiques employées dans le vocodeur à faible redondance [8].

3.2 Synthèse par indices : principes de la méthode

La reproduction du signal à partir des indices consiste à effectuer l'inverse des opérations de l'analyse : le signal source excite un banc de filtres controlés par un ensemble de paramètres (gains, largeurs de bande, fréquences centrales). Ces paramètres ne sont pas donnés explicitement par les indices. Il est donc nécessaire d'établir une transformation mathématique permettant de les générer. D'autre part les effets non linéaires de l'analyse doivent être réintroduits soit au niveau de cette transformation, soit par le couplage des filtres du synthétiseur.

Afin de simplifier le problème nous considérons la transformation linéaire et les filtres de synthèse indépendants. L'étage d'analyse doit être modifié en conséquence : l'enveloppe spectrale est déterminée par un banc de filtres non couplés. Le codage sur les énergies spectrales (ou énergies dans les canaux d'analyse) dans R(Nc) est projeté sur le sous-espace des 6 indices dans R(6) (fig.1) selon le calcul précédent (3.1). Les valeurs des indices présentent plus de dispersion en raison de l'absence de couplage des filtres. D'autre part les différences entre les deux analyses seront accusées si l'échelonnement et la résolution fréquentielle des filtres ne sont pas équivalents. Le nombre de filtres du banc non couplé n'influe pas sur la valeur des indices à cause de la normalisation (3.1). Les évolutions des indices dans le temps sont comparables. Par conséquent le passage de l'analyse par modèle d'oreille à l'analyse par canaux pose le problème de l'équivalence des indices du point de vue de leur performance pour l'encodage et le décodage.

Le synthétiseur, dual de l'analyseur, contient le même banc de filtre (fig.1) : les largeurs des canaux et leurs fréquences centrales sont fixées pour l'analyse. Le vecteur de commande est composé des gains des filtres $(E'_1,...,E'_{Nc})$ et calculé par la transformation indices → paramètres de commande (énergies spectrales) dont le rôle est de régénérer une certaine redondance spectrale éliminée lors du codage dans R(6).

L'ensemble analyseur/synthétiseur (fig.1) est appelé "vocodeur à indices". Il peut être classé dans la famille des vocodeurs à canaux à faible redondance [8] [9]. La validation des indices par la synthèse consiste à trouver une transformation linéaire telle que les signaux synthétisés à partir des 2 représentations vectorielles $(E_1,...,E_{Nc})$ et (AG, FO, BD, EC, DS) soient perceptivement identiques. Dans le paragraphe 4 nous présentons les caractéristiques principales du banc de filtres numériques d'ou sont issus les deux formes de codage $(E_1,...E_{Nc})$ et (AG, FO, BD, EC, DS). Le paragraphe 5 décrit la recherche d'une transformation linéaire qui minimise la perte d'information produite par la réduction de l'enveloppe spectrale sur les 5 indices.

4. SIMULATION D'UN VOCODEUR A CANAUX NUMERIQUE

La simulation est réalisée sur un PDP 11 -73 (DEC). Les filtres de Butterworth d'ordre 4 ont été choisis car ils offrent un bon compromis entre une réponse fréquentielle assez sélective et une réponse temporelle donnant des lobes secondaires peu nombreux et d'amplitude faible [10] [11] ; ceci afin d'éviter des effets de réverbération sur le signal synthétique.
Pour l'analyse il nous a semblé important d'adapter la largeur des filtres au fondamental moyen du locuteur. Cette précaution réduit les distorsions lors de la mesure de l'amplitude des harmoniques. Le logiciel permet de redéfinir un banc approprié à chaque locuteur traité. Pour la synthèse, afin d'éviter des effets de réverbération dus à des temps de propagations inégaux, les filtres (de même ordre) doivent avoir une largeur de bande constante.
Les qualités d'un vocodeur de synthèse sont les suivantes [12] [13] [14] :
- réponse globale en amplitude plate (∓3 dB d'ondulation)
- phase globale continue (pas de sauts de phase aux frontières des bandes passantes) et quasi-linéarité.

Afin de réunir toutes ces conditions le banc de filtres est élaboré suivant la méthode de [12] étendue aux filtres numériques par l'emploi de la transformée bilinéaire. Le principe de construction n'autorise qu'une répartition linéaire des fréquences centrales des canaux. La détection du fondamental est effectuée par le même algorithme [6].

5. COMMANDE DU VOCODEUR A INDICES

5.1 Résolution numérique du système

Soit Nc le nombre de canaux du banc de filtres. La commande du synthétiseur nécessite la donnée de l'enveloppe spectrale codée sur le vecteur $(E'_1,...,E'_{Nc})$ déterminé à partir des 5 indices (AG, FO,..). L'indice continu/discontinu (CD) n'est pas pris en compte dans cette étape.

Le système linéaire à résoudre peut s'écrire sous la forme:

$$\begin{bmatrix} AG \\ FO \\ BD \\ EC \\ DS \end{bmatrix} = \begin{bmatrix} A_{11} & & A_{1Nc} \\ & & \\ & & \\ & & \\ A_{51} & & A_{5Nc} \end{bmatrix} X \begin{bmatrix} E'_1 \\ \\ \\ \\ E'_{Nc} \end{bmatrix}$$

Le nombre de solutions est indéterminé, d'ordre Nc-5. Afin de limiter de façon simple, le choix des vecteurs possibles, nous ajoutons au système un critère linéaire pour sélectionner les solutions proches du spectre incident $(E_1,...,E_{Nc})$. Pour celà nous supposons que le sous-ensemble des vecteurs solutions est tel que la différence entre la somme de ces coordonnées E'i et l'énergie totale détectée à l'analyse est minimale.

La recherche de ces solutions est effectuée par une méthode d'optimisation appliquée aux programmes linéaires : le simplexe. L'algorithme du simplexe [15] fournit une solution extrémale du programme linéaire. Ces solutions extrémales sont les vecteurs de R(Nc) correspondant aux sommets du polyèdre formé par la

réunion des contraintes. Cette technique de résolution conduit à une exploration rapide des solutions. Cependant elle comporte certains inconvénients :
- l'opérateur ne peut accéder à l'ensemble des solutions.
- l'algorithme a tendance à annuler des variables (E'i) ; par conséquent la forme de la courbe spectrale solution est très accidentée. Or les solutions attendues sont lissées et sans zéro dans le spectre. Le lissage serait une autre forme de contrainte non linéaire, qu'il est difficile d'intégrer dans le cadre présent.
- Les conditions optimales d'utilisation du simplexe le restreignent aux systèmes de faible degré d'incertitude, ce qui n'est pas le cas ici.

Ces constatations nous ont amenés à adjoindre au système initial des contraintes devant faire converger les solutions de 2 façons : en se rapprochant globalement du spectre incident et en maitenant des écarts d'amplitude faible entre les coordonnées.

Soit $(W_1,\ldots,W_{Nc})$ un échantillon de l'analyse spectrale proche de la solution recherchée et appelé spectre de référence. Nous supposons que les coordonnées du vecteur optimal E'i sont déduites des énergies spectrales de référence par des opérations linéaires :
Les contraintes sont formalisées comme suit :
$Wi - Sv \leqslant E'i \leqslant Wi + Sv$ pour $1 \leqslant i \leqslant Nc$ $(Wi - Sv, Wi + Sv)$ est un intervalle de variation limitant la divergence des coordonnées E'i. Le problème est ramené ici à la détermination du spectre de référence et du seuil de variation spectrale Sv.

5.2 Premiers résultats

Le calcul de la suite des enveloppes spectrales est initialisé sur un échantillon spectral pris arbitrairement dans le mot analysé ; le degré de variation (Sv) est d'abord fixé à une valeur faible (3dB). En cas d'échec on réitère le processus de recherche de la solution en tolérant un seuil Sv de plus en plus grand (5,7,9,11 dB...) jusqu'à détection d'un vecteur optimal. Au fur et à mesure du traitement dans le temps les spectres obtenus s'éloignent du spectre de référence et la parole reconstruite à partir de ces derniers est inintelligible.

Bien que nous ne cherchions pas absolument une superposition exacte des spectres incidents et des vecteurs du simplexe, il s'est avéré indispensable de réinitialiser la recherche des solutions chaque fois que le spectre d'analyse présente une grande instabilité. Ainsi les équations permettant d'accéder à des solutions plus satisfaisantes sont assimilées à des contraintes de continuité spectrale dans le temps.

Par conséquent l'ajustement de la transformation linéaire reliant les indices aux paramètres de commande s'appuie sur le relevé des instants présentant une forte discontinuité spectrale (blocs d'initialisation). Cette recherche est guidée par le souci de minimiser le nombre de spectres de référence fournis au système linéaire, afin de valider au mieux l'information contenue dans les indices.

5.3 Recherche des blocs d'initialisation

Les informations dont nous disposons sont présentées sur la figure 2.Afin d'éviter des erreurs et de conférer une certaine souplesse au logiciel d'analyse/synthèse la recherche des instants d'initialisation est opérée manuellement par la lecture et observation des informations de l'analyse (fig.2) Nous enregistrons des étiquettes notées 'IN' (initialisation) qui séparent les segments à l'intérieur desquels l'algorithme n'est pas réinitialisé. Ces segments stables ont une taille qui varie du phone à la syllabe.

Dans le paragraphe suivant nous illustrons ce principe d'étiquetage sur le mot /sɔmə/ codé sur la figure /SOME/.

6. EXEMPLE DE SYNTHESE PAR INDICES

Les marques positionées (fig. 2) ont été fixées après plusieurs essais, en gardant pour objectif la ressemblance des signaux synthétiques. La production du phonème initial /S/ donne 2 phases acoustiques :
Des échantillons 3 à 11, on note l'établissement de la friction caractérisé par l'instabilité spectrale intense représentée par l'indice continu/discontinu (CD) ; des échantillons 12 à 26, on distingue une zone stable avec bruit de friction. Les indices aigu, strident, et diésé sont les plus manifestes. La frontière entre /S/ et /O/ fait apparaître de grandes variations formantiques. Si l'on calcule la suite des spectres du /O/ par référence à un vecteur de la transition (échantillons 27 à 28) le /O/ synthétisé est très distordu. Le bloc d'initialisation est choisi à 29 à partir duquel le changement des paramètres est modéré. A l'intérieur de la syllabe /ME/ (syllabe fermée, /e/ élidable) l'évolution des paramètres spectraux est très progressive. L'information des 5 indices associée à la donnée du premier spectre (échantillon 42) suffit à reproduire un signal perceptuellement très proche de celui obtenu par les canaux. Le bruit de la phase finale de faible intensité est intégré à la syllabe lors du traitement par le simplexe. La figure 3 permet de comparer un spectre issu de l'analyse par vocodeur (24 canaux) et sa reconstruction à partir des indices.

7. CONCLUSION

La méthode choisie pour restaurer une information sur le spectre à partir des 5 indices introduit certaines limitations. Cependant, l'étiquetage soigné du signal permet d'adapter la transformation linéaire de façon à optimiser le résultat perceptif. Dans ce contexte, la perte d'information subie lors du codage de l'enveloppe spectrale sur les 5 indices est négligeable. Le traitement du signal vocal proposé ici rappelle les techniques mathématiques (transformation linéaire et critère des moindres carrés) dans le vocodeur à faible redondance. L'originalité de ce travail est de s'appuyer sur les propriétés acoustiques et phonétiques de la parole afin de générer une transformation satisfaisante. L'intégration du contexte (adaptation des filtres au locuteur et recherche des événements pertinents sur le spectre) sont les traits principaux que nous développerons afin d'orienter le logiciel actuel sur la recherche des invariants au travers des indices.

REFERENCES

[1] P. Delattre et Al "Acoustic loci and transitionnal cues for consonants" J. Acoust.Soc.Am. 24, pp 597-606, 1955.

[2] G. Murillo, P. Badin "Methodes et logiciels pour la synthèse de parole de haute qualité". XII JEP, GALF-CNRS, pp 106-115, Bruxelles, Mai 1984.

[3] R. Carre, "Contribution aux études sur l'analyse et la synthèse de la parole. Rôle et importance des formants". Thèse de docteur Es-Sciences physiques, Grenoble 1971.

[4] C. Bietry "Synthèse multilocuteur de haute qualité". XII JEP, GALF-CNRS, pp 79-88, Bruxelles, Mai 1984.

[5] G. Caelen, N. Vigouroux "Les indices de distribution spectrale : étude comparative au moyen de deux analyses discriminantes monolocuteur et interlocuteur". Speech Communication Vol 2, N°2-3, pp 133-136, juillet 1983.

[6] J. Caelen, P. Cazenave "Mesure du fondamental par filtrage variable", VIII JEP, première partie, pp 54-59, Aix en Provence, 1977.

[7] J. Caelen, G.Caelen "Indices et propriété dans le projet ARIAL II" Actes du séminaire Encodage et Décodage Phonétique" pp 129-143, Toulouse, septembre 1981
[8] H.P Kramer, M.V. Mathews "Linear coding for transmitting a set of correlate signals" I.R.E. Trans. Inform. Theory IT-2 pp41-46, 1956.
[9] J.L.Flanagan "Speech analysis, synthesis and perception" Ed.Springer, Verlag, Berlin, 1972
[10] L.Rabiner, R.Shafer "Digital processing of speech signals" Prentice Hall, New-Jersey, 1978
[11] L. Rabiner, B. Gold "Theory and applications on digital processing" Prentice Hall, New-Jersey, 1978
[12] R.M. Golden "Vocoder filter design : practical considerations", J.Acoust.Soc.of Am. Vol43,4,pp 803-810, 1968.
[13] F:Zurcher, "Le vocodeur à canaux : une nouvelle jeunesse?" Recherches/Ac. CNET, Vol 6, pp 21-40, 1979/1980
[14] B. Gold, C.Rader "The channel vocoder", IEEE Trans. on Audio and Elec., Vol AU-15, No4, décembre 1967
[15] M.Simonnard "Programmation linéaire" Ed Dunod, 1978.

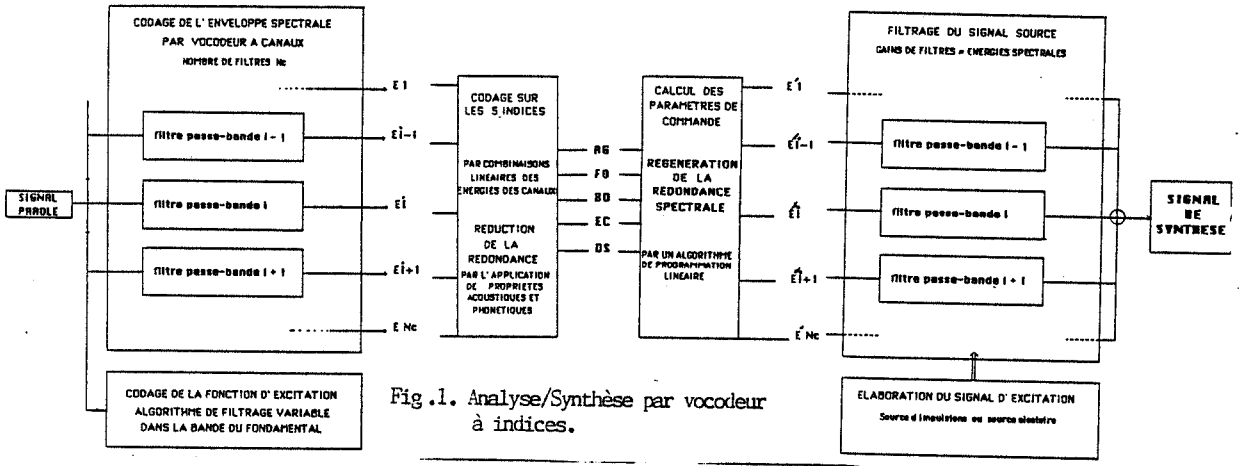
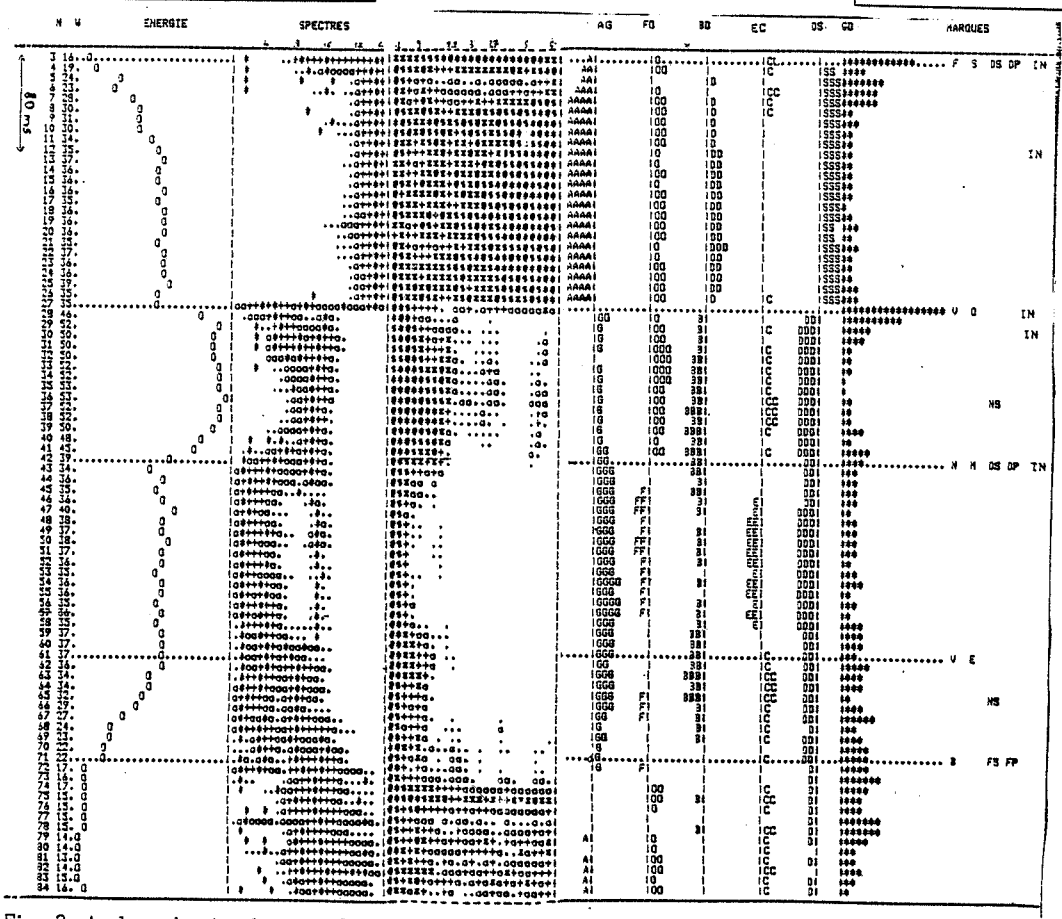


Fig .1. Analyse/Synthèse par vocodeur à indices.



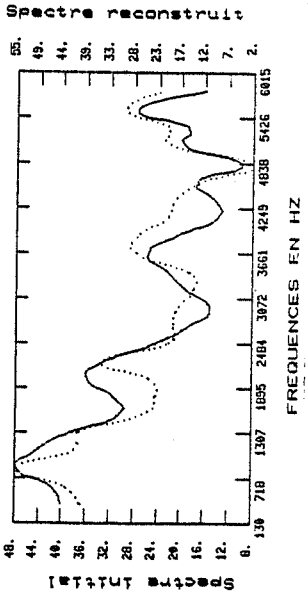


Fig .3. Echantillon spectral No 32. Spectre initial en ligne continue et spectre reconstruit en pointillés.

Fig .2. Analyse du signal vocal. De gauche à droite : numéro de l'échantillon, énergie en dB, spectres par modèle d'oreille et par vocodeur 24 canaux, indices spectraux, indices continu/discontinu, marques de segmentation (acoustique, phonémique, syllabique, prosodique), marque d'initialisation.

****
* SEGMENTATION *
****

INTRODUCTION A UNE SEGMENTATION CINEMATIQUE

J. Caelen

Laboratoire C.E.R.F.I.A. UA CNRS n° 824 Université P. Sabatier
118 Route de Narbonne 31062 Toulouse CEDEX

ABSTRACT

This paper describes a new approach for speech segmentation, using cinematic concepts. Instead of researching the discontinuity on the signal, we localize acoustic-targets on the spectral-path. An algorithm (segmentation automaton) is based on this principle.

1. INTRODUCTION

La plupart des méthodes de segmentation de la parole en unités phonétiques (phones, phonèmes, syllabes...) utilisent une fonction qui mesure les discontinùités locales du signal ou de son spectre, dans le temps. Ces discontinuités ne sont pas toutes synchrones selon le système de paramètres choisi et les critères retenus. On aboutit alors à des conceptions différentes selon que l'on considère ces discontinuités de manière analytique (événements) ou globales (segments). Or ces deux "manières" ne sont pas forcément en accord avec le geste articulatoire: on sait en effet que les mouvements des articulateurs ne sont pas tous synchrones, ils ont des inerties qui introduisent des retards (vs. anticipations) [1] dans l'établissement de certaines propriétés acoustiques. On sait d'autre part que ces mouvements, du fait d'une certaine compensation acoustique, ne concourent pas tous, avec un égal degré de pertinence, à la production du son. De plus, lorsque le débit d'élocution croit, ces gestes sont moins précis et la "cible" articulatoire est moins souvent atteinte: seule une inflexion (intention) est alors perceptible, pourtant suffisante à l'auditeur humain pour le décodage phonétique. Dans ce dernier cas, la structure "discontinue" de la parole s'atténue rendant plus incertaine la démarcation des frontières des segments.

Le problème de la segmentation de la parole doit donc se poser en termes de localisation des cibles acoustiques (ou articulatoires) et non en termes de recherche de discontinuités (frontières, variations...). Dans un espace de représentation adéquat, cette idée sous-tend celle de trajectoire (cinématique du point): il devient alors tentant de s'inspirer des méthodes utilisées en mécanique pour résoudre le problème.

2. TRAJECTOIRE DANS L'ESPACE DE REPRESENTATION

Dans tout système de traitement de la parole, le signal est analysé et/ou codé dans un espace de représentation, en général R^p où p est le nombre de paramètres choisis. (énergies spectrales, coeffs. cepstraux...)

L'analyse, répétée régulièrement sur une fenêtre glissante, produit à l'instant t_n, un vecteur $X_n=(x_{1n},x_{2n},\ldots,x_{pn})$ de R^p. Le point $M_n=M(X_n;t_n)$ décrit dans R^p une trajectoire discrète $T=(M_o,M_1,\ldots,M_n,\ldots)$ que l'on peut étudier à travers les concepts développés en cinématique du point matériel. Si la trajectoire est toute entière contenue dans un sous-espace R^s de R^p, le point M_n possède s degrés de liberté. Il est animé d'une vitesse v_n et d'une accélération γ_n et on peut supposer qu'il se dirige (ou aboutit) à tout instant vers une "cible" qui est une région déterminée de l'espace R^s. Si le geste articulatoire est "parfaitement" réalisé, on peut admettre que la cible acoustique est atteinte: le poit M_n s'y stabilise (trajectoire stationnaire) avant de se diriger à nouveau vers la prochaine cible. On peut admettre aussi que les transitions d'une cible vers la suivante se font par des trajectoires sans point singulier (stationnaire, rebroussement, double,...). Ces hypothèses se rapprochent de celles qui sont faites habituellement, à savoir que la parole est une succession de portions stables et instables (sur lesquelles s'appuie la recherche des discontinuités). Mais les phénomènes de coarticulation déforment cette trajectoire "idéale" (Fig. 1) qui ne peut plus se décomposer aussi simplement. Il devient alors nécessaire d'étudier plus précisément les trajectoires réellement observées.

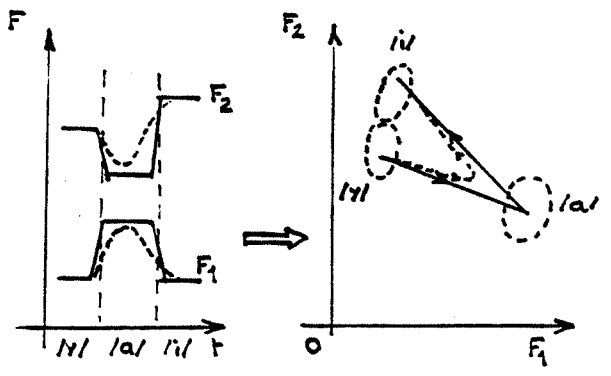


Fig. 1: Trajectoire formantique
—— idéale ---- observée

3. LE PROBLEME

Le problème posé par le décodage acoustico/phonétique se ramène à localiser les cibles décrites ci-dessus à partir de la trajectoire connue. Un sous-problème, celui de la segmentation, est de détecter seulement la présence des cibles sans les localiser précisément.

4. APPROCHE PAR LA MECANIQUE ANALYTIQUE

4.1. Grandeurs usuelles

Le choix d'une métrique dans R^p permet de définir les grandeurs v_n (vitesse) et γ_n (accélération) qui sont, en discret, des distances: $v_n=d(M_n,M_{n-1})/(t_n-t_{n-1})$

$\gamma_n=(v_n-v_{n-1})/(t_n-t_{n-1})$

avec en général $t_n-t_{n-1}=dt=$constante.
La courbure de la trajectoire au point M_n est l'inverse du rayon du cercle passant par les points M_{n-1},M_n,M_{n+1} dont le centre est dans le plan des vecteurs v_n et v_{n+1}. La torsion au point M_n est l'angle formé par les plans définis par les vecteurs (v_nv_{n-1}) et (v_nv_{n+1}). On rejoint ici les concepts de corrélation locale (torsion nulle) de trajectoire rectiligne (courbure nulle) de trajectoire stationnaire (courbure indéterminée) etc... Il est évident que les méthodes de l'analyse des données, s'appuyant sur les mêmes concepts, seront tout à fait appropriées dans ce domaine: R^s n'est que le sous-espace des s premiers facteurs en ACP (Analyse en Composantes Principales) par exemple.

En associant une masse au point M_n on peut tenter de décrire la trajectoire au moyen de forces s'exercant sur ce point. On fait l'hypothèse que les forces de liaison sont sans frottement et que le point matériel M_n reste soumis aux lois générales de la dynamique: $m.\gamma_n$ est donc la résultante des forces d'attraction des cibles sur le point M_n de masse m.

4.2. Modélisation

Hypothèses de travail:

(a) le point matériel ne se dirige que vers une seule cible à la fois,

(b) la cible est atteinte lorsque la trajectoire est stationnaire (ou quasi-sationnaire)

(c) les forces de laison sont sans frottement

(d) les forces appliquées sont des forces d'attraction dues à la présence rapprochée de cibles,

(e) la vitesse moyenne est proportionnelle au débit d'élocution,

(f) les forces ont une intensité décroissante avec le débit,

(g) un changement de direction dénote la présence d'une cible non atteinte.

4.3. Expérimentation

Dans une première étude,les paramètres retenus pour caractériser l'espace de représentation sont les indices [2] AG aigu-grave, FO fermé-ouvert, BD bémolisé-diésé, EC écarté-compact, DS doux-strident calculés toutes les 8 ms sur les sorties du modèle d'oreille comportant 24 canaux. Ces 5 indices sont donc les coordonnées instantanées du point M_n considéré ci-dessus. Sur un corpus de phrases et mots isolés prononcé par 6 locuteurs il est apparu que: (Fig. 2 et 3)

(a) les indices sont localement corrélés, ce qui met en évidence l'existence de forces de liaison locales (et/ou de contraintes),

(b) la courbure de la trajectoire est faible dans les phases de transition entre cibles (trajectoire quasi-linéaire),

(c) la trajectoire est quasi-stationnaire quand une cible est atteinte. Un mouvement desordonné s'installe autour du centre de la cible,

(d) une cible non atteinte mais présente correspond à un changement de direction net du point matériel,

(e) un débit d'élocution important "rapproche" les cibles

(f) le point matériel accélère en quittant une cible et décélère en s'approchant de la suivante,

(g) la coarticulation propre à chaque locuteur se déduit de la position relative des cibles.

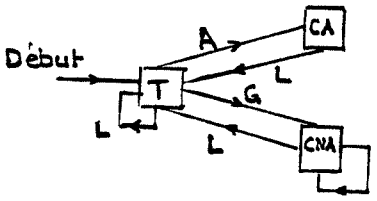
4.4. Un automate de segmentation

A partir des observations (a-g) on peut diviser les trajectoires en trois catégories, suffisantes en vue de la segmentation:

1. trajectoire quasi-linéaire: transition entre deux cibles,
2. trajectoire aléatoire: cible atteinte,
3. trajectoire "anguleuse": cible présente, non atteinte, débit d'élocution localement important.

Ces trois catégories définissent les trois états d'un automate de segmentation:

T: transition, CA cible atteinte, CNA cible non atteinte. Les transitions sont données par trois procédures de calcul sur la trajectoire: L trajectoire quasi-linéaire, A trajectoire aléatoire, G trajectoire anguleuse. Ceci conduit à l'automate suivant:



5. CONCLUSION

Ces premiers résultats permettent d'envisager l'étude de la parole sous un angle cinématique voire dynamique. Certes il est nécessaire de poser et de vérifier un modèle acoustico-articulatoire pour quantifier et préciser les grandeurs utiles (forces, masses...) mais d'ores et déjà cette approche semble riche de prolongements pour décrire une syntaxe des "déformations" locales en vue du décodage phonétique. Pour la segmen-

tation le problème parait plus simple du fait de la netteté des trajectoires. Ceci reste à confirmer sur de grands corpus et de nombreux locuteurs.

REFERENCES

[1] J.F.P. Bonnot, Variations d'encodage et oppositions de quantité: une étude contrastive. Actes du séminaire "Processus d'encodage et de décodage phonétiques", GALF-CNRS pp. 42 - 53, 1981.

[2] J. Caelen, G. Caelen-Haumont, Indices et propriétés dans le projet ARIAL II, Actes du séminaire "Processus d'encodage et de décodage phonétiques", GALF-CNRS, pp. 129-143, 1981.

[3] G. Fant, Speech Sounds and Features. MIT Press, Cambridge, 1973.

[4] K.S. Fu, Syntactic Pattern Recognition, Springer Verlag, 1982.

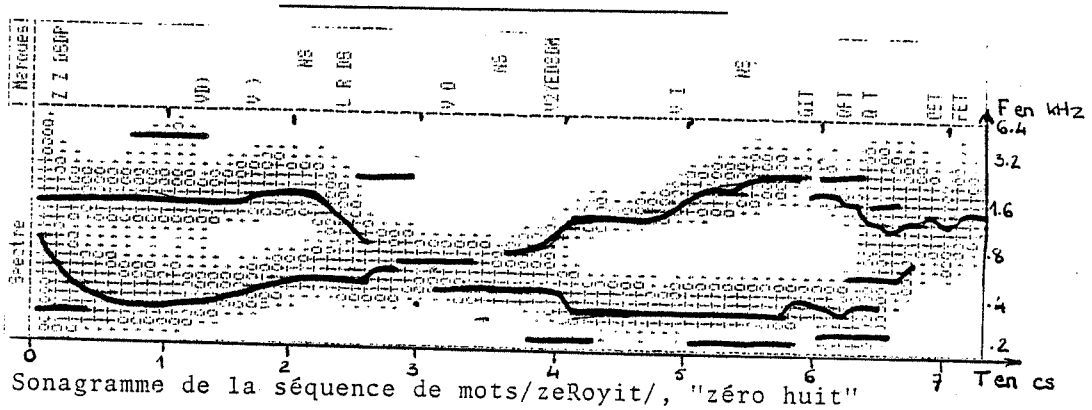


Fig. 2: Sonagramme de la séquence de mots/zeRoyit/, "zéro huit"

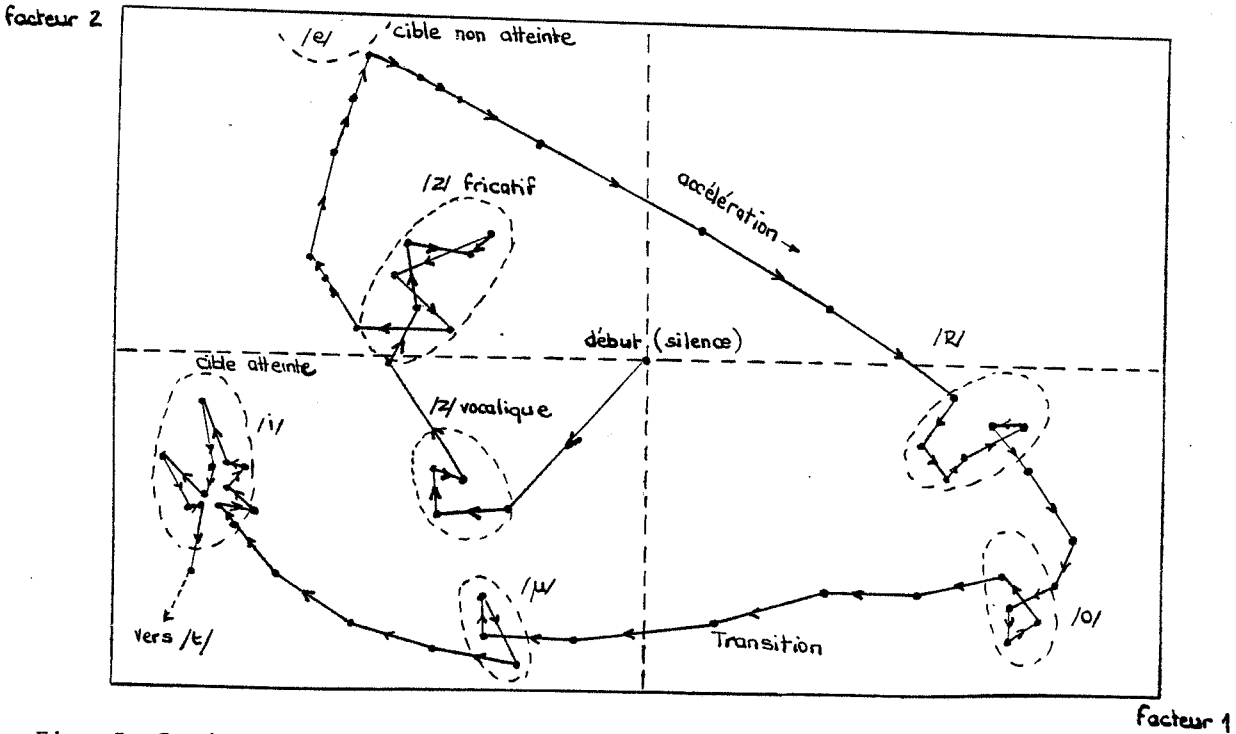


Fig. 3: Projection de la trajectoire des indices sur le plan factoriel formé par les 2 premiers facteurs (début arbitraire). Les cibles sont matérialisées par des ellipses, les échantillons spectraux sont localisés par des points.

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

C. Abry, C. Benoît, L.J. Boë, R. Sock

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

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

SEGMENTATION, REPRESENTATION ET PRODUCTION

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

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

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

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

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

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

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

EVENEMENTS & COORDINATIONS vs. DISCONTINUITES ET SEGMENTS

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

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

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

---

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

** Et à différentes échelles...

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

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

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

Ainsi, alors qu'un segment minimal n'est qu'un contenu temporel (durée positive ou nulle) entre deux discontinuités successives, l'analyse de celles-ci en événements offre une réelle gamme de relations coordinatives (binaires ou n-aires) entre ceux de ces événements qui peuvent refléter les coordinations linguistiques apprises par le sujet parlant. On peut rappeller certaines de celles-ci comme la tenue des consonnes, le VOT (6), le VTT (Voice Termination Time (1), pour nous relation C-VT, soit voiced closure ou closure voicing in (4)), etc., sachant qu'il en reste plus d'une à construire ou à reformuler.

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

EVENEMENTS ACOUSTIQUES, PHONETIQUES ET ARTICULATOIRES

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

début de friction) seront bien entendu étiquetés CFO (Consonantal Frication Onset : début consonantique fricatif).

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

LECTURE ARTICULATOIRE

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

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

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

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

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

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

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

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

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

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

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

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

. R0, pour les voyelles orales ;

. R1, pour les consonnes nasales et latérales ;

. R2, pour les approximantes.

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

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

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

- en initiale d'énoncé :

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

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

- en finale :

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

PERSPECTIVES EN AVAL ET EN AMONT

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

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

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

3) Construire les classes phonétiques correspondant à ces coordinations.

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

1) Sursegmenter les segments entre nos événements *.

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

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

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

---

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

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

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

REMERCIEMENTS ET REDEVANCES

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

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

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

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

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

REFERENCES

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

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

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

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

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

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

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

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

R. Sock, C. Benoît & C. Abry

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

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

INTRODUCTION

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

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

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

CORPUS ET PARAMETRES RETENUS

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

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

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

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

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

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

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

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

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

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

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

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

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

RESULTATS ET DISCUSSION

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

CONCLUSION

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

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

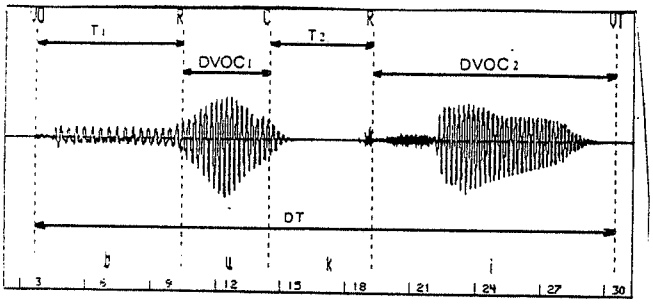


Figure 1

Logatome 'buki' par M.

DT : Durée totale du logatome

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

* Evénéments phonétiques :
  VO : Début de voisement
  R : Relâchement de degré 2
  C : Clusion de degré 4
  VT : fin de voisement

Corrél. :	DVOC1 (--) DT1		DVOC1 (--) DT2		DVOC2 (-) DT2	
Locuteur	(baki)	(buki)	(baki)	(buki)	(baki)	(buki)
B	**********	*******	*******	********	**********	**********
D	*******	**********	*******	*******	**********	**********
M	**********	**********			*******	*****
S	******				**********	**********

Tableau A

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

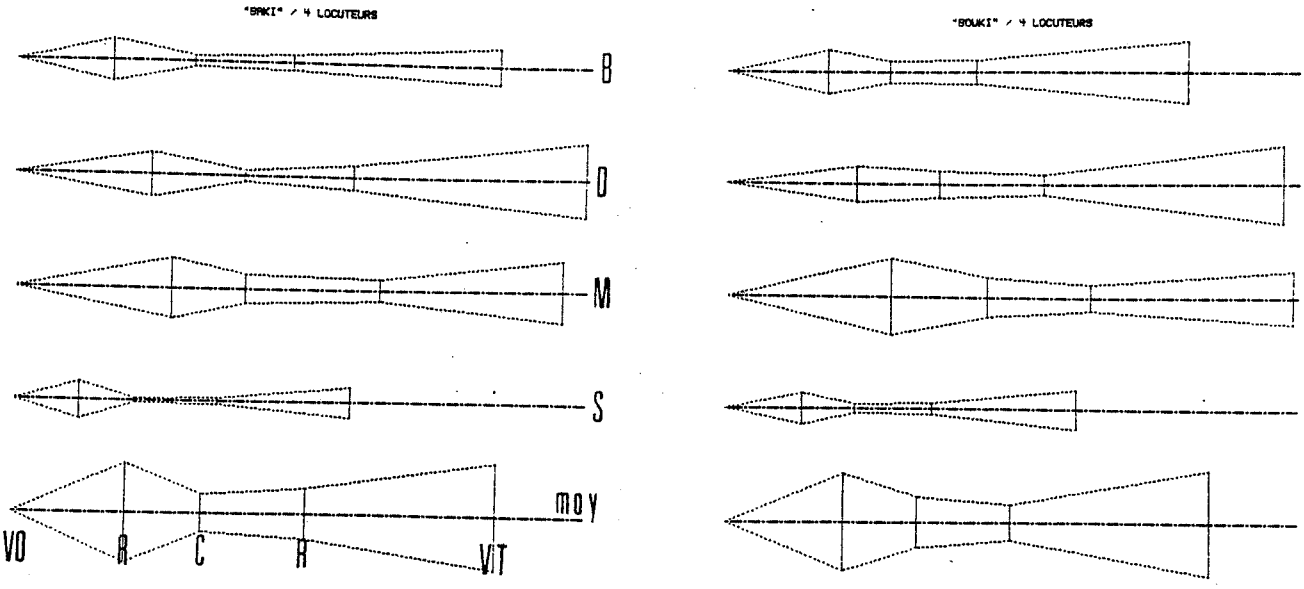


Figure 2

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

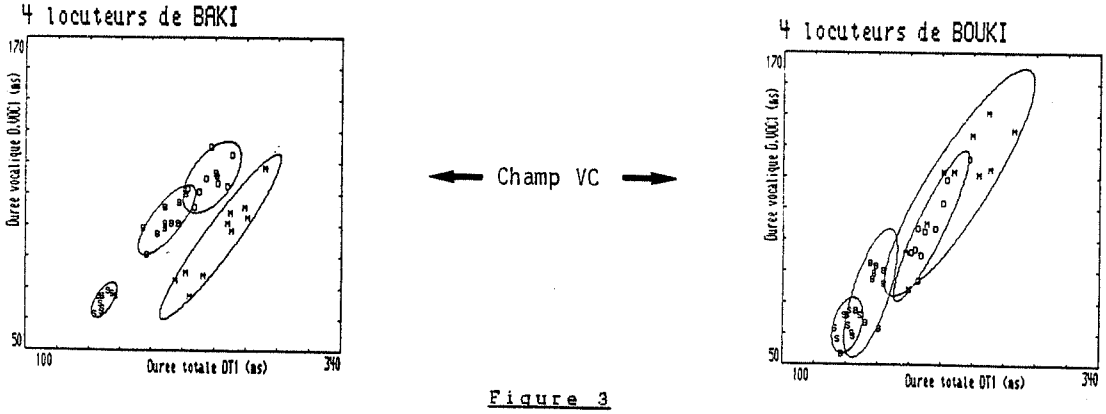
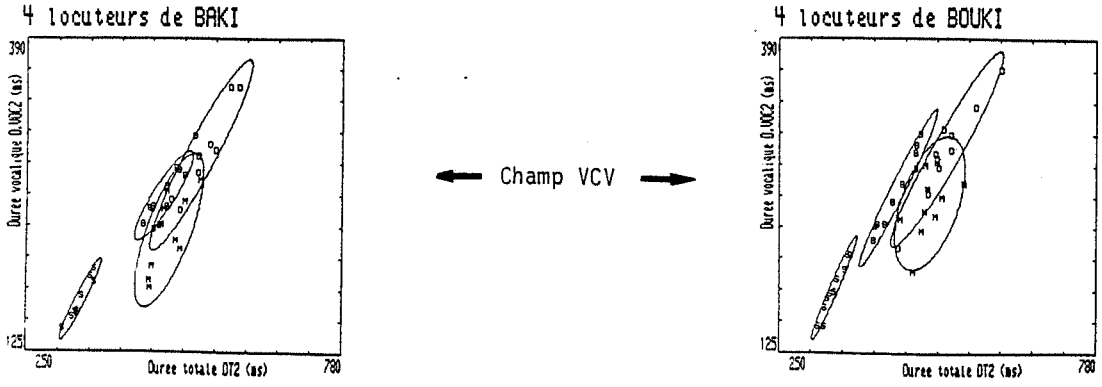


Figure 3

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



SEGMENTATION EN EVENEMENTS PHONETIQUES ET EN UNITES SYLLABIQUES

G. Pérennou M. de Calmès

CERFIA - TAPT - UA 824 DU CNRS
118, Route de Narbonne, 31062 TOULOUSE-CEDEX

ABSTRACT

In this paper we describe a phonetic and phonological decoding model in which:
- the homogeneuous phones resulting from the phonetic analysis are regrouped in phonetic events,
- the phonetics events are regrouped in syllabic units ("semi syllabes"),
- scores are attributed to the abstract units at the request of the lexical component.

This paper only deals with the first two points.

The first part is a description of the regrouping mecanisms using the **TEX** system (**EX**pert **T**ransducer). TEX has something in common with an expert system. The knowledge necessary to carry out tasks such as decoding tasks is given in a declarative form, namely, a definition of cues and features as well as a definition of production rule that control the regrouping in higher-level units.

In the second part, we give the results obtained after analysing a corpus of continuous speech.

INTRODUCTION

L'un des problèmes cruciaux du traitement automatique de la parole est actuellement le décodage phonético-phonologique où se manifeste un déficit de performances par rapport à l'auditeur humain interdisant véritablement la réalisation de systèmes quelque peu ambitieux en compréhension automatique de la parole.

Dans le projet ARIAL II ([1],[2],[3]), c'est le point sur lequel nous faisons porter notre effort ([4],[5],[6]). Il est vraissemblable que les systèmes de décodage phonético-phonologique élaborés comporteront plusieurs stratégies complémentaires et c'est ce que nous envisageons dans ARIAL II. L'approche que nous décrivons ici comporte:

(i) un traitement acoustico-phonétique segmentant le signal vocal en phones homogènes, chacun décrit par un vecteur d'indices divers: aigu-grave, doux-strident...(sur ce point on se reportera à Caelen [5], Vigouroux [6]); nous considérons que les phones reflètent les situations articulatoires successives du conduit vocal.

(ii) une représentation en événements phonétiques; ceux-ci reflètent les gestes articulatoires successifs, et par conséquent les changements de cibles qui les sous-tendent.

(iii) une représentation en unités syllabiques.

Il ne sera pas question ici d'accès lexical. Mais nous pouvons indiquer que nous l'envisageons à partir des trois niveaux précédents, de manière ascendante ou descendante. Dans cette communication nous nous intéresserons aux points (ii) et (iii) vus plus précisément sous l'angle de la segmentation.

Les événements phonétiques

Le point de vue schématique exprimé précédemment, selon lequel une succession de cibles articulatoires déterminerait une suite de gestes articulatoires, discernables dans la succession des phones par les changements de tendance, doit être probablement nuancé ou corrigé sur divers points:

1- Les événements successifs $y_1\ y_2 \ldots y_n$ dégagés par des traitements envisagés reflètent des commandes sous-jacentes $x_1\ x_2 \ldots x_m$ pouvant être déduites de la représentation phonologique de surface; dans cette formulation le décodage doit prendre en compte différentes possibilités de fautes: insertions, fusions (délétions), substitutions résultant tout à la fois du contenu spectral, des limites de notre description acoustico-phonétique et des critères retenus pour la détermination des événements.

2- Entre autres faits de nature phonétique qui s'opposent à une vision par trop simplifiée de la correspondance des x_i aux y_j mentionnons que:

- certains faits, pertinents au plan phonologique, pouvant être réalisés en tant que segments, ne sont en fait observables sur le spectre que comme simple modulation d'un événement; c'est le cas par exemple pour des explosions non marquées, surtout si elles ne demandent pas de grande modification de la configuration articulatoire (ex.: [di] si d n'est pas aspiré); certaines réalisations de semi-voyelles ou de voyelles fermées atones se présentent aussi de cette manière.

- d'autres faits, non pertinents au plan phonologique, se manifestent comme des événements très marqués: ainsi, dans le cas d' une voyelle tenue des modulations peuvent apparaître en liaison avec divers traits prosodiques.

Les unités syllabiques

Le niveau syllabique peut être utile pour l'interprétation des événements en phonèmes. Dans l'approche présentée ici ils sont regroupés en unités syllabiques [3] qui sont:

- des syllabes croissantes ouvertes (SCO),
- des fermetures syllabiques (FS),
- des éléments pré ou post syllabiques.

Ainsi [#stridãt#] pourra être observé comme:

$$[[_{SYLL}\ (_{PRESYLL}{}^{s})\ (_{SCO}{}^{tri})]\ [_{SYLL}\ (_{SCO}{}^{dã})\ (_{FS}{}^{\tilde{}})(_{POSTSYLL}{}^{t})]]$$

où ~ représente, pour certain type de réalisation, une consonne nasale prolongeant la voyelle nasale.

Les unités syllabiques sont dégagées d'après les alternances de croissance et de décroissance sur lesquelles nous reviendrons plus loin.

TRANSDUCTEURS GENERALISES - LE SYSTEME TEX

Un transducteur généralisé ([4],[7]) peut être vu comme une machine avec un nombre fini d'états, un ruban d'entrée et de sortie. Cette machine peut lire sur le ruban, déplacer la tête de lecture, écrire sur le ruban de sortie. Pour en préciser le fonctionnement, nous devons le doter:

- d'une fonction de transition qui définira les changements d'états,
- d'un opérateur qui écrit sur le ruban de sortie,
- d'un opérateur qui gère les déplacements de la tête de lecture.

Un transducteur généralisé est constitué de sextuplets de la forme:

(e, ALT, e', COND, ACTION, SUI)

avec e: état origine
e': état but
ALT: stratégie locale:
m: arc ayant la meilleure plausibilité,
f: tous les arcs ayant une plausibilité suffisante (recherche en faisceau),
etc...
COND: une constante (comme VRAI) ou un prédicat simple ou complexe ou la clause SINON ou un nom de réseau.
ACTION: une ou plusieurs actions consistant par exemple à placer des valeurs dans un registre ou à écrire sur le ruban de sortie,...
SUI: examen des arcs sortants du noeud avant (option implicite) ou après le déplacement de la tête de lecture (option notée par F).

La fonction de sortie, écriture sur le ruban de sortie, est provoquée par une action prédéfinie. Le parcours de l'automate n'étant pas forcément déterministe, les sorties appartenant à une même interprétation pourront être chainées entre elles sur le ruban de sortie.

La philosophie de ces automates s'apparente à celle des réseaux ATN (Augmented Transition Network) de Woods [8] dont la particularité est de munir chaque arc d'un automate fini récursif de conditions et d'un nombre déterminé d'actions.

Le sytème TEX (Transducteur EXpert), développé en langage PASCAL sur DPS8-Multics, intègre les principes de transduction généralisée et autorise différentes facilités utiles plus particulièrement dans le décodage phonétique, en particulier la définition d'un ensemble de paramètres divers (traits ou indices).

L'utilisateur de TEX doit:

- préciser les données en entrée,
- définir les traits ou indices sous forme déclarative: $trait_i = f_i$(entrée),
- décrire les procédures de décodage sous forme de réseaux, ce qui revient à déclarer un ensemble de sextuplets.

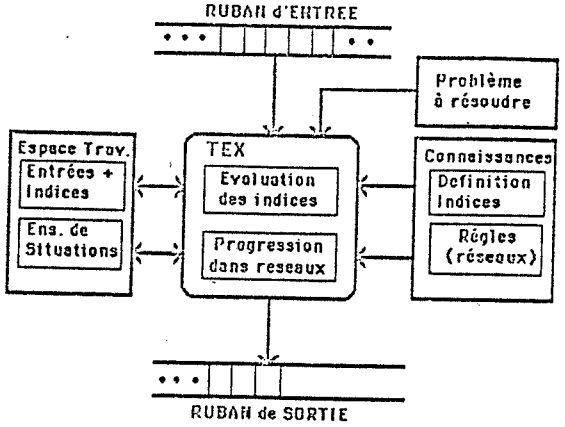


Fig.1 Le Système TEX

Cette tâche de définition est facilitée par les fonctions diverses prédéfinies - On notera qu'elles apparaissent en notation polonaise préfixées et qu'elles peuvent être combinées par composition. Voici quelques exemples de telles fonctions:

mul, add, ret (resp. multiplier, additionner, retrancher) ....

$ou(X_1,...X_n) = sup(X_1,...X_n)$,

$et(X_1,...X_n) = inf(X_1,...X_n)$, non(X) = 1-X,

supf(a,b,X) à valeur dans [-1,1] (voir fig. 2)

etc...

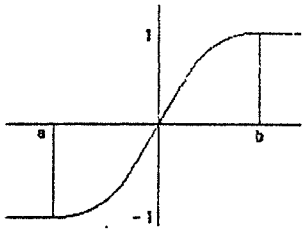


Fig.2
Graphe de supf(a,b,X)

Cette dernière fonction permet de normaliser les paramètres en ramenant leurs variations dans l'intervalle [-1,1].

Les formes d'évolution des paramétres jouent un rôle déterminant dans le décodage phonétique. Dans TEX elles peuvent être approchées en utilisant des combinaisons de fonctions prédéfinies sur les données présentes sous la tête de lecture, ainsi que celles qui précèdent ou qui suivent. Soit X un paramètre, X(n) sa valeur courante (la tête de lecture étant placée sur la nième entrée) est désignée par X. De même X(n+i) {resp. X(n-i)} est désignée par X+i {resp. X-i}.

Voici un exemple d'utilisation sur les données de notre analyse où les phones sont décrits par des

indices divers: aigu-grave, écarté-compact, ..., intensité, durée (notée DUR). On pourra par exemple déclarer au sujet d'un paramètre X:

```
DELTA = div(add(DUR+1, DUR), 2);
```

(intervalle de temps entre le milieu du phone courant et le milieu du suivant)

```
X' = div(ret(X+1, X), DELTA);
X" = div(ret(X', X'-1), DELTA);
```

(dérivées discrètes première et seconde)

Il est ensuite possible de définir, entre autre, une dissemblance entre le phone courant et le suivant au moyen de:

```
DIS = add(abs(X'), abs(Y'),...);
```

(X, Y... étant des paramètres; X', Y'... leurs dérivées obtenus comme précédemment)

```
NDIS = supf(a, b, DIS);
```

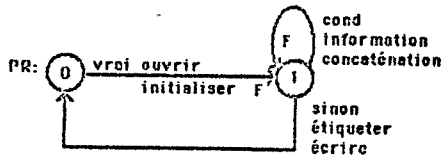
(où NDIS est un indice de ressemblance-dissemblance à valeur dans l'intervalle [-1,1] pour un choix convenable de a, b). Enfin pour appécier si un phone est plus voisin du précédent ou du suivant, il est possible d'introduire:

```
AGGL = ret(NDIS, NDIS-1);
```

Quand certaines valeurs intermédiaires ne sont pas utilisées il est possible de combiner plusieurs étapes pour déclarer la valeur finale intéressante, par exemple:

```
NDIS = supf(a, b, add(abs(X'), abs(Y'),...));
```

La même possibilité de combiner les fonctions existe au niveau du réseau. A titre d'exemple voici un réseau créant des événements par utilisation de AGGL: si AGGL > 0 le phone courant se rattache à l'événement courant, sinon il est le premier d'un nouvel événement. Au passage les informations relatives à l'événement doivent être élaborées. Si un événement est terminé il faut l'écrire sur le ruban de sortie aprés l'avoir étiqueté. Avant d'examiner un nouveau phone il faut de plus "ouvrir" pour un nouvel événement. Voici comment nous pouvons faire ceci en TEX.




```
0   1   vrai   ouvrir; initisliser;F
1   1   cond   informations; concaténation;F
-   0   sinon  étiqueter; écrire;
```

Fig.3 Réseau principal simplifié PR pour la construction d'événements

Dans ce petit réseau, où l'on néglige quelques difficultés, il y a deux paramètres X et Y dont on veut affecter les valeurs maximales {resp. mimales} à MAX=(MAX_X,MAX_X',MAX_Y,MAX_Y') {resp. à MIN=...} (dérivées inclus):

```
initialiser -> aff(MAX,X,X',Y,Y');
               aff(MIN,X,X',Y,Y');
```

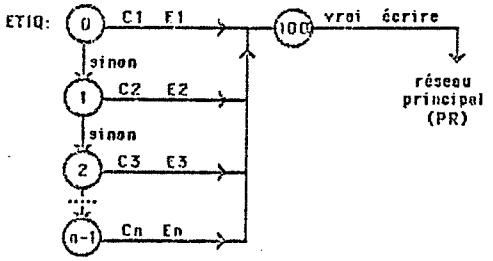
De même:

```
information ->
        aff(MAX,ou(X,MAX_X),ou(X',MAX_X'),...);
        aff(MIN,et(X,MIN_X),et(X',MIN_X'),...);
cond ->  >(AGGL,0)
```

Si l'étiquetage est trop complexe un sous-réseau peut être utilisé; appelons le ETIQ. Dans ce cas la dernière transition s'écrit:

```
-   ETIQ   sinon;
```

Le réseau en question pourrait être de la forme




```
0    100   C1  E1;
-    1     sinon;
1    100   C2  E2;
-    2     sinon;
2    100   C3  E3;
-    3     sinon;
.....
100  PR    vrai écrire;
```

Fig.4 Sous-réseau d'étiquetage ETIQ

donnée dans la figure 4, où Ci est une condition, Ei une affectation d'étiquette, PR le nom du réseau principal.

Comme exemple de Ci Ei (simplifié) on pourrait avoir l'étiquetage de l'occlusion à partir de l'intensité.

```
1  100  et(>(MIN_INTENSITE",0),>(MAX_INTENSITE',0),
          <(MIN_INTENSITE',0),
          <(MIN_INTENSITE,SEUIL))
       aff(ETQ,OCCLUSION);
```

RESULTATS

Recherche des événements

Dans le cadre de cet exposé nous ne pouvons donner que quelques résultats et conclusions qui s'en dégagent. Un réseau du type précédent a été construit sous TEX. Les paramètres utilisés étaient: l'intensité, la durée, les indices aigu-grave, écarté-compact et de friction.

phonè-mes / eff. événem.	p t k	b d g	ũ õ ɔ̃ ɛ̃ œ̃ /cons	l	r	m n	f s ʃ	v z ʒ	Totaux	%
nbre total	20	20	22	9	17	21	22	21	152	100%
nbre à 1 évén.	5	9	3	5	6	19	16	19	82	54%
nbre à 2 évén.	15	11	19	4	9	2	3	2	65	43%
nbre à 3 évén.	0	0	0	0	2	0	3	0	5	3%

La plupart des réalisations en trois événements correspondent à des articulations affaiblies de finale comme dans #novembr'#. D'une manière générale les événements obtenus sont interprétables d'un point de vue articulatoire:

- les occlusives sourdes s'obtiennent comme: $\begin{bmatrix}+occl.\\+sourd\end{bmatrix} /\!\!-\!\!- +expl.$ ou comme $\begin{bmatrix}+occl.\\+occl.\end{bmatrix} + \begin{bmatrix}aspiré\\+expl.\end{bmatrix}$
- les occlusives voient leur occlusion réduite dans le contexte /[+nas]—.
- les voyelles nasales apparaissent en contexte /—[+cons] comme: [+syll] ou [+syll]+[+voc].

Recherche des unités syllabiques

Cette recherche vise à répartir les événements en:

SCO: syllabe croissante ouverte,
FS: fermeture syllabique,
CROIS: attaque ou SCO non marquée,
SISFL: bruit ou souffle,
FRIC: friction,
TRANS: transition.

Dans le système mis en place sous TEX la syllabation a été repartie en trois sous-réseaux: BR, SCO et FS, munis de conditions d'entrée et de conditions de sortie fondées essentiellement sur les étiquettes des événements (voir annexe 2, 3).

Les événements dégagés par l'analyse précédente ont pu être étiquetés selon les catégories exclusives suivantes:

NY: caractère +syllabique,
NYF: tendance +syll moyenne,

(ces étiquettes sont conditionnées par [-fric] et [+intensité,-concave])

OV: occlusion voisé [+occl,+voi],
OC: occlusion non OV,
CU: closion non OV [+clos,+voi],

(ces étiquettes sont conditionnées par la concavité de l'intensité et l'intensité elle-même)

FRS: friction sourde, FV: friction voisée,
FVD: friction voisée douce,
XX: transition non marquée,
SIO, SIF: silence occlusif ou fricatif,

Un même événement peut recevoir plusieurs étiquettes. C'est ainsi que le caractère explosif est recherché pour les occlusions et les silences:

EXV: explosion voisée, EX: explosion,
SD: attaque sourde,

(ces étiquettes sont conditionnées par l'apport d'aigu en fin d'occlusion ou de silence).

Les vocaliques sont elles-mêmes caractérisées par le timbre qui peut être II: type [i], AA: type [a], OU: type [u], EE: timbre moyen (central).

Les fricatives peuvent avoir une caractérisation complémentaire: STR pour strident, JZ pour le type [j] ou [z] moyen, CY pour le type [j] [ʒ] moyen.

Dans le corpus étudié (voir annexe1) nous n'avons pas trouver d'étiquetage que l'on peut considérer comme faux si l'on considère comme naturel des faits prévisibles comme:

[+occl] -> [+clos] / [+nas]—

Ainsi les [d] du corpus sont apparus comme (OV EXV) ou (CU EXV), ce dernier cas dans des contextes /[+nas]—.

De même les voyelles fermées et consonnes vocaliques ne sont pas toujours bien démarquées et se retrouvent souvent sous l'étiquette NYF. Cependant bien des ambiguïtés peuvent être levées au niveau syllabique. Ainsi deux événements étiquetés NYF + NY s'interprètent par [+cons] [+syll].

Le cas des syllabes enchaînées sans consonnes marquées demeure cependant difficile à traiter. A titre d'exemple la séquence <<21 janvier>> a été analysée comme suit (en anticipant sur la méthode de syllabation utilisée).

Il apparait que le groupe [e ɛ̃] ne se syllabe pas de manière conventionnelle, sans que l'on puisse vraiment dire si l'analyse fournie par notre système est vraiment fautive. On observera de plus l'étiquette CROIS pour la syllabe [vje] qui provient de l'enchaînement de deux syllabes sans fermeture syllabique détectée.

phonèmes	v ɛ̃		t e		ɛ̃	ʒ ɑ̃	v j e	:	#
événem.+ étiquettes	CU NY EE	NYF EE	OV FVD NY EX EE SD		NY EE	FVD NY NYF AA OU	FVD CU NY JZ FVD II CY	NYF CU EE CY	SIO
temps	10 11 10 17	6	7 5 7 12 7		20 20	10 14 16 10 30	8 6 13 8 6 29	10 6	
syllabes	SCO	FS	SCO		FS	SCO	CROIS	FS	
temps	27	6	19		20	40	27	16	
	33		39			40	43		

Signalons encore un autre type de difficulté: l'enchaînement des voyelles fermées et des nasales ou des latérales. Le risque ici est la fusion de deux phonèmes dans un seul événement, ce qui s'est produit dans des séquences comme <<numéro>> où nous avons observé soit la fusion de [ny] dans le segment NY, soit de [ym] dans NYF.

Mis à part ces difficultés la syllabation ascendante donne d'excellents résultats.

CONCLUSION

Cette étude montre l'intérêt de l'analyse en événements phonétiques. Ils permettent l'accès lexical par des méthodes appropriées (traitant des possibilités d'insertion et de fusion entre autres). De plus les syllabes peuvent être bien dégagées en tant que suites d'événements, ce qui permet aussi d'envisager l'accès lexical utilisant l'information syllabique.

Cependant l'orientation que nous avons prise s'est révélée fragile lorsque nous avons voulu passer à d'autres corpus par le fait que les diverses conditions portant sur les indices sont sensibles aux caractéristiques générales de la voix: force, timbre...

C'est pourquoi nous mettons l'accent à l'heure actuelle sur des réseaux de structures plus simples, plus répétitives, modulables par jeux de paramètres caractérisant le locuteur. Cette étude est menée en liaison avec le GRECO de la communication parlée en utilisant la base de données des sons du francais.

REFERENCES

[1] G. Pérennou, 1980, "ARIAL II: system for speech recognition",IAPR, "Automatic Speech Analysis and Recognition", Reidel Pub. Cy., pp. 269-275.

[2] G. Pérennou, 1982, " The ARIAL II speech recognition system", J.P. Haton (ed).

[3] G. Pérennou, M. de Calmès, M. Bui Van, 1981, "Décodage lexical et composante phonologique dans ARIAL II", 12ème JEP, Montréal, pp. 49-61.

[4] G. Pérennou, M. de Calmès, 1981, "Le décodage au niveau phonologique dans ARIAL II", actes du séminaire GALF "Processus d'encodage et de décodage phonétiques", pp.191-203.

[5] J. Caelen, N. Vigouroux, G. Pérennou, 1983, "Structuration des informations acoustiques dans le projet ARIAL", Speech Communication, n° spécial 2,3 pp.219-222.

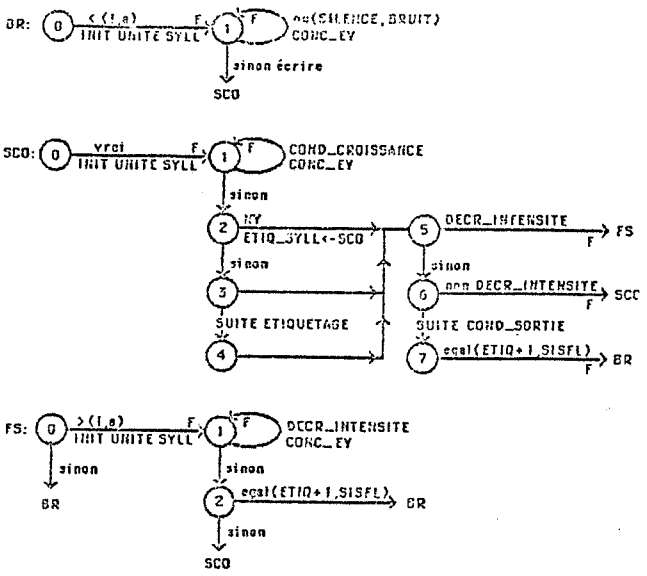
[6] N. Vigouroux, J. Caelen, 1985, "Segmentations en vue de l'organisation d'une base de données acoustique et phonétique", 14ème JEP, GALF-CNRS, Paris.
[7] G. Pérennou, M. de Calmès, 1983, "Spécifications pour un système générateur de modèles de décodage phonétique", Speech Communication, n°2, pp.227-230.
[8] W.A. Woods, 1970, "Transition network grammary for langage analysis", Communication ACM, vol. 13, pp. 591-606.

ANNEXES

Annexe 1- Corpus Date-Téléphone

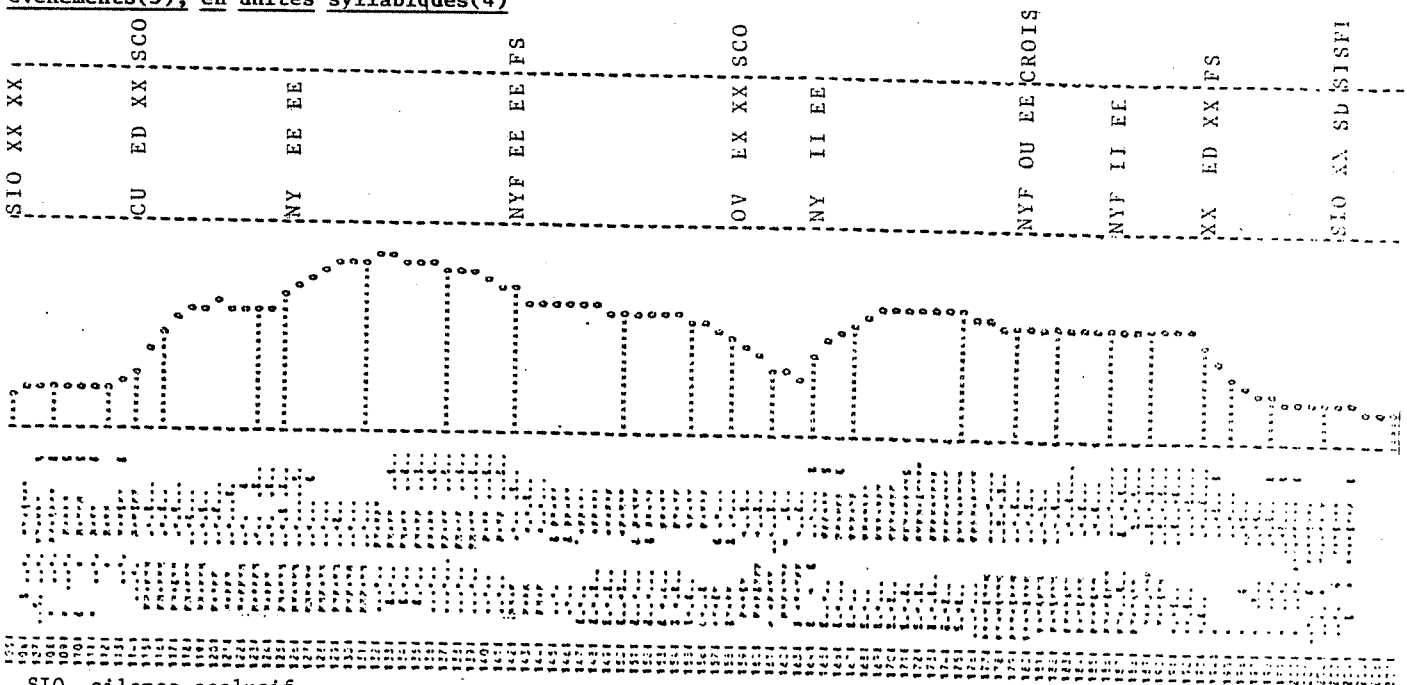
1- ## lundi # 28 février # 1979 ##
2- ## mardi 21 janvier # 1898 ##
3- ## mercredi ## 14 mars 15 cent 16 ##
4- ## jeudi 21 avril ##
5- ## vendredi 12 mai ##
6- ## samedi 13 juin ##
7- ## dimanche 8 août ##
8- ## 10 juillet ##
9- ## 6 septembre ##
10- ## 20 octobre ##
11- ## 31 novembre ##
12- ## 7 décembre ##
13- ## numéro de téléphone de mr Lavayssière ##
14- ## numéro de téléphone # de mme Jouany ##
15- ## numéro de téléphone # 25 # 32 # 21 ##

Annexe 2- Forme schématique des réseaux de syllabation ascendante



Notation
INIT UNITE SYLL: action d'initialisation d'une unité syllabique.
CONC_EY: adjonction d'un événement supplémentaire à l'unité syllabique.
COND_CROISSANCE: critère utilisant l'énergie, sa dérivée et sa dérivée seconde ainsi que le caractère non fricatif.
DECR_INTENSITE: inverse du précédent.
a: une constante.
I: l'intensité.

Annexe 3- Analyse de <<lundi>> comportant une représentation spectrale (1) en phones(2), en événements(3), en unités syllabiques(4)



SIO silence occlusif
CU occlusion faible
OV occlusion voisée
NY noyau
NYF noyau faible

ED explosion douce
EX explosion
II voyelle fermée diffuse
OU voyelle fermée non diffuse

XX consonnantique non typé
EE vocalique non typé
SD sourd

SCO syllabe croissante ouverte
FS fermeture syllabique
CROIS attaque consonnantique croissance d'énergie vocalique
SISFL silence ou souffle

PROPOSITIONS POUR UNE SEGMENTATION ET UN ETIQUETAGE HIERARCHISES APPLICATION A LA BASE DE DONNEES ACOUSTIQUE DU GRECO COMMUNICATION PARLEE

D. Autesserre et M. Rossi

Institut de Phonétique Générale et Appliquée - L.A. 261 C.N.R.S.
Université de Provence I - 29, avenue Robert Schuman, 13621 Aix

In this paper we propose a method of segmentation and labelling for the acoustic data base of the GRECO Parole.
In a first time, we give a broad phonetic transcription (normative) of the speech data. This transcription brings to the data base a phonetic information which completes usefully the labelling.
Then, we illustrate a method of segmentation founded on the identification of discontinuities which stay in the acoustic speech signal.
The corresponding acoustic events or segments are labelled hierarchically: first in macro-classes defined by fundamental cues (oral vowel, stop consonant, etc.), then in micro-classes constituted by the successive phases of the acoustic events (onset, etc.) and by attributes (like burst, friction, aspiration, etc.).

Les phonéticiens maîtrisent de longue date les techniques de segmentation du signal de parole; les spécialistes de la reconnaissance automatique, de leur côté, segmentent le signal acoustique à partir de critères tels que l'instabilité spectrale. Les uns et les autres s'appuient sur des discontinuités acoustiques indéniables. Mais nous devons souligner avec force que les évènements repérés ne sont pas en correspondance biunivoque avec les unités discrètes de la taille du phonème : la segmentation en phonèmes apparaît comme une opération de haut niveau chez l'auditeur.

Si la segmentation en phonèmes n'est pas reflétée dans le signal de parole, il n'en demeure pas moins que l'auditeur n'est pas insensible aux discontinuités acoustiques auxquelles nous venons de faire allusion. Il les traite à l'aide d'un ensemble de processus assimilable à un système de règles lui permettant d'identifier les unités nécessaires à la reconnaissance du signal.

De nombreux travaux concernant la perception de la parole tendent à montrer que les traits phonétiques constituent l'interface entre le signal acoustique et le phonème [1] . Dans cette perspective, on saisira tout l'intérêt qu'il y a à partir des discontinuités acoustiques pour une segmentation du signal plus directement reliée à cette unité intermédiaire indispensable qu'est le trait phonétique. Ce lien apparaît plus nettement si l'on fait intervenir la notion de contraste réservée par R. Jakobson "aux cas où la polarité de deux unités est mise en reliefpar leur contigüité dans l'expérience sensorielle": ainsi par exemple "le contraste grave-aigü dans la séquence /pi/ ou le même contraste mais avec inversion dans l'ordre des éléments dans la séquence /tu/ [2] ". Toutefois il convient de garder présent à l'esprit que les traits ne sont pas organisés linéairement mais simultanément et présentent ainsi des caractéristiques de chevauchement. Celles-ci proviennent du fait que les traits s'organisent au sein de l'unité supérieure qu'est la syllabe.

La méthode de segmentation que nous présentons est donc fondée, d'abord, sur l'identification d'évènements acoustiques, de segments, qui correspondent à des macro-classes (définies par des "traits fondamentaux" [3]), elles-mêmes précisées dans leurs phases successives par des attributs (micro-classes) [4] .

TRANSCRIPTION PHONETIQUE

La segmentation proprement-dite est précédée d'une transcription phonétique. Les symboles de cette transcription apparaîtront dans la base de données. Cette transcription large, de type normatif, reflète non pas les réalisations phonétiques concrètes de la parole mais les unités de référence de la langue (les phonotypes).

Par exemple, dans la séquence "mon ami", la voyelle "on" de "mon", quelle que soit sa réalisation sera transcrite comme voyelle nasale et recevra l'étiquette VN. De même, les voyelles atones ne seront pas transcrites avec leur timbre précis mais à l'aide de symboles d'ordre archiphonémique: dans le mot "soleil" la voyelle "o", quelle que soit sa réalisation concrète, ouverte ou fermée, sera notée [O] (en majuscule). Il appartiendra ensuite à l'analyste de préciser si, dans le premier cas, la voyelle est nasale ou pas, dans le second cas, si elle est ouverte ou fermée. [5]

Ce choix ne peut être opéré au moment de la segmentation car les traits en question n'apparaissent pas temporellement sur le signal comme des discontinuités acoustiques directement visibles.

Pour les groupes de consonnes, les phénomènes d'assimilation ne seront pas représentés dans

la transcription (cf. plus bas l'exemple développé "mēd(e)cin").

En effet, une transcription étroite à ce niveau serait dangereuse : elle introduirait des critères subjectifs et ferait double emploi avec l'étiquetage qui identifie des évènements acoustiques fins. Le transcripteur n'a pas pour fonction de se substituer à l'analyste.

Pour cette raison, il est proposé de placer le symbole de transcription phonétique au milieu du segment correspondant: les faits de coarticulation qui imposent un chevauchement des unités pourront conduire ultérieurement l'analyste, en fonction de ses besoins, à interpréter tel micro-segment comme un évènement de la consonne ou de la voyelle.(cf. plus bas sur ce point les relations entre transcription phonétique, étiquetage et accès aux micro-segments.- et figure -)

Les laboratoires chargés de la mise au point de la méthode de segmentation et d'étiquetage (Aix, Grenoble et Toulouse) se sont mis d'accord sur cette conception de la transcription qui apporte à la base de données une information phonétique complétant utilement l'étiquetage.

SEGMENTATION

Nous illustrerons par un premier exemple la méthode de segmentation que nous préconisons.- Soit la séquence "attaque" [atak] , schématisée sur la figure .

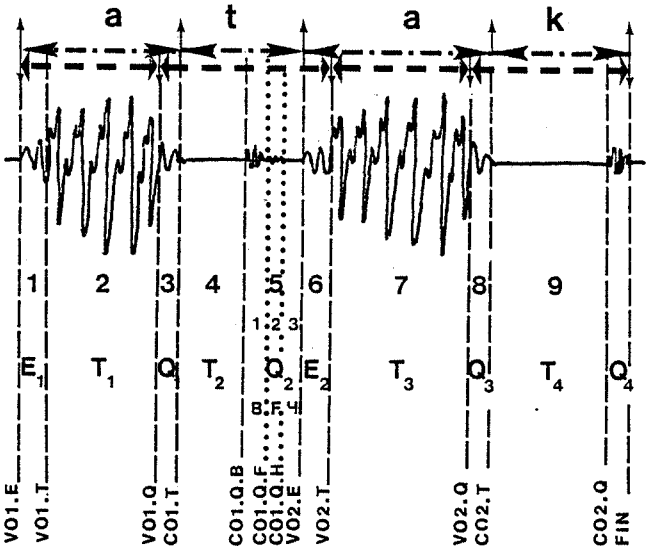


Figure . Représentation oscillographique schématisée de la séquence "attaque".

Les symboles de transcription phonétique sont placés au dessus et au milieu des segments auxquels ils sont attribués.

Les lettres, renvoyant au texte, indiquent les différents évènements isolés.

Les étiquettes sont placées verticalement au début du segment qu'elles spécifient.

Les flèches indiquent les possibilités d'extension des segments étiquetés afin d'inclure, à droite et/ou à gauche, les moments de coarticulation compris dans les segments limitrophes.

Sur la voyelle [a_1] on peut identifier trois évènements :

1. La voyelle s'établit avec un certain délai. En effet, les premières périodes ne contiennent pas toutes les caractéristiques acoustiques qui déterminent le timbre de cette voyelle. Nous appellerons cette phase E : établissement de la voyelle (dans ce cas E1). Ce type d'établissement peut être précédé, à l'initiale, d'un coup de glotte (?) ou d'un souffle (H).

2. Viennent ensuite les périodes caractéristiques de la voyelle pour lesquelles les deux premiers formants (F1 et F2) sont nettement définis. Cette partie correspond au début de la tenue proprement-dite T (ici T1).

3. A la fin de la voyelle les périodes perdent l'une et/ou l'autre de leurs composantes spectrales (d'abord F2 puis F1). Cette zone de changement correspond au début du relâchement Q (Q1 pour l'exemple représenté).

Viennent ensuite les évènements et les phases qui caractérisent la consonne occlusive [t] .

4. L'occlusive non voisée [t] commence au début de la tenue silencieuse (T2). En effet nous décidons provisoirement d'attribuer l'étiquette de macro-classe VO à l'intervalle E1 (établissement de la voyelle)- T2 (tenue de l'occlusive). Dans le cas où E serait précédé d'un souffle (H) ou d'un coup de glotte (?), alors VO serait compris d'une part entre H et T2 et d'autre part entre ? et T2.Le choix des caractères VO plutôt que VN est directement lié à la transcription phonétique préalable du signal (cf. plus haut).

5. La discontinuité qui apparaît maintenant est en relation avec le relâchement de l'occlusion (Q2) habituellement appelé explosion dans le cas d'une consonne occlusive.

Q2 peut comporter trois phases :

5.1 Le bruit d'explosion proprement-dit désigné par B (bruit ou burst).

5.2 B est suivi éventuellement d'une phase de friction F

5.3 qui peut être suivie elle-même d'un souffle H.

6. Après l'une de ces trois phases B, F ou H se présentent alors une ou deux périodes de voisement pour lesquelles F1 et/ou F2 ne sont pas établis. D'ailleurs, ces périodes peuvent comporter aussi des bruits. C'est l'établissement de la voyelle [a_2] noté E2.

7. et 8. renvoient aux deux phases suivantes de la voyelle [a_2] : sa tenue T3 et son relâchement Q3.

9. Tenue silencieuse de [k] : T4.

L'intervalle compris entre T2 et E2 porte l'étiquette CO (macro-classe de la consonne occlusive), l'intervalle compris entre E2 et T4 est étiqueté VO (macro-classe de la voyelle orale).

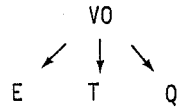
La décision d'attribuer à la macro-classe VO un segment partant de l'évènement E2 peut paraître arbitraire. Il serait tout aussi légitime d'inclure l'intervalle E2-T3 dans la macro-classe CO. Il n'appartient pas au responsable de la segmentation d'en décider : cet intervalle correspond à un moment de coarticulation. En effet, les frontières entre les segments sont nécessairement floues en raison du chevauchement des unités. Pour arbitraire qu'elle ait pu paraître, la déci-

sion prise a au moins le mérite d'inclure dans un même segment tous les évènements voisés.

L'essentiel à notre avis sera de fournir à l'utilisateur de la base de données des portions de signal plus larges que les segments étiquetés et qui incluront, à droite et à gauche, les moments de coarticulation compris dans les segments limitrophes. De la sorte, si l'utilisateur demande les consonnes occlusives CO il aura à sa disposition les intervalles compris entre Q1 et T3 où l'on trouve Q1 et E2, qui ont pourtant été attribués aux voyelles adjacentes.

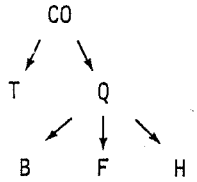
ETIQUETAGE HIERARCHISE

On aura remarqué que l'étiquetage proposé est hiérarchisé (cf. tableau). En effet une macro-classe telle que VO domine les évènements E, T, Q:



Les macro-classes sont notées par au moins deux caractères alphabétiques; les évènements,eux, ne sont notés que par un seul caractère.

La hiérarchie que nous proposons peut comporter plus de deux niveaux. C'est le cas, par exemple, pour les consonnes occlusives CO

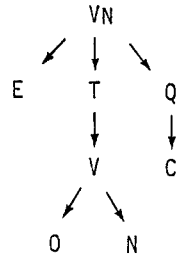


où B (bruit d'explosion ou burst), F (bruit de friction), H (souffle)sont inclus dans une phase plus large de relâchement (Q).

Nous tenons à souligner que les symboles unités tels que B, F, H désignent des évènements (discontinuités acoustiques) et qualifient aussi l'intervalle compris entre deux discontinuités : ainsi B désigne le bruit d'explosion qui commence au point B et se termine au point F. Il en est de même de F et de H. Ces symboles représentent donc des évènements, des micro-segments et des attributs.

La phase de relâchement n'est pas forcément présente après la tenue de l'occlusive. Par exemple dans "méd(e)cin" [mEd(ə)sɛ̃] elle disparait en position implosive : B n'est plus inclus dans Q, il indique alors non pas la fin de l'occlusive [d] mais bien le début de la constrictive [s] .

Un autre exemple intéressant permet d'illustrer la validité des symboles d'étiquetage proposés. Les voyelles nasales sont très souvent suivies d'un appendice consonantique, lorsqu'elles précèdent une consonne, et ceci même dans des usages non méridionaux du français. De nouveau, la macro-classe VN domine les évènements E, T, Q qui dominent eux-mêmes les deux attributs V (voyelle) et C (consonne). Le segment V comporte lui-même une partie orale (O) initiale suivie d'une partie nasale (N), l'une et l'autre séparées par un pic d'instabilité spectrale. Nous obtenons, cette fois-ci, une hiérarchie à quatre niveaux :



Pour noter ce même étiquetage nous proposons le formalisme suivant :

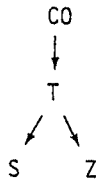
VN.E

VN.T.V.O

VN.T.V.N

VN.Q.C

Ce dernier exemple montre clairement que les symboles à une lettre, qui désignent à la fois des évènements et des micro-segments, représentent également des attributs : C (consonne) est l'attribut du relâchement Q du macro-segment VN. Ce sera aussi le cas des symboles Z désignant le voisement et S le non voisement. Dans l'exemple déjà développé de [mEd(ə)sɛ̃] , la consonne [d] étiquetée CO peut être voisée sur la première partie de sa tenue et dévoisée dans la deuxième partie, à proximité de la constrictive subséquente : S et Z sont bien alors des attributs de la tenue T et l'étiquetage de l'occlusive est réduit à :



soit : CO.T.S et CO.T.Z . Ainsi le voisement n'apparait pas dans les étiquettes de macro-classes mais constitue un attribut de troisième niveau.

Ceci met bien en évidence la complémentarité entre la transcription large, où la consonne occlusive demeure représentée par le symbole [d] , et l'étiquetage fin qui implique sa réalisation comme mi-sourde. Dans ce cas précis, les discontinuités présentes dans le signal permettent d'affiner la transcription large, mais il n'en est pas toujours de même évidemment.

Dans la liste desmacro-classes (cf. tableau) apparaissent certains symboles qui appellent une explication complémentaire : ainsi CA, réservé aux consonnes approximantes. Certaines constrictives, en français, peuvent connaître une réalisation plus aperte, non accompagnée de bruit. C'est le cas, par exemple de la consonne "v" en position

intervocalique, réalisée [ʋ] dans un mot tel que "avec" qui sera transcrit alors [aʋɛk] . Nous reprendrons, pour désigner ces variantes apertes des consonnes constrictives, le nom d'approximantes, déjà proposé par P. Ladefoged [6] .

Les consonnes comprises dans les macro-classes CA (approximantes) mais aussi CV (consonnes vocaliques) sont souvent difficiles à segmenter. Ainsi la séquence voyelle - consonne vocalique glissante (telle que [j]) - voyelle ne contient pas toujours de discontinuité acoustique susceptible d'indiquer une frontière entre ces trois unités. Le passage de la voyelle à [j] ou inversement se fait graduellement. En l'absence totale de discontinuité acoustique cette séquence ne peut pas être segmentée. On la fait alors précéder d'un symbole d'indétermination segmentale X. Dans ce cas l'étiquetage est simplifié : à la place de la notation complète faisant intervenir successivement les symboles VO, CVJ, VO par exemple, on utilise pour cette séquence non segmentable la notation réduite XVJV.

TABLEAU RECAPITULATIF DES SYMBOLES
DE L'ETIQUETAGE HIERARCHISE

Premier niveau : Macro-classes

VO : voyelle orale
VN : voyelle nasale
CVN : consonne vocalique nasale
CVL : consonne vocalique latérale
CVR : consonne vocalique de type R
CVJ : consonne vocalique glissante ([j , ɥ , w])
CA : consonne approximante (type [ʋ])
CS : consonne constrictive
CO : consonne occlusive

X : indétermination segmentale

Deuxième niveau : Micro-classes (phases)

E : délai d'établissement de la voyelle
T : tenue d'une voyelle ou d'une consonne
Q : relâchement d'une voyelle ou d'une consonne

Troisième niveau : Micro-classes (attributs phonétiques de 1er ordre)

V : voyelle
C : consonne
Z : voisé
S : non voisé
H : souffle
? : coup de glotte
B : bruit d'explosion
F : bruit de friction

Quatrième niveau : Micro-classes (attributs phonétiques de 2ème ordre)

O : oral
N : nasal

Les procédures de segmentation et d'étiquetage que nous proposons pour la base de données des sons du français parlé (BDSONS) identifient les discontinuités acoustiques ou évènements acoustiques sans les relier directement aux évènements du domaine articulatoire qui se situent à un autre niveau. La définition des relations entre niveau articulatoire et niveau acoustique pose des problèmes difficiles qui nécessiteront des études spécifiques.

L'identification proposée des évènements acoustiques présente un double avantage :

1°/ elle permet de repérer les macro-classes qui se réfèrent au mode d'articulation. Il est bien clair que la limite entre ces macro-classes n'est en aucune façon un reflet des frontières des unités phonologiques discrètes. Les décisions que nous avons prises pour indiquer le début et la fin des macro-classes sont essentiellement opératoires. En raison de la coarticulation, les divers segments désignés par les symboles de macro-classes sont l'objet de chevauchements courants dans la chaîne parlée.

2°/ La segmentation opérée en utilisant l'étiquetage de deuxième et troisième niveau, fournit une information sur un domaine particulier de la coarticulation, celui qui a trait aux indices de mode d'articulation. En revanche, la personne chargée de segmenter et d'étiqueter la base de données ne peut intervenir sur la spécification des lieux d'articulation qui relèvent, eux, de l'analyse et apporteront l'information complémentaire nécessaire pour bien connaître l'organisation de la coarticulation.

A un autre niveau, le problème de la syllabe ne se pose donc pas. En effet si l'on considère cette dernière comme une unité de fusion ([7] et [8]), l'information spectrale concernant les indices des traits de tonalité s'avère indispensable pour indiquer les limites syllabiques.

REFERENCES BIBLIOGRAPHIQUES

[1] M. Rossi, "Les traits acoustiques", La Linguistique, vol. 13, fasc. 1, pp.63-82, 1977.

[2] R. Jakobson, Essais de Linguistique Générale, Paris, ed. de minuit, p.105, 1963.

[3] R. Jakobson, G.M. Fant et M. Halle, Preliminaries to speech analysis, Cambridge (Massachusetts) MIT Press, p.18, 1963.

[4] Pour d'autres points de vue concernant la segmentation cf. les communications à ces mêmes journées d'étude de la parole de C. Abry, C.Benoît et L.J. Boë d'une part et de J. Caelen, G. Caelen-Haumont, J. Malet, N. Vigouroux et C. Barrera d'autre part. Les laboratoires de recherche d'Aix, Grenoble et Toulouse se sont mis d'accord sur une méthodologie d'ensemble qui sera présentée à ces mêmes journées dans une communication commune, sous le titre "Propositions pour la segmentation et l'étiquetage d'une base de données des sons du français".

[5] Des indications concernant les différences de timbre vocalique dans les usages régionaux sont fournies dans l'ouvrage de F. Carton, M. Rossi, D. Autesserre et P. Léon, Les accents des français Paris, Hachette, 94p., 1983.

[6] P. Ladefoged, notamment dans A course in Phonetics, New-York, Harcourt - Brace - Jovanovich, p.10, 1975.

[7] B. Malmberg, "Gémination, force et structure syllabique en latin et en roman", Orbis Litterarum supplementum 3, Etude romane dédiée à A.Blinkenberg, 1963; repris dans Phonétique générale et romane, La Haye - Paris, Mouton, pp.343-348, 1971.

[8] G. Faure, "Analyse acoustique de deux allophones du l final anglais", Papers in Linguistics and Phonetics to the memory of P. Delattre, La Haye - Paris, Mouton, pp.117-127, 1972. Plus précisément en ce qui concerne le problème de la syllabe, pp.124-126.

SEGMENTATIONS EN VUE DE L'ORGANISATION D'UNE BASE DE DONNEES ACOUSTIQUES ET PHONETIQUES

Vigouroux N. - Caelen J.

CERFIA-TAPT - UA au CNRS,No 82
Université Paul Sabatier - 118, route de Narbonne - 31062 Toulouse Cedex

ABSTRACT

This article purports to describe the two segmentation devices used at the acoustic decoding level in the ARIAL Project:
- automatic-segmentation device
- manual-segmentation device assisting an experienced phonetician's work.

Automatic segmentation is performed on the basis of a spectral variability criterion relying on temporal evolution. The segments secured are infra-phonemic and homogeneous (stable).

Manual segmentation is performed with a spectrogram editor which enables a phonetician to define phonetic facts by placing flag-marks of four types: acoustic, phonetic, syllabic and prosodic.

1. INTRODUCTION

En reconnaissance, le problème du choix de l'unité phonétique ou linguistique de décision reste posé: trait, segment acoustique (infra-phonème), phonème, etc [1]. En analyse, il semble donc raisonnable de partir d'une unité neutre infra-phonémique entièrement objective pour former des unités de plus en plus larges en tenant compte des modèles de perception qui sous-tendent les systèmes de reconnaissance.

En parallèle à cette segmentation en phones, il est nécessaire de déceler les faits phonétiques par une phase de codage dite "d'expertise". Il en résulte des méthodes de travail fondées sur la définition et la normalisation des critères de codage.

Dans cet article, nous exposons les principes relatifs à ces deux segmentations: en phones et en faits phonétiques.

2. LA SEGMENTATION AUTOMATIQUE: LES PHONES

2.1 - Historique

La première segmentation proposée dans le projet ARIAL 1 était une segmentation en éléments acoustiques, fondée sur des variations d'indices [2] et avait pour fonction de définir la nature de l'échantillon: stable, transitoire ou frontière i.e. de grande instabilité. Notons que la difficulté majeure de ce traitement réside dans le fait que l'instabilité spectrale est par nature plus grande à l'intérieur de certains phonèmes (fricatives, ...) que dans d'autres (liquides ...). Cette fluctuation (discontinuité) spectrale explique la difficulté à adapter les seuils d'instabilité (variation) vis-à-vis des valeurs d'indices prises lors de la catégorisation de l'échantillon. Ce principe a par conséquent, été abandonné à cause d'une part, du mauvais séquencement des suites liquide-liquide par exemple, et d'autre part en raison de la trop forte dépendance des seuils aux réalisations spectrales.

Cette segmentation a donc été remplacée par une segmentation dite "infraphonémique" dont le principal objectif est d'obtenir des segments courts indépendamment d'une unité phonétique donnée pour ne pas figer dès le niveau acoustique, le choix de l'unité minimale de reconnaissance.

Par ce procédé, les risques d'erreurs classiques (insertions, fusions, substitutions, etc) d'unités phonétiques sont supprimés puisqu'aucune décision n'est prise au niveau acoustique. Les segments (appelé phones) produits seront ensuite transformés en unités de plus en plus larges par les niveaux linguistiques [3].

2.2. - Les indices de segmentation

Notre méthode de segmentation actuelle repose sur l'évolution temporelle de paramètres acoustiques et prosodiques en chaque point du spectre échantillonné obtenu par modèle d'oreille [2]. Les paramètres acoustiques utilisés sont des indices non formantiques [4]. Ce sont respectivement les indices: Aigu-Grave, Fermé-Ouvert, Bémolisé-Diésé, Ecarté-Compact, Doux-Strident. Ces indices ont été retenus pour:
- leur facilité de calcul,
- et leur qualités descriptives [2].

A cet ensemble, nous avons ajouté l'énergie du signal et la dérivée spectrale en raison de leur pouvoir de séparer grossièrement les phonèmes.

2.3. - La fonction de segmentation

Son rôle est de quantifier la variation globale des indices acoustiques et prosodiques.

Elle opère sur chaque échantillon de 8, 4 ou 2ms, comme suit:

a) Codage des indices,

b) Incrémentation de la fonction de segmentation à chaque variation d'indices,

c) Calcul d'un seuil et positionnement éventuel d'une marque.

a) Codage des indices: Trois types de codage des variations d'indices ont été retenus:

(I) le codage dérivée:

$$k_i^n = (p_i^n - p_i^{n-1})/\Delta_i,$$

(II) le codage delta:

$$k_i^n = [p_i^n/\Delta_i] - [p_i^{n-1}/\Delta_i],$$

où [ ] représente la partie entière

(III) le codage "delta modifié":

$$k_i^n = [p_i^n/\Delta_i] - [\hat{p}_i/\Delta_i],$$

avec

k_i^n, le codage de la variation de l'indice i,

p_i^n, la valeur de l'indice à l'échantillon n,

p_i^{n-1}, la valeur de l'indice à l'échantillon n-1,

$\hat{p}_i$, la valeur du dernier extrêmum de l'indice,

Δ_i, le pas de codage de l'indice i.

Ce pas est indépendant de la discontinuité spectrale et/ou du locuteur.

b) Fonction de segmentation:

$$Fseg^n = \sum_{\sim} k_i^n.$$

c) Calcul du seuil et principe de positionnement:

Un seuil variable S^n est calculé à chaque nouvel échantillon. Ce seuil dépend de l'intervalle de temps entre deux segmentations: plus la distance entre l'échantillon courant et la dernière marque est importante, moins le seuil est grand, rendant par là, le système plus sensible aux faibles variations (Voir Schéma ci-contre). Dès que $Fseg^n > S^n$, une marque de segmentation est positionnée.

2.4. - Comparaison des codages

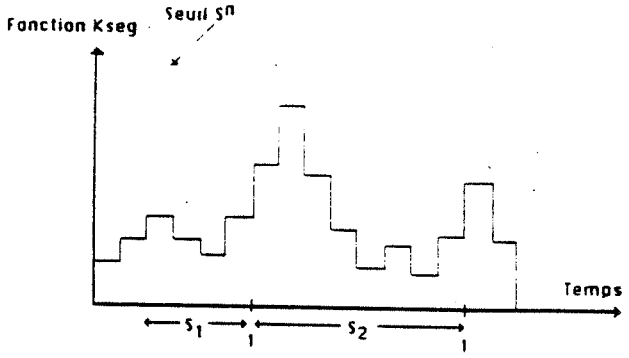
Nous considérons comme homogènes non seulement les zones stables mais également les phases de transition dont la variation est lente et uniforme. Un critère de comparaison doit donc tenir compte de ces exigences et s'attacher notamment aux cas des phonèmes fluides: liquides, semi-voyelles, voyelles orales mais aussi au cas de phonèmes discontinus: occlusion.

Si l'on considère par exemple comme "idéale" une segmentation des voyelles en contexte CVC, en trois phases: (e) établissement, (t) tenue et (c) coda, alors il est nécessaire de favoriser le codage II. En effet, ce codage est sensible aux faibles variations continues. Mais en contrepartie, il peut produire une segmentation trop fine (Fig. 1, Flèches 1 et 2).

Inversement les fortes discontinuités sont mieux détectées par le codage I qui n'effectue pas de lissage temporel mais qui au contraire, peut masquer les discontinuités mineures. Ce codage produit donc des segments relativement longs dans lesquels peuvent se trouver plusieurs zones de faibles discontinuités spectrales (par exemple, phase transitoire liquide- voyelle).

La décroissance du seuil S^n est choisie pour que l'on obtienne des segments de durée $\leq$ 60 ms. En effet, lorsque ce seuil s'annule, il déclenche le positionnement d'une marque arbitraire même lorsqu'il n'y a aucune discontinuité dans le spectre. Notons que si l'on n'imposait pas une telle marque, la plupart des voyelles orales seraient rassemblées en un seul phone (Codage I, Flèches 3 et 4).

En conclusion, nous avons retenu un compromis entre le codage I et le codage II: le codage III "delta modifié" en réduisant la portée du lissage entre deux extrêma pour répondre à nos objectifs d'homogénéité: le signal parole (ici le phone) est stationnaire sur des intervalles courts. On peut de ce fait, caractériser un phone par un vecteur moyen d'indices et de traits ce qui rapelle les techniques actuelles de codage vectoriel [7].



3. L'ETIQUETAGE SEMI-AUTOMATIQUE: LES FAITS PHONETIQUES

Une autre segmentation est de s'appuyer sur des connaissances phonétiques pour définir des segments en relation avec les concepts habituels des phonéticiens. Pour cela depuis plusieurs années, à travers le problème du décodage phonétique, le laboratoire CERFIA s'est attaché à définir des concepts d'étiquetage phonétique et par conséquent à réaliser des procédures semi-automatiques de segmentation reposant sur les phones décrits précédemment.

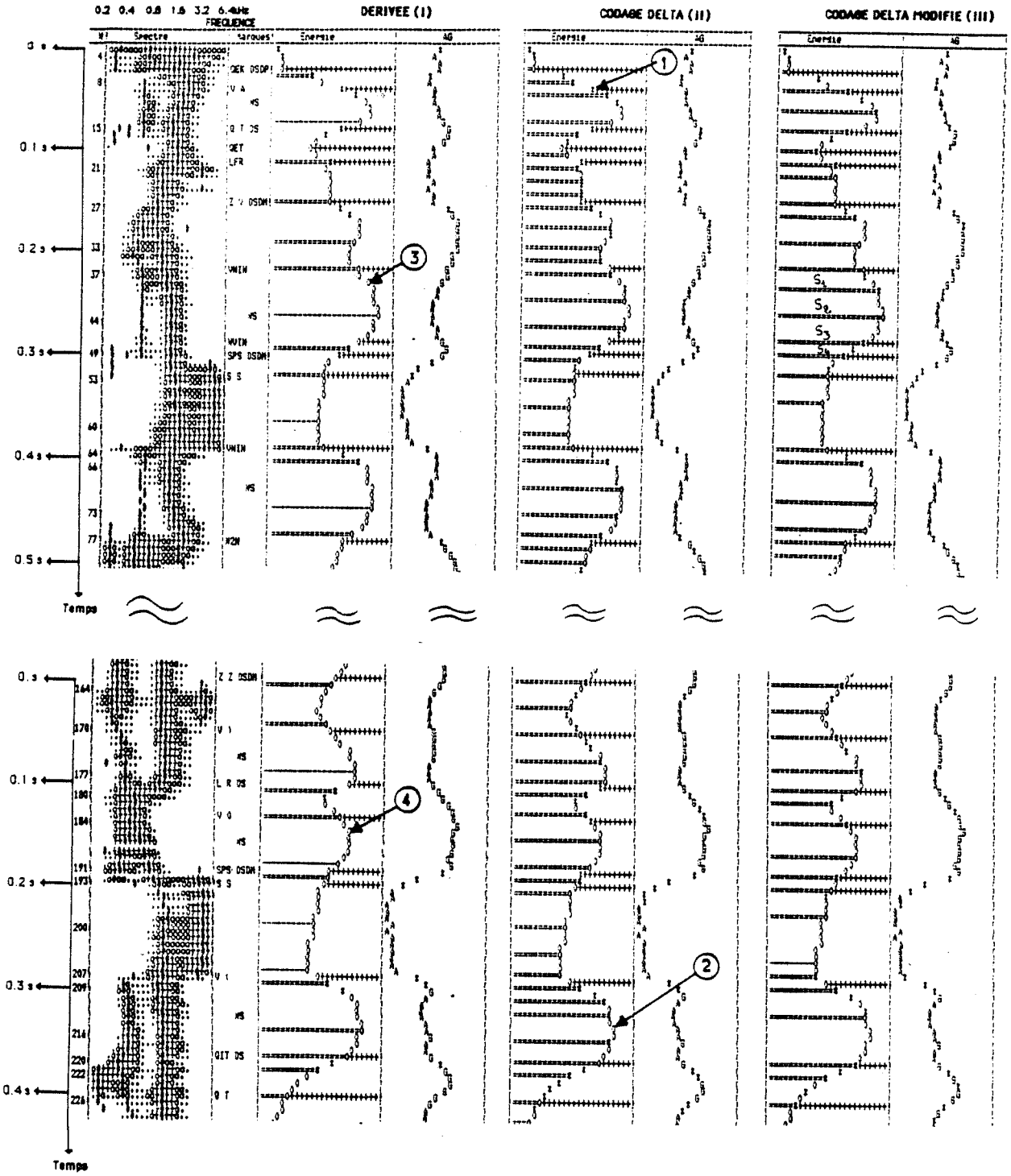


Figure 1 - Segmentation automatique et manuelle de la séquence /katr(ə)vɛ̃sɛ̃ zerosɛt/.

Sucessivement de la gauche vers la droite, on lit:
- L'axe des temps gradué en secondes,
- Le numéro de l'échantillon dans le fichier,
- Le spectrogramme simplifié (bandes d'octave),
- Les marques acoustiques, phonémiques, syllabiques et prosodiques.
- Les variations de l'énergie du signal et de l'indice Aigu-Grave (AG) (2 échantillons sont séparés par 8,8ms) pour chacun des codages mis en oeuvre. Chaque fois que le caractère '*' est imprimé cela signifie que la variation de l'indice a été supérieure ou égale au Δ_i.

Les frontières des événements acoustiques, phonémiques, syllabiques sont indiquées à l'aide du trait horizontal matérialisé par une séquence de '+++++' tandis que les phones sont délimités par une suite de '=======' ou de '-----' si marque arbitraire.

Cette segmentation est effectuée par un phonéticien au cours de laquelle il définit des événements conceptuels (segments phonétiques). Sa tâche consiste donc à aligner les phones (éléments objectifs) sur les unités phonétiques qu'il détecte en fonction de sa propre interprétation des faits en déterminant les événements les plus significatifs et en positionnant les attributs aux frontières des phones. De ce fait, les segments produits et les phones ont des frontières communes.

Chaque fait est défini par un étiquetage "hiérarchisé" (Cf. Colonnes "Marques" de la Figure 1):

- une étiquette acoustique qui se décompose en:
  . un attribut phonétique. Ce macro-trait donne la catégorie phonétique de réalisation.
  . un attribut acoustique ou articulatoire dit de 1ère modalité.
- une étiquette phonémique (nom du phonème),
- une étiquette syllabique: délimiteurs des frontières et du noyau syllabique,
- une étiquette syntactico-prosodique.

Ainsi, il est possible grâce aux frontières communes des phones et des segments d'élaborer des règles de réécriture: de type conceptuelle (I) et objective (II). Par exemple, pour le phonème /Ɛ̃/, codé IN dans notre système de codage, nous avons (Voir figure 1):

(I): IN <--- VN, NV
avec VN Voyelle nasale (partie stationnaire),
NV Nasalité post-vocalique.
(II): IN <--- S_1, S_2, S_3, S_4 d'où
VN <--- S_1, S_2, S_3
NV <--- S_4 .

D'autre part, l'expérience acquise par l'équipe de codage [5], lors de l'étiquetage de corpus nous a amené à:

- normaliser les critères d'étiquetage,
- dresser un inventaire des faits rencontrés et notamment des difficultés et des ambiguïtés possibles,
- confronter les diverses interprétations des "experts".

Ces remarques impliquent que d'un point de vue pratique, une session d'étiquetage doit comprendre une session de codage proprement dite, suivie d'une session de synthèse. A l'issue de cette dernière, un dossier récaptulatif est élaboré ce qui nous permet:

- a) de définir des critères convergents (normalisés) d'étiquetage et
- b) d'y consigner une connaissance phonétique sous forme de règles opératoires utile à une procédure automatique de décodage phonétique.

4. CONCLUSION

A l'opposé de la première segmentation en phones entièrement acoustique, la deuxième produit des faits phonétiques plus abstraits mais en s'appuyant sur les mêmes frontières. La confrontation de ces deux segmentations permet de décrire les faits objectifs (phones) acoustiques en termes de faits conceptuels (unités phonétiques) et réciproquement. Cette réécriture est un atout précieux pour tenter de s'affranchir de la dualité abstraite-concrète du phonème. C'est pourquoi, nous avons opté pour une solution extrême: séparer le plus possible l'aspect acoustique de l'aspect phonétique de la parole en réduisant au maximum les inconvénients de cette dualité.

Ainsi au moyen de ces deux segmentations, il est possible d'obtenir deux sources de connaissance (conceptuelles et objectives) et de les intégrer dans une base de connaissance acoustico- phonétique [6].

5. REFERENCES BIBLIOGRAPHIQUES

[1] **J.E. Shoup**, Phonological Aspects of Speech Recognition, in Trends in Speech Recognition, W. LEA ed., Prentice-Hall 1980.

[2] **J. Caelen**, Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique. Thèse d'état, Toulouse, Juillet 1979.

[3] **G. Perennou**, Segmentation en événements phonétiques et en unités syllabiques, 14ème JEP, Galf-Cnrs, Paris, Juin 1985.

[4] **J. Caelen, G. Caelen-Haumont**, Indices et propriétés dans le projet ARIAL II. Actes du séminaire "Encodage et décodage phonétiques" Galf-Cnrs, Toulouse, Septembre 1981, pp 129-143.

[5] **Article** collectif, Propositions pour la segmentation de la base de données des sons du Gréco, 14ème JEP, Galf-Cnrs, Paris 1985.

[6] **J. Caelen, N. Vigouroux**, Une base acoustique et phonétique hiérarchisée: des faits aux connaissances, Séminaire Franco-Suédois, Grenoble, Avril 1985.

[7] **A. Tassy, L. Miclet**, Quantification vectorielle et reconnaissance de mots multilocuteurs, Séminaire Franco-Suédois, Grenoble Avril 1985.

PROPOSITIONS POUR LA SEGMENTATION ET L'ETIQUETAGE D'UNE BASE DE DONNEES DES SONS DU FRANÇAIS

C. Abry[1], D. Autesserre[2], C. Barrera[3], C. Benoît[1],
L-J. Boë[1], J. Caelen[3], G. Caelen-Haumont[3], M. Rossi[2],
R. Sock[1], N. Vigouroux[3].

[1]. Institut de la Communication Parlée Grenoble
[2]. Institut de Phonétique d'Aix-en-Provence
[3]. C.E.R.F.I.A. Toulouse

Abstract

A set of proposals has been formulated and tested to meet the needs of the French language Database (in segmenting and labeling the speech signal).

Two approaches are to be compared : one in the time domain (manual), the other using spectral cues (semi-automatic).

The first one - after centering phonetic symbols (broad transcription) on the "corresponding" parts of the signal - locates in relevant discontinuities one or more events. Relations or coordinations between events, which are interpretable articulatorily, are hence constructible. Each elementary segment ("time content" between two successive discontinuities or events) is macroclassed, set in a phase and described into attributes.

Automatic segmentation obtained by delta coding of spectral cues (through an ear model) delivers homogeneous or coherent phones.

They are finally compared to manual segmentation and expertized.

INTRODUCTION

La segmentation et l'étiquetage constitueront une phase importante dans l'élaboration de la base de données des sons du français mise en place par le GRECO Parole [2].

C'est de la qualité des opérations qui seront menées à ce stade que dépendra en grande partie la richesse de la consultation ultérieure.

Dans l'état actuel de l'art, il ne semble pas possible de mener toutes ces opérations avec le même degré d'automatisation. Il importe donc de bien cerner quelles seront les approches automatisables et quelle sera la part du manuel, ceci à la fois pour compléter l'automatisation et l'évaluer.

Dans ce but, deux expériences devront être menées de front et confrontées :

- la première se situe dans le domaine temporel : elle opère directement sur l'onde acoustique;
- le seconde procède à partir d'une représentation fréquentielle.

Il nous a semblé en effet que ces domaines d'analyse et de traitement de la parole étaient complémentaires - à la fois au niveau des pratiques et des besoins - en analyse-reconnaissance et en synthèse.

Des résultats de ces deux représentations par rapport à la précision temporelle découlent les affectations suivantes :

- l'approche temporelle est réservée à l'analyse manuelle; elle doit servir de référence.
- Quant à l'analyse spectrale, c'est elle qui contient actuellement le plus d'informations pour fournir ses données à une procédure plus ou moins automatisable.

Différentes dans leurs techniques, ces deux démarches le sont aussi au niveau de leurs exigences de départ en ce qui concerne la segmentation. L'une ne récupère que les discontinuités majeures du signal de parole; l'autre évalue les zones d'homogénéité ou de cohérence spectrales.

Ceci nous amène à préciser l'état des conceptions courantes en ce qui concerne la segmentation de la parole [3] (voir Fig. 1).

L'approche la plus triviale consiste à supposer à ce signal une structure proche de celle de la séquence de symboles correspondant à une transcription phonétique plus ou moins large ("beads conception"). (Fig. 1a).

Une approche plus réaliste consiste à observer des segments de taille infraphonémique ou suprahonémique. (Fig. 1b).

La complexité des interactions dues à la production (coarticulation) amène à une correspondance chevauchante (overlapping) entre ces segments observés et les segments de la représentation phonétique; les frontières de ces recoupements restant le plus souvent floues. (Fig. 1c).

Il reste possible tout de même, avec une représentation phonétique relativement "basse", d'obtenir une projection - sinon de limites, tout au moins de "centres" - pour les réalisations de segments de taille phonémique.

La base de données doit-elle se contenter purement et simplement de ces centres ? Cette étape

- indispensable pour le repérage des réalisations en contexte phonémique dit large* - devra être enrichie d'un repérage plus fin qui tienne compte des connaissances et des exigences actuelles concernant le signal de parole. En effet, si les limites des influences phonémiques sont - comme nous l'avons dit - relativement floues, il reste possible d'établir des segments (infra ou supraphonémiques) à partir des discontinuités que présente le signal.

Dans ces discontinuités nous verrons des événements et nous montrons qu'il est possible de constituer, à partir de ceux-ci des segments minimaux (macroclassés) et/ou des coordinations.

Enfin, ces différentes unités, qui auront été établies par observation des discontinuités, pourront servir de référence pour une comparaison avec les homogénéités délivrées par l'analyse fréquentielle (semi-automatique).

1. SEGMENTATION DANS LE DOMAINE TEMPOREL

1.1. Evénements et coordinations

Il est possible de concevoir certaines discontinuités observables dans le signal de parole, non pas comme des discontinuités globales, mais comme des ruptures dues à l'activité d'au moins une des unités du système articulatoire. Cette conception du signal de parole, qui correspond assez bien à ce qu'on connaît de sa production, amène à une formulation plus ou moins explicite de ce signal comme un produit de canaux (traits, etc.) ([3] voir Fig.1d). Dans ces conditions, il est possible de considérer une rupture acoustique comme étant le produit d'un ou de plusieurs événements. Cette conception permet ainsi - contrairement à la conception en segments, où les discontinuités ne sont là que pour marquer le début ou la fin de cette unité - d'analyser ces discontinuités en événements.

*. Pour l'instant, il est envisagé de repérer ces centres de réalisations par un symbole de transcription phonétique large normative, ce qui correspondra le plus souvent à la requête du consultant. Dans un premier temps, cette transcription fait abstraction des réalisations contextuelles, individuelles, régionales, etc. Seules, les différences en nombre de segments symboliques sont prises en compte (élisions entre parenthèses, épenthèses entre crochets). Toutes "délicates" décisions sur la nature de ces segments (en particulier les timbres vocaliques en position atone) sont ainsi reportées aux repérages fins à partir des discontinuités réellement observables sur le signal. En effet, seul le consultant déjà "au courant" des possibilités de réalisations pourra être amené à accéder directement à leurs éléments (plosions, silences, etc.). La transcription phonétique segmentale comportera des indicateurs prosodiques (accents) et démarcatifs (syllabes, mots, mots-outils, groupes intonatifs). Il est prévu, bien entendu, d'affecter plusieurs segments à un même centre (p. ex. deux voyelles consécutives non segmentables; ou deux consonnes non voisées, la première sans détente, affectées au même silence).

Un exemple bien connu permettra d'illustrer cette approche.

Les nombreuses études consacrées au V.O.T. [5] montrent que ce concept, contrairement aux apparences, n'est pas qu'un segment. En effet, il est possible d'obtenir :

- non seulement un V.O.T. nul, soit une coïncidence entre deux événements (glottique : Voice Onset; supra-glottique : Release).
- mais encore un V.O.T. négatif; dans ce cas, il est clair qu'il ne peut plus s'agir d'un segment mais d'une relation temporelle entre deux événements, l'un des deux servant de référence*.

Une lecture événementielle du signal de parole peut ainsi aboutir à une conception qui prend en compte les délais (avance ou retard) des unités de production les unes par rapport aux autres.

Elle peut aussi aboutir plus généralement à une lecture des résultats acoustiques des coordinations apprises par le sujet parlant pour maîtriser les réalisations de son système linguistique.

Ces coordinations restent encore à construire. Même la plus célèbre d'entre elles que nous venons d'évoquer - le V.O.T. - est toujours sujette à définitions variables (p. ex. [4]). Il n'empêche que la constitution de telles unités est indispensable pour servir à l'intelligibilité de l'organisation temporelle (timing) du signal de parole.

Il semble convenable de proposer une liste d'événements qui correspondent aux discontinuités majeures repérables sur le signal de parole. Ces événements, rappelons-le, ne peuvent avoir d'existence justifiée que par une conception analytique "en canaux". Mais il serait téméraire, dans l'état actuel de l'art, de prétendre à coup sûr identifier dans le signal de parole les origines articulatoires des discontinuités observées. Il importe donc de réserver deux niveaux de lecture : l'un autant que possible strictement acoustique, l'autre doté d'une composante interprétative articulatoire.

Savoir que le repérage des discontinuités acoustiques dépend de nos connaissances articulatoires pour décrypter le signal acoustique, n'est pas s'enfermer dans un cercle vicieux. C'est tout simplement reconnaître que connaître est un processus interactif.

Rappelons d'emblée à ce propos que ce type de lecture du signal de parole présuppose toujours chez le phonéticien une représentation phonétique. En effet, il ne s'agit jamais pour lui de repérer autre chose sur ce signal que la réalisation d'un code dont il est censé connaître les moindres prolongements acoustiques. Sa connaissance s'accroît en effet dans ce domaine dans la limite où ses hypotheses (le signal auquel il s'attend) sont démenties.

*. Autrement dit : une coordination sera une relation temporelle entre événements (binaire, n-aire); un segment est un contenu temporel, une durée entre 2 événements.

Il ne s'agit donc nullement pour lui d'une tâche de reconnaissance visuelle [7].

Notre expérience en lecture du signal de parole nous a conduit à retenir une dizaine d'événements au minimum.

Après ce qui vient d'être dit, on ne s'étonnera pas que ces 10 événements fassent appel à des catégories acoustiques et aussi souvent à des catégories phonétiques de haut niveau comme "vocalique" et "consonantique".

Ces événements sont les suivants* :
- VO : début de voisement (Voice Onset)
- VT : fin de voisement (Voice Termination)
- FO : début de friction, de plosion, d'aspiration (Frication Onset)
- FT : fin de friction, de plosion, d'aspiration (Frication Termination)
- VVO : début vocalique voisé (Vocalic Voiced Onset)
- VVT : fin vocalique voisée (Vocalic Voiced Termination)
- CFO : début consonantique fricatif (Consonantal Fricative Onset)
- CFT : fin consonantique fricative (Consonantal Frication Termination)
- CVO : début consonantique voisé (Consonantal Voiced Onset)
- CVT : fin consonantique voisée (Consonantal Voiced Termination)

Certains de ces événements peuvent être, dans un second temps, interprétés en termes articulatoires. Nous avons retenu deux changements principaux C (closion) et R (détente, Release) correspondant, respectivement, à une augmentation et à une diminution de <u>striction</u> du conduit vocal.

Nous avons besoin, dans le cas du français, d'au moins six degrés de striction**.
-1 pour [ã, ...,õ] (voyelles nasales)
0 pour [i, ...,u,j,w,ɥ] (voyelles et semi-voyelles)
1 pour [m,n,l] (nasales et latérale)
2 pour [v̬] (v français habituellement constrictif non bruissant ou approximant)
3 pour [f,s,ʃ,z,ʒ](constrictives bruissantes***)
4 pour [p,t,k,b,d,g](tenue occlusive des plosives)
/R/ polymorphe, étant dans l'une ou l'autre de ces classes, suivant qu'il est à battements (closions), constrictif, etc.

*. Pour une justification et un exposé plus détaillé de ces événements, cf. ABRY, BENOIT & BOË dans ces mêmes J.E.P.

**. Ce nombre est variable depuis SAUSSURE, au début de ce siècle, jusqu'à LADEFOGED (in [6]).

***. Y compris les plosions (bursts) des occlusives.

Les dix événements acoustico-phonétiques, munis - pour ceux d'entre eux qui s'y prêtent - de leur interprétation articulatoire*, peuvent donner lieu, selon les besoins du consultant de la base de données, à plusieurs syntaxes de lecture, livrant, outre les coordinations susdites :
- des segments minimaux (ou <u>micro-segments</u>) entre deux discontinuités successives;
- des segments non minimaux (ou <u>macro-segments</u>) comprenant plus d'un segment minimal.

Nous nous attacherons, comme il se doit dans un premier temps, à proposer un ensemble de micro-segments.

1.2. Segments et macro-classes

Chaque segment entre deux discontinuités successives peut être étiqueté à plusieurs niveaux (étiquetage hiérarchisé).

D'abord d'après la <u>macro-classe</u> à laquelle il appartient. La macroclasse donne la catégorie phonétique de la réalisation.

Nous utilisons actuellement une dizaine de ces macroclasses** :
- Vo : voyelle orale
- Vn : voyelle nasale
- Cv : consonne vocalique
- Cvl : " " latérale
- Cvn : " " nasale
- CvR : " " de type R
- Cvj : " " glissante [j,ɥ,w]
- Ca : consonne approximante (ex : v français déjà cité)
- Co : consonne occlusive
- Cs : consonne constrictive

La macroclasse peut être bien entendu une catégorie différente de celle à laquelle appartient le symbole phonétique de la transcription large normative affecté au centre de cette réalisation : par exemple, dans [y n œ R e d (ə) m i] "une heure et demie", réalisé [y n œ R ɛ n m i] , où [n] est macroclassé comme consonne nasale mais transcrite [d].

L'appartenance d'un ou de plusieurs segments successifs à une macroclasse ayant été reconnue, il s'agit de spécifier ensuite la <u>phase</u> de la réalisation à laquelle appartient chaque segment.

Nous distinguons, pour les voyelles et les consonnes :
- E : l'établissement (onset),
- T : la tenue,
- Q : la fin ou coda (relâchement ou offset).

Chacune de ces phases peut s'étendre sur plusieurs segments : p. ex., la tenue (T) d'une

*. C et R, indicés par le degré de striction atteint (C_1, C_2, etc.).

**. Pour une définition des classes de ces segmentations, cf. AUTESSERRE et ROSSI dans ces mêmes J.E.P.

voyelle macroclassée nasale (Vn) peut recouvrir les segments suivants : oral (O), nasal (N) et consonantique (C) (voir Fig. 2b).

La nature de chacun de ces derniers doit ainsi être précisée :

a) en attributs phonétiques acoustiques :
- B : bruit de plosion
- F : friction
- H :souffle (aspiration)
- ? : coup de glotte

b) en attributs phonétiques de plus haut niveau :
- V : vocalique
- C : consonantique
- O : oral
- N : nasal

Cet étiquetage est hiérarchisé en 3 niveaux : la macro-classe (ex : Vo), les phases (ex : T), l'attribut (ex : B). L'arborescence (figure 3) est l'illustration de la segmentation de la voyelle nasale [ã] de "endive" de la figure 2. Notons que ne sont pas indiqués, dans le troisième niveau, les attributs implicites : p. ex., le second segment de la tenue de [ã] ne sera pas étiqueté N, puisqu'il est déjà macroclassé Vn. Mais il n'en va pas de même pour le premier segment de cette tenue, qui réalise lui une macro-classe Vn, tout en étant oral (O).

Plus généralement, les étiquettes segmentales ne font double emploi ni avec les symboles phonétiques (centres), ni avec les étiquettes événementielles.

Elles précisent d'une part les réalisations des symboles (cf., plus haut, "une heure et demie").

D'autre part, elles spécifient :

a) la macro-classe des segments en sous-catégorisant les grandes classes de striction retenues pour l'interprétation articulatoire des événements (p. ex. distinction entre nasales et latérales, entre glissantes et voyelles, etc.);

b) la phase dans laquelle se trouve les segments, information qui ne peut être déduite des interprétations C et R des événements qu'avec des règles de coordination;

c) enfin, la nature phonétique des segments, sans faire, là non plus, double emploi. Ainsi, un attribut de voisement ne sera pas utilisé car il peut être déduit des étiquettes événementielles VO et VT. En revanche, les attributs B, F et H ne sont pas redondants avec les événements FO et FT puisqu'ils spécifient différentes qualités phonétiques de bruit.

A ce stade de l'étiquetage, on peut considérer que le signal ne comporte pas toutes les informations dont il pourra être encore muni, notamment celles dérivées des représentations spectrales. Mais les repérages temporels sont déjà suffisamment nombreux pour permettre toute une gamme de consultations.

2. SEGMENTATION DANS LE DOMAINE FREQUENTIEL.

2.1. Segmentation automatique

L'analyse acoustique s'effectue à l'aide d'un modèle d'oreille qui fournit les paramètres spectraux habituels (24 énergies dans la bande 150-6000 Hz) et les paramètres prosodiques : Fo et intensité sur une fenêtre de 8 ms. Sur ce spectre, par des calculs simples de rapport d'énergie dans certains canaux privilégiés [1] on évalue des "indices spectraux" dont les propriétés s'apparentent à celles des indices utilisées en phonétique. Ce sont : aigu/grave, fermé/ouvert, bémolisé/diésé, écarté/compact, doux/strident, continu/discontinu.

Dans un deuxième temps, on calcule les variations temporelles de ces indices, mesurées par codage "delta" (ce qui a l'avantage de détecter de faibles variations à long terme). Ceci rend possible la définition d'un deuxième lot d'"indices contextuels", puisque la variation ainsi codée dépend fortement du passé et donc du contexte antérieur. A ce vecteur d'indices viennent s'ajouter des informations plus classiques comme les formants (crêtes du spectre) qui permettent par exemple de déterminer des degrés de "friction", de "nasalité", etc.

A l'aide des variations des indices spectraux et de l'énergie totale du signal mesurées par codage delta, il est possible de découper le signal en unités homogènes ou cohérentes appelées "phones" (figure 4). Tous les indices sont comptabilisés dans cette opération et leur variation globale est comparée à un seuil décroissant avec le temps de manière à rendre le système plus sensible aux faibles variations à long terme*. Sur chaque phone ainsi constitué, est calculé un vecteur d'indices à partir des échantillons qui le composent.

A l'issue de ces traitements, on obtient, en correspondance avec le signal, un fichier comprenant :
- les échantillons du signal,
- les échantillons spectraux et les paramètres prosodiques (intensité, fréquence du fondamental, durée),
- les vecteurs d'indices et de traits associés à chaque phone.

2.2. L'étiquetage semi-automatique

Les connaissances du phonéticien sont, à ce stade, requises pour une deuxième segmentation : sa tâche est d'aligner les phones délivrés automatiquement sur les segments (conceptuels) qu'il repère sur le signal en fonction de sa propre interprétation des faits. Son travail consiste donc à positionner les étiquettes de phones mais surtout, à sélectionner les segments significatifs. Pour un exemple, voir la figure 5.

3. LES PROBLEMES DE NORMALISATION DES CRITERES.

D'importantes masses de données devant être traitées, une bonne coordination entre les équipes

*. Voir VIGOUROUX & al. dans ces mêmes J.E.P.

de codage est nécessaire pour définir des règles précises dans la vérification de la segmentation automatique par la segmentation manuelle. A l'heure actuelle, trois équipes se sont constituées dans ce but. Leur travail se décompose en deux sessions :

a) la session de codage qui comprend :
   - l'étiquetage d'un corpus,
   - la vérification des fichiers,
   - la constitution du dossier pour b),

b) la session d'expertise.

Cette dernière réunit tous les membres des équipes afin de noter les différentes interprétations, les cas particuliers, et de dresser un récapitulatif des problèmes rencontrés. Cette confrontation est indispensable pour normaliser les critères de sélection des étiquettes, pour affiner le degré d'interprétation des faits phonétiques et rassembler les connaissances sur les phénomènes rencontrés.

Après avoir noté, au début, une certaine dispersion dans l'interprétation des faits, surtout dans le niveau de détail, petit à petit, un consensus s'est dégagé sur les critères de sélection et les règles opératoires, en examinant cas après cas.

Une session d'expertise comprend ainsi :

1. la vérification croisée du codage et l'exament des arguments ayant conduit à la prise de décision pour les cas ambigus : cela conduit à affiner les règles et les critères;
2. un relevé des cas particuliers pour alimenter les connaissances sur les locuteurs ou les effets de contexte;
3. une mise à jour du savoir accumulé sur les observations.

4. CONCLUSION

La segmentation et l'étiquetage du signal de parole nécessitent, non seulement des logiciels adaptés à un codage souple et rapide, mais aussi une grande coordination des équipes. Notre expérience confirme que l'on peut atteindre une bonne homogénéité dans ce codage même s'il se situe à un niveau de détail important. Il nous paraît donc non seulement raisonnable, mais tout à fait enrichissant que le GRECO coordonne une action dans cette direction. Au-delà des données produites pour une base de travail, la confrontation des méthodes et des points de vue ne peut qu'être positive dans le décodage phonétique de la parole.

BIBLIOGRAPHIE

[1] J. Caelen & G. Caelen-Haumont, "Indices et propriétés dans le projet ARIAL II". Actes du Séminaire "Encodage et Décodage Phonétique", GALF-CNRS, Toulouse, pp 129-143, 1981.

[2] R. Carré, R. Descout, M. Eskenazi, J. Mariani & M. Rossi, "The French Language Database : Defining, Planning and Recording a Large Database". Proc. IEEE. ICASSP, San Diego, U.S.A., 1984.

[3] G. Fant, "Speech Sounds and Features". M.I.T., Cambridge, London, 1973.

[4] D.H. Klatt, "Voice Onset Time, Frication and Aspiration in Word-Initial Consonant Clusters", Journal of Speech and Hearing Research 18, pp 686-706, 1975.

[5] L. Lisker & A. Abramson, "A Cross Language Study of Voicing in Initial Stops : Acoustical Measurements", Word 20, pp 384-422, 1964.

[6] K. Williamson, "Multivalued Features for Consonants", Language 53, pp 843-871, 1977.

[7] V. Zue & R. Cole, "Experiments in Spectrogram Reading", Proc. IEEE Inst. Conf. ASSP, Wash. D.C., 1979.

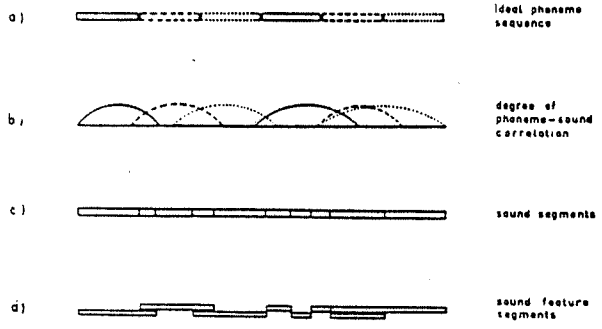


Fig. 1 - Différentes conceptions segmentales d'après [3].

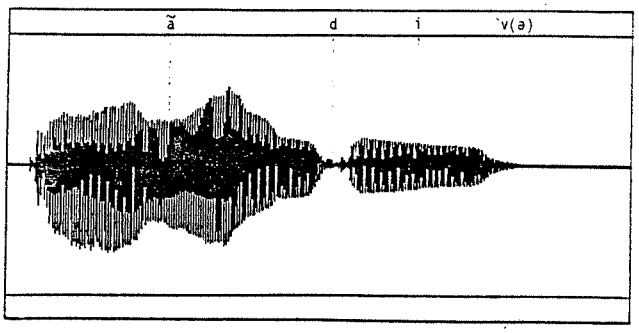


Fig. 2a - Oscillogramme de la réalisation de "endive" dans : "C'est : 'endive' ? Ça ? " (Loc.S.S.). Repérage des centres (pointillés) par les symboles phonétiques (dans le cartouche supérieur).N.B.: l'exemple est choisi à dessein pour se prêter à d'autres interprétations ...

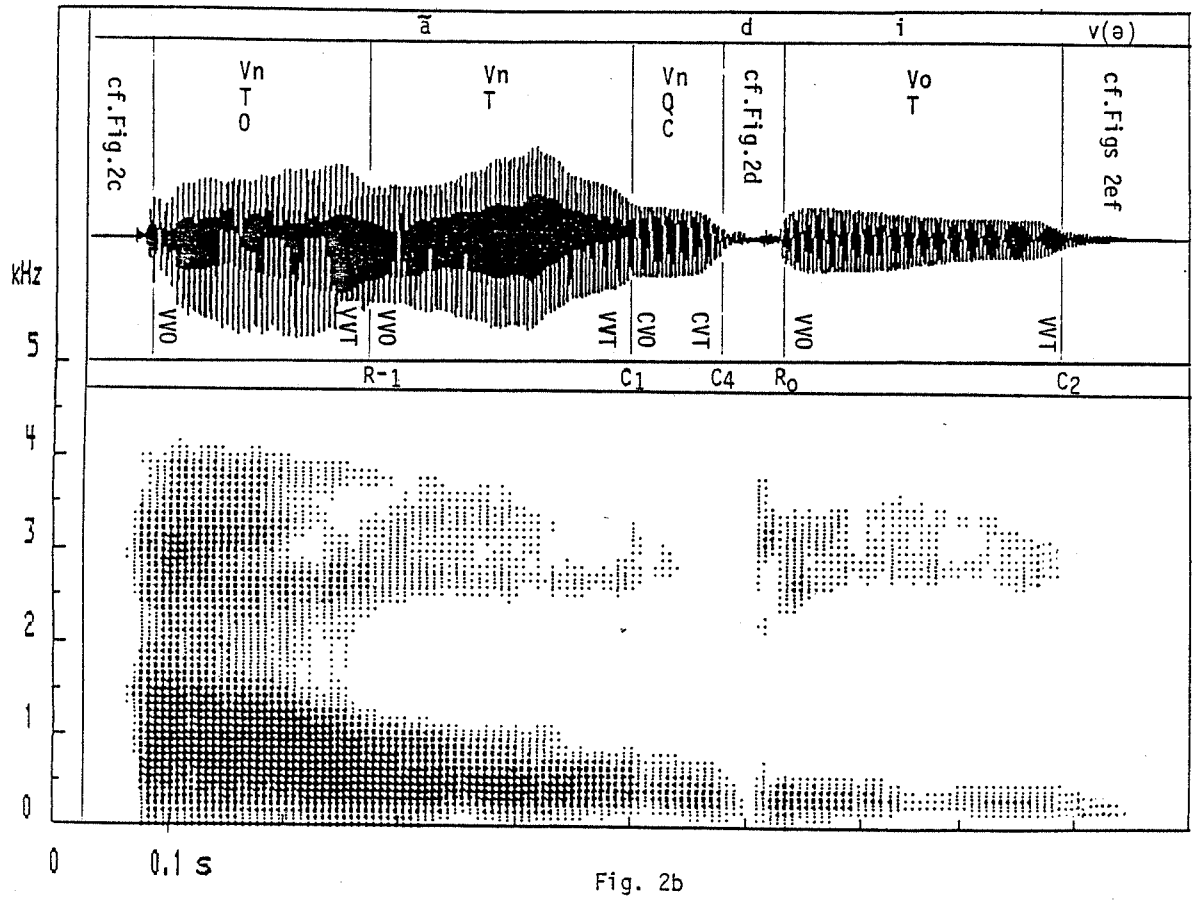


Fig. 2b

Fig. 2b (cf.Figs 2c-f) - Pose des <u>événements</u> (en bas, sur les barres verticales), avec leurs interprétations articulatoires en <u>strictions</u> (cartouche central). En haut (entre les barres verticales et en ordre descendant), les <u>macroclasses</u>, les <u>phases</u> et les <u>attributs</u>. (En aide de lecture, les fengogrammes).

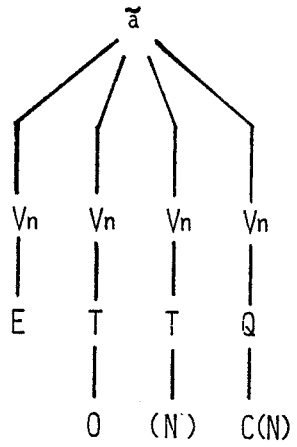


Fig. 3 - Etiquetage <u>segmental</u> de la voyelle [ã] de "endive", en trois niveaux.

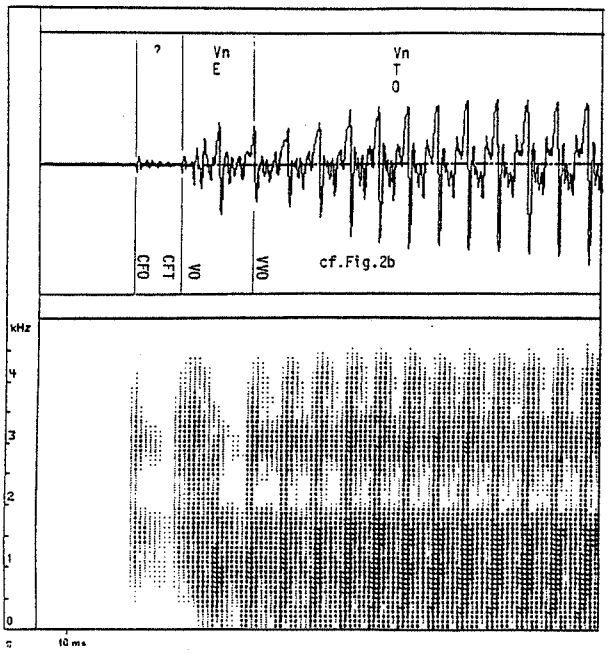


Fig. 2c

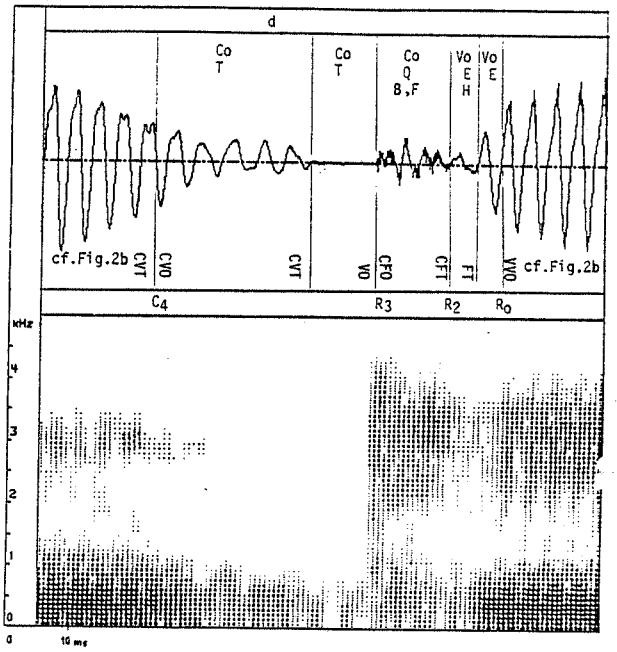


Fig. 2d

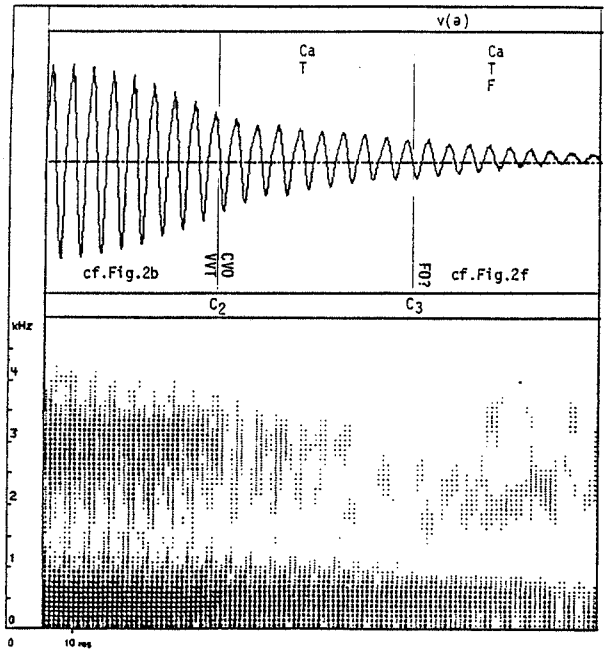


Fig. 2e

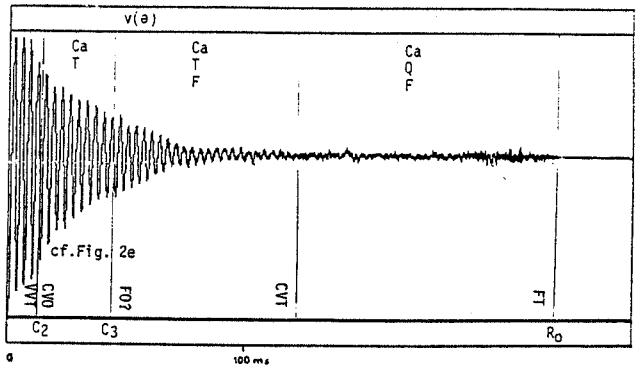


Fig. 2f

Fig. 2c-f - Gros plans sur l'attaque (2c), l'intervocalique (2d) et la finale (2e-f) de "endive" (cf. légende Fig. 2b).

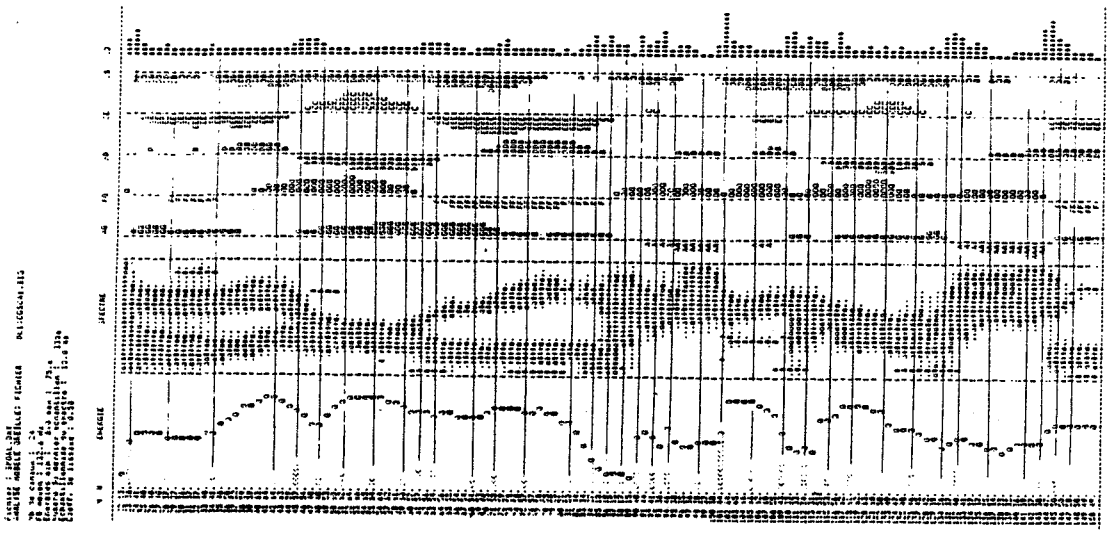


Fig. 4: Segmentation en phones homogènes avec de bas en haut:
la courbe d'énergie échantillonnée à 10 ms et les marques
de segmentation notées par '*' dont le nombre est proportionnel
au degré de discontinuité spectrale,
le sonagramme squelettisé,
les indices représentés sur 9 niveaux: AG, FO, BO, EC, DS,
L'indice CD (continu-discontinu) sous forme d'histogramme.
Les phones sont délimités par deux traits verticaux.

Phrase prononcée: 'zéro huit cinquante six' (08-56).

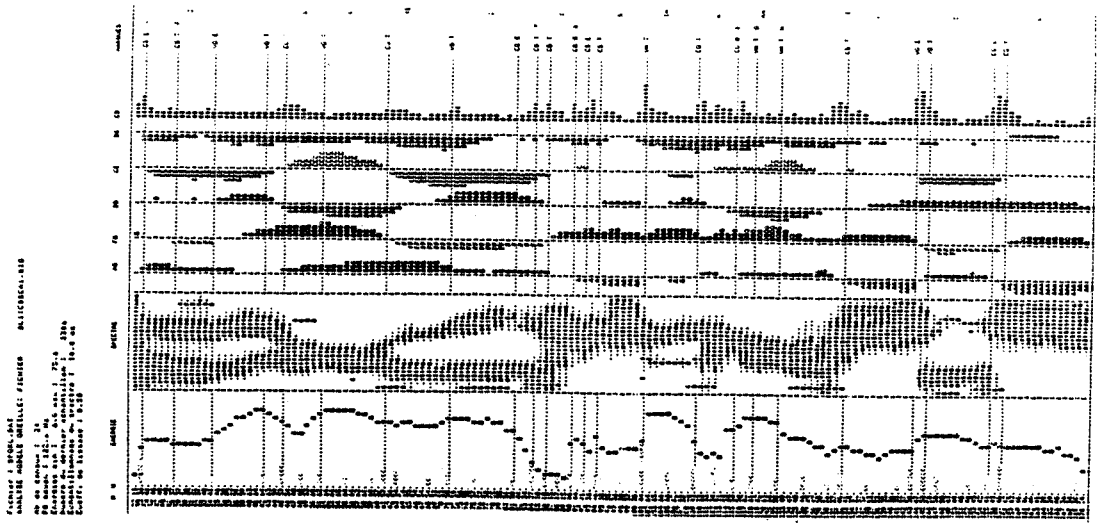


Fig. 5: Etiquetage phonétique de la phrase 'zéro huit cinquante six'.

METHODES DE SEGMENTATION SYLLABIQUE EN RECONNAISSANCE DE LA PAROLE

D. Fohr*, J.P. Haton*, F. Lonchamp**, L. Sauter***

*CRIN - ** Institut de Phonétique de Nancy - *** CGE Marcoussis

ABSTRACT

Segmentation at the acoustic-phonetic level is a major step in continuous speech recognition. We examine such processes in two different contexts : pattern matching based recognition and knowledge based recognition.

Segmentation into syllables seems to be an interesting step, considering the importance of the syllable in the phonation process. Also, the phonetic expert relies heavily on the syllable for his reasoning.

We present two methods for segmenting into syllable-based segments, developped independently in Marcoussis and Nancy. These methods are compared on common data.

INTRODUCTION

Les problèmes de segmentation au niveau acoustico-phonétique constituent un point majeur de la reconnaissance de parole continue. Nous envisageons ici ces processus comme une étape dans le décodage phonétique par des méthodes de comparaison de prototypes d'une part, et de raisonnement fondé sur une base de connaissance d'autre part.

La segmentation syllabique apparaît comme une étape intéressante compte tenu de l'importance de la syllabe dans le processus de phonation. De plus, l'expert phonéticien accorde également une place importante à la syllabe dans son raisonnement.

Nous allons présenter ici deux méthodes de segmentation syllabique développées indépendamment à Marcoussis et à Nancy, avec des finalités complémentaires. Ces méthodes ont été comparées sur des données communes.

SEGMENTATION ET COMPARAISON DE PROTOTYPES

Utilisation de la segmentation pour la comparaison de prototypes

Une des approches de la reconnaissance de la parole utilise des techniques de reconnaissance des formes : elle consiste à comparer des portions plus ou moins longues du signal de parole à des prototypes.

Si les portions sont très courtes, représentant un spectre à court terme, le résultat du traitement sera, par exemple, une suite d'étiquettes de noms de phonèmes. C'est le cas des méthodes dites centisecondes. Un phonème pourra être hypothésé lorsqu'une densité suffisante d'étiquettes identiques sera trouvée. La connaissance d'une segmentation du signal pourra donner des points d'ancrage pour l'émission d'hypothèses. Dans RAPACE [1], une segmentation en demi-syllabes permet d'éviter que plusieurs voyelles soient hypothésées pour un seul noyau vocalique. D'autre part, on tient compte du fait qu'en français, seules certaines suites de consommes peuvent initier ou terminer une demi-syllabe.

On utilise souvent une portion de signal correspondant à l'élocution d'un mot entier : il s'agit alors de la reconnaissance de mots isolés. La programmation dynamique permet de trouver un chemin de distorsion temporelle qui aligne le mot inconnu et les prototypes. Il est connu que cette technique est améliorée lorsqu'on ajoute des contraintes réalistes de pente pour le chemin de distorsion. La réduction de l'espace de recherche qui en résulte diminue à la fois la complexité des calculs et le nombre d'erreurs de reconnaissance en évitant qu'un chemin fantaisiste permette d'aligner deux mots différents.

Si les deux mots à comparer sont segmentés de façon analogue, alors l'espace de recherche peut encore être réduit en ajoutant comme contraintes d'aligner les points correspondants.

Si les segments sont peu sensibles au contexte, comme c'est le cas pour les syllabes et les demi-syllabes, alors il sera possible de remplacer les prototypes de mot par des concaténations de prototypes de segment.

Un certain nombre de systèmes de reconnaissance ont été simulés à partir des considérations précédentes. En ce qui concerne l'approche centiseconde, une segmentation en demi-syllabes a été utilisée dans RAPACE. En reconnaissance de mots isolés, le même algorithme de segmentation a permis d'utiliser un corpus de segments multilocuteurs comme ensemble de référence [2].

Nous décrivons maintenant succintement l'algorithme de segmentation tel qu'il a été utilisé dans ces deux cas :

Algorithme de segmentation

Notre algorithme de segmentation est fondé sur le principe suivant : tout maximum et tout minimum significatif de l'amplitude subjective du signal de parole est un candidat potentiel pour être une frontière entre deux demi-syllabes. Une mesure de l'amplitude subjective peut être obtenue en calculant l'énergie du signal de parole après un filtrage approprié.

Selon ce principe, la parole est segmentée en recherchant des minimums et des maximums successifs de l'énergie du signal. Ceux-ci correspondent respectivement aux groupes de consonnes et aux noyaux vocaliques. Afin d'éliminer des extremums non significatifs, nous utilisons la méthode suivante :

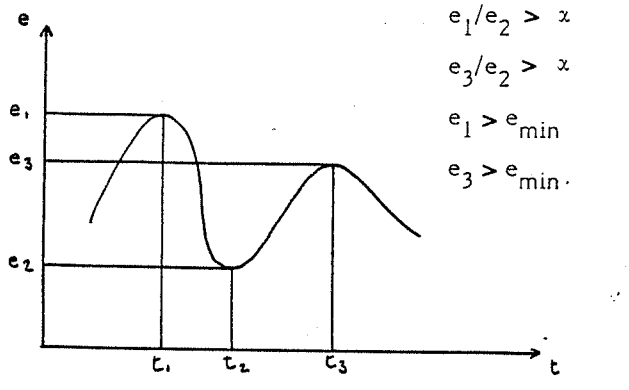
Une suite d'extremums de l'énergie est recherchée vérifiant les contraintes suivantes (cf fig. 1) :

- la longueur de chaque demi-syllabe doit dépasser un temps minimum (de l'ordre de 40 millisecondes)

- le rapport entre les valeurs de l'amplitude au début et à la fin de chaque demi-syllabe doit dépasser une certaine valeur.

- les maximums retenus doivent correspondre à une énergie suffisante.

Figure 1 : Segmentation



contraintes

$t_{i+1} - t_i > \ell min$

$e_1/e_2 > \alpha$

$e_3/e_2 > \alpha$

$e_1 > e_{min}$

$e_3 > e_{min}$.

La fonction utilisée pour mesurer l'énergie est obtenue après filtrage passe-haut du signal de parole (fréquence de coupure : 500 Hz), donnant ainsi une mesure correspondant à l'amplitude subjective. Cette fonction est mesurée toutes les 8 ms puis légèrement lissée.

Résultats

Cet algorithme a été évalué sur un corpus de 10 phrases équilibrées extraits de [3]. (Note : il ne s'agit pas du même corpus que ci-dessous. Des résultats sur le même corpus seront disponibles ultérieurement). Sur ces données, le nombre d'insertions est inférieur à 5 %, le nombre d'omissions est inférieur à 6 %.

SEGMENTATION ET RAISONNEMENT FONDE SUR UNE BASE DE CONNAISSANCE

Démarche de l'expert et fonctionnement général du système

Nous développons depuis deux ans un système pour le décodage acoustico-phonétique de la parole. Notre but est le décodage multilocuteur en parole continue. Pour ce faire, nous utilisons les connaissances d'un expert phonéticien (François Lonchamp) et nous formalisons son expertise pour le système expert SYSTEXP. François Lonchamp est capable, à partir d'une représentation visuelle du signal acoustique de parole (spectrogrammes), de décoder des phrases prononcées de manière naturelle avec un taux de reconnaissance nettement supérieur à celui des algorithmes actuellement disponibles [4].

Quand l'expert phonéticien décode un spectrogramme, il commence par jeter un regard sur l'ensemble du spectrogramme pour calculer la durée moyenne vocalique. Cette notion lui est indispensable pour segmenter les paquets vocaliques (plusieurs voyelles et sonnantes sans frontières marquées) ou les zones qui peuvent être segmentées de plusieurs manières (consonne ou voyelle très longue).

D'autre part, l'expert ne semble pas avoir de problème de segmentation dans les cas non ambigus : il y a séparation entre segmentation et étiquetage : segmentation grâce à des frontières nettes puis étiquetage phonétique en appliquant des règles contextuelles en fonction d'indices pertinents. Dans les cas difficiles (paquets vocaliques) segmentation et identification vont de pair, avec essai de différentes possibilités et choix de celle qui semble la plus probable.

Le système SYSTEXP se fonde sur la démarche de l'expert humain. Dans une première phase, nous effectuons un prétraitement pour obtenir la durée vocalique moyenne et une segmentation grossière en grandes classes phonétiques. Puis, dans une deuxième phase, nous utiliserons un moteur d'inférence et les règles obtenues avec l'expert pour identifier et segmenter finement. Les règles peuvent appeler les procédures de traitement du signal pour extraire les indices nécessaires pour prendre une décision. De plus les règles peuvent modifier la segmentation, c'est-à-dire changer les limites d'un segment, scinder un segment en deux ou regrouper deux segments. La première phase est implantée sous forme procédurale alors que la deuxième est sous forme d'un système expert (cf figure 2).

	Nom	Résultats	Implantation
(1)	NOVOCA	durée vocalique moyenne noyaux vocaliques + limites	procédural
	PLOSI	plosives + limites	procédural
	FRICA	fricatives + limites	procédural
(2)	EXP	phonèmes frontières précises	système expert

(1) Prétraitement
(2) Identification et segmentation

Figure 2 : Synoptique du système SYSTEXP

Fonctionnement du système de segmentation NOVOCA

Le but de Novoca est de trouver tous les noyaux vocaliques contenus dans une phrase et de déterminer la durée moyenne vocalique. L'expert nous a tout d'abord décrit les critères qu'il utilisait pour reconnaître les noyaux vocaliques sur un spectrogramme : forte intensité, présence de formants (pics d'énergie dans le spectre) dans une certaine bande de fréquence ... Nous avons décidé d'utiliser un critère sur l'énergie. Après avoir numérisé (10 kHz 10 bits) et préaccentué le signal de parole, nous calculons tout d'abord l'énergie dans une bande de fréquence. Cette bande a été choisie de manière à défavoriser les sons ayant principalement de l'énergie en très basse fréquence (par exemple les nasales) et ceux qui ont de l'énergie en haute fréquence (par exemple les fricatives). Il fallait inclure dans cette bande la zone du premier et du deuxième formant des voyelles. Nous avons essayé différentes fréquences (250-1500 Hz, 250-2230 Hz, 250-2500 Hz) et la bande retenue a été la bande 250-2350 Hz. Puis nous cherchons les pics de cette courbe qui vérifient les critères ci-après :

- le nouveau pic doit atteindre au moins 55 % du pic précédent (deux noyaux successifs ne peuvent pas avoir des énergies trop différentes),
- la vallée de part et d'autre du pic est fonction de la hauteur du pic (plus un pic est important, plus la vallée doit être importante),
- au moins 50 % des échantillons du noyau vocalique doivent être voisés.

Quand un pic vérifie tous ces critères, on recherche le début et la fin du noyau correspondant. On ne décrira que la recherche de la fin du noyau car la recherche du début est symétrique.

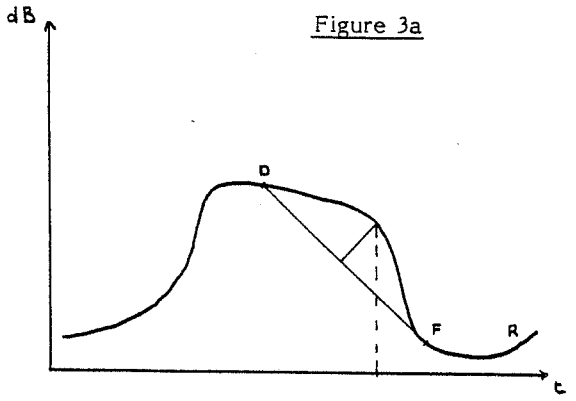
(1) A partir du pic, on recherche le point D où l'énergie est inférieure ou égale à E(PIC)-SEUIL 1 (point de début de la descente).

(2) On recherche ensuite le point R où l'énergie commence à remonter.

(3) A partir de R, on recherche le point F qui vérifie E(F) > E(R)+SEUIL2

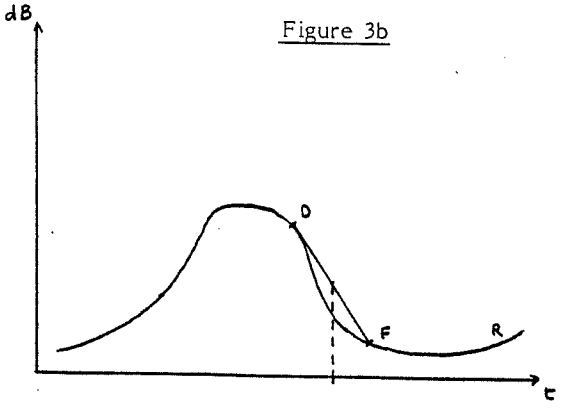
(4) On trace le segment D,F et on cherche le point de la courbe de l'énergie situé au-dessus de cette droite et qui est à la plus grande distance de cette droite. Si le point trouvé est entre D+(F-D)/4 et F-(F-D)/4 c'est le marqueur de fin de noyau sinon c'est (D+F)/2. La figure 3 présente deux exemples.

Figure 3a



Exemple d'une courbe d'énergie présentant une épaule

Figure 3b



Exemple sans épaule

On effectue 2 passes :

une première passe en imposant une vallée importante de part et d'autre du pic (pics ayant une très forte probabilité d'être des noyaux mais taux d'omission de noyaux important).

une deuxième passe en imposant une faible vallée de part et d'autre du pic (taux d'omission très faible mais quelques insertions).

Résultats

Le corpus est constitué de 57 phrases équilibrées de Combescure [3], prononcées à un rythme naturel d'élocution par cinq locuteurs masculins non professionnels. Les phrases étaient lues par un locuteur expérimenté et répétées de mémoire. Le rythme et la prosodie étaient imposés (> 14 phonèmes par seconde). Le corpus comprenait 581 noyaux vocaliques, les résultats sont donnés figure 4.

	Trouvés	Insertion	Omissions
1ère passe	563 (97 %)	18 (3 %)	18 (3 %)
2ème passe	572 (98 %)	26 (< 5%)	9 (< 2%)

Figure 4 : Résultats

Les erreurs sont dues principalement à 5 raisons :

- omission de noyaux faibles situés à côté de consonnes très intenses et insertion de cette consonne, (dans 'toujours' /ʒ/ plus intense que /u/)
- omission de /i/ presque complètement assourdi, en tête ou en queue d'énoncé,
- omission de phonèmes trop brefs (rythme très rapide),
- découpage de voyelles nasales en deux parties (importante fluctuation d'énergie au cours de la voyelle). Ceci est compté comme une insertion,
- insertion d'un noyau supplémentaire dans les logatomes /bR/ /dR/ (remontée d'énergie après la plosive puis chute dans le /R/).
- insertion de quelques consonnes (/l/ /m/ /n/) très vocaliques.

CONCLUSION

Nous avons présenté deux méthodes de segmentation syllabique développées indépendamment avec des finalités différentes. Nous pensons qu'il serait intéressant d'examiner dans quelle mesure leur utilisation conjointe pourrait permettre d'obtenir une segmentation syllabique de qualité encore meilleure.

REFERENCES

[1] L. Sauter, "RAPACE : un système de reconnaissance analytique de parole continue", 4 ème congrès AFCET RFIA, Paris, janv. 1984.

[2] L. Sauter, "Speaker independent isolated word recognition using a segmental approach", Proc. ICASSP 85, Tampa, mars 1985.

[3] P. Combescure, 20 listes de phrases phonétiquement équilibrées, Rev. d'Acoust. n° 56, pp. 34-42, 1981.

[4] N. Carbonnel, D. Fohr, J.P. Haton, F. Lonchamp, J.M. Pierrel, "An Expert System for the Automatic Reading of french Spectrograms", Proc. ICASSP 84, San Diego, mars 1984.

ANALYSE PHONETIQUE ET PHONOLOGIQUE ET SEGMENTATION DU SIGNAL ELECTROMYOGRAPHIQUE

J-F.P.Bonnot[1] et C.Chevrie-Muller[2]

1: Institut de français, université du Maine, 72017 Le Mans
2: Lab. d'explorations neurophysiologiques Inserm U3, Paris

In this paper we discuss some problems linked to the segmentation of the electromyographic signal during speech. We deal first with the question of the experimental method. Then, a number of results obtained from the levator veli palatini, palatopharyngeus and orbicularis oris muscles enable us to shed some light on the inter-subject variability and on the variable power of discrimination of the parameters. Finally, we replace the datas in the framework of the theory of motor equivalence, and we emphasize some linguistic constraints (syllable structure, organization of the phonological rules).

INTRODUCTION

Les problèmes posés par la segmentation du signal EMG ne sont pas limités aux seuls aspects techniques de repérage de points permettant de marquer un début ou une fin d'activité ou de localiser un sommet d'amplitude. Dans un récent article de synthèse, Kent [1] insiste sur un fait capital: la vision que nous avons de l'organisation segmentale est très étroitement liée à la méthode d'analyse. Comment peut-on passer de la description indicielle à la qualification phonétique des traits? Comment ne pas perdre de vue ce qui est l'essentiel, c'est à dire la recherche de l'unité par-delà l'apparente(?)diversité?

1. ANALYSE PHONETIQUE DU SIGNAL EMG

1.1. Décours temporel du signal

Une première série de difficultés est liée à l'approche expérimentale. Suivant le type d'électrodes utilisé et en fonction de la position in situ, un certain nombre de variations peuvent apparaître. O'Dwyer et al.[2] notent que l'amplitude du signal et sa richesse en hautes fréquences diminue à mesure que la distance entre les électrodes et les fibres musculaires augmente. Ils soulignent que, par rapport à des électrodes de surface, plus grosses et plus espacées, des électrodes intramusculaires fines et rapprochées enregistrent des "spikes" ayant une plus grande amplitude, une durée plus brève, et provenant d'une zone beaucoup plus réduite (p.273). Cependant, Bouisset et Maton [3] ont pu mettre en évidence une relation linéaire entre les activités EMG intégrées de surface et intramusculaire. Cette relation est indépendante de la vitesse et de l'inertie opposée au mouvement. Pour ces auteurs, la mesure de l'EMG de surface équivaut à la mesure de l'EMG intramusculaire, multipliée par un coefficient constant (p.294).
D'autre part, O'Dwyer et al.[2], d'accord en cela avec Doran et Baggett [4], Painter [5], etc., remarquent que la description anatomique des muscles pose souvent problème: les désignations standard ne recouvrent pas nécessairement une fonction unique. Ainsi le génioglosse, par exemple, peut agir comme une unité globale ou comme deux, voire trois entités indépendantes (p.284). Dans la même perspective, Künzel [6] note que la production de la parole constitue un cas particulier dans la mesure où les muscles impliqués ont d'abord une fonction végétative, ce qui entraîne un comportement spécifique et un recrutement d'unités motrices généralement limité (p.349). Nos travaux sur l'orbiculaire ont permis de confirmer cette ambivalence fonctionnelle (Bonnot et al.[7,8,9]). Chez certains sujets (JFB) le décours du signal est structuré en fonction des phases articulatoires de fermeture, tenue et ouverture, pour les occlusives. Nous rejoignons ici les remarques de Gentil et Boë [10], de Delaire et al.[11] et de Leanderson et Lindblom [12]: l'orbiculaire des lèvres est actif à la fois dans l'occlusion et dans la protrusion. Nous estimons qu'il peut s'y ajouter une activité de freinage (pour une discussion, cf.Bonnot et al.[7], p.4). Leanderson et Lindblom mettent d'ailleurs en évidence, chez l'un de leurs sujets, une activité de l'orbiculaire supérieur lors de la phase de rupture (phénomène de co-contraction). Certes, le signal est nettement moins ample, et quelquefois même absent. Ces auteurs interprètent les

faits comme étant la marque d'une inhibition réciproque entre muscles antagonistes (p.366). Ces remarques nous amènent à examiner la place qu'il convient de donner à la variabilité et aux caractéristiques semi-idiosyncrasiques dans le cadre d'une théorie de la segmentation du signal.

1.2. Variabilité du comportement.

1.2.1. Au-delà des considérations proprement physiologiques, ces particularités peuvent éclairer les processus d'encodage de la parole. Chez le sujet JFB, par exemple, pour des logatomes ayant la forme CVCVCV, la première bouffée d'activité, ainsi que l'intervalle, possèdent un fort pouvoir discriminant en ce qui concerne l'opposition [b~m]. Au contraire, la seconde bouffée est beaucoup moins liée au contraste phonologique (Bonnot et al.[9]), et peut être considérée comme un "composant fixe" visant à équilibrer le mécanisme de l'ouverture labiale (voir 1.1.).
L'interprétation des indices recueillis au terme de la procédure de segmentation doit donc être conduite avec prudence, afin de distinguer ce qui est porteur d'une information linguistique de ce qui n'est "que" physiologique.

1.2.2. Tout récemment, en mars 1985, nous avons enregistré l'activité EMG du palatopharyngien, de l'élévateur du voile et de l'orbiculaire des lèvres chez deux sujets, DA et JFB (Bonnot et al.en prép.[13]). Le palatopharyngien est assez bien décrit par plusieurs auteurs. Pour Fritzell [14], il a son origine dans les parois postérieure et latérale du pharynx, ainsi que dans le cartilage thyroïde; certaines fibres convergent vers le palais; d'autres (fibres antérieures), orientées verticalement constituent le pilier postérieur. Voir aussi Hirose [15],p.83, Bell-Berti [16],p.227. Néanmoins, les chercheurs ne s'accordent pas pour définir le rôle de ce muscle dans la parole. Pour Fritzell [14], son activité est associée aux sons oraux; elle serait assez mal reproductible (p.33). De plus, des différences inter-locuteurs marquées sont relevées ([14],p.70; Fritzell [17],p. 96). Bell-Berti [16],p.238, pense que la variation des potentiels EMG serait liée, dans une certaine mesure, à la qualité de la voyelle: l'amplitude du signal décroît de [a] à [i], et surtout à [u]. Une première analyse qualitative des tracés, basés sur un corpus de logatomes CVCVCV, où C= [p t k g R] et V= [i u], permet de mettre en évidence une opposition entre DA et JFB. Chez ce dernier, les logatomes [RiRiRi] et [RuRuRu] sont caractérisés par une très forte activité du palatopharyngien, alors que le recrutement est très faible, voire nul, pour l'ensemble des autres occurrences. Par contre, chez DA, le pattern est bien différent: bien que l'on constate une mise en oeuvre pour [RiRiRi], c'est essentiellement lors de la production de [u] que ce muscle entre en action. De ce point de vue, nos résultats ne concordent pas avec ceux de Bell-Berti [16]. Chez nos deux témoins, le palatopharyngien est actif pour des articulations qui peuvent nécessiter un rétrécissement latéral et une élévation du pharynx (rapprochement des piliers postérieurs)(Legent et al.[19]). De plus, la reproductibilité (de 5 à 8 répétitions pour chaque séquence) est excellente.
Il s'ensuit à nouveau qu'un même type de signal, pour lequel des procédures de segmentation identiques peuvent être adoptées, peut véhiculer une information linguistique variable.

1.2.3. Notre troisième illustration ressortit au mécanisme de l'élévateur (fig.1,2). Lors de la réalisation de groupes consonantiques, on observe que le rapport "suppression de l'activité de l'élévateur/durée acoustique" est presque toujours plus élevé pour les groupes où l'occlusive orale précède la nasale: on a -VCorCnasV- vs -VCnas CorV- (ex. [apma-ampa]). Cette particularité n'apparaîtrait pas si l'on n'avait pas mis en rapport trace EMG et signal acoustique. On doit donc souligner que la prise en compte isolée d'une trace graphique n'autorise que des conclusions limitées: une durée absolue de contraction ou de décontraction, si elle n'est pas comparée au signal de parole et, si possible à une autre ou à d'autres activité(s) musculaire(s), ne détient qu'un assez faible pouvoir informatif, au moins pour ce qui regarde l'organisation du timing.

1.2.4. Un autre problème, qu'il faut absolument évoquer, est constitué par la migration de l'activité EMG hors du domaine qui devrait être le sien, si l'on s'en rapporte aux procédures de segmentation du signal oscillographique de parole (Bonnot et Chevrie [19],Bonnot et al.[20]). Ces faits ont une incidence à plusieurs niveaux. Il convient d'abord de réfléchir au rôle des articulateurs, les lèvres et le voile ayant une "indépendance" par rapport à l'organisation sérielle de la chaîne parlée que ne possède pas la langue (Kent [11],p.67-68; Bothorel [21]). Il faut ensuite reconsidérer les modèles phonologiques présentés, certaines interprétations allant jusqu'à remettre en cause la description en traits distinctifs (Moll et al.[22]).

2. INTERPRETATION DES INDICES

2.1. Ces faits, qui montrent la complexité de l'analyse, doivent-ils conduire à abandonner toute tentative d'interprétation linguistique? Certes non! A partir du moment où l'on dispose d'un ensemble suffisant de données coordonnées, il devient possible de proposer un modèle cohérent. Ainsi, Tuller et al.[23] ont montré que la

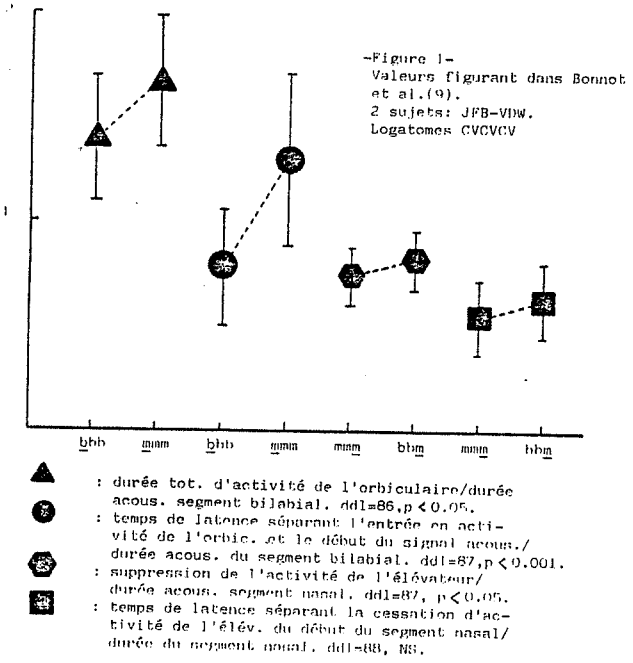
modélisation temporelle des composants musculaires d'une structure était indépendante des variations de la durée et de l'amplitude absolue de l'activité EMG des muscles considérés séparément. Kelso et al.[24],p. 155, notent que la malléabilité du système est due à l'ajustement des valeurs paramétriques non essentielles. Ceci permet de conserver intact le schème d'encodage moteur de base. Voir aussi Gentil et al.[25]. Des stratégies différentes peuvent donc s'expliquer dans le cadre d'une théorie des structures coordonnées. Une grande partie de l'interprétation repose sur les notions d'équivalence motrice et de synchronisation de l'activité musculaire. Nos travaux confirment partiellement cette analyse. Cependant, nous avons mis en lumière une influence non négligeable de la position, un certain degré de désynchronisation pouvant être introduit pour la réalisation de C3 dans C1VC2VC3V par rapport à C1. D'autre part, il est clair que le champ de variabilité diffère suivant les muscles. Enfin, certains indices sont plus opérants en C1, bien qu'un relais puisse être pris par certaines sous-composantes, au niveau de C3 (Bonnot et al.[9]). Cette opposition peut s'expliquer par le fait que la consonne initiale est porteuse d'une information concernant l'ensemble du logatome. Nos résultats, obtenus chez le sujet sain, peuvent être mis en parallèle avec les données de Klich et al.[26], qui ont travaillé sur l'apraxie de la parole. Pour ces auteurs (p.464) l'encodage de la consonne initiale est associé à une grande quantité d'information phonétique: cette complexité peut expliquer le nombre élevé de substitutions de consonnes initiales chez les patients apraxiques.

2.2. On peut d'autre part estimer que le pouvoir de discrimination détenu par les indices est non seulement fonction de la position, mais encore de la nature de la syllabe et du nombre de règles nécessaires à l'obtention de l'output phonique. Mackay [27] a proposé un modèle génératif de structure syllabique, d'où il ressort qu'un ensemble comme CV n'est pas plus complexe que la syllabe monosegmentale V (application de 3 règles) et qu'il l'est moins que VC (5 règles). Le même raisonnement vaut pour CCV (5 règles) vs VCC (7 règles). Ces règles de réécriture sont du type Syllabe(S) →Groupe Cons Init (GCI)+Groupe Voc(GV). Ainsi, VCC [ark] est obtenu comme suit: S→GCI+GV; GCI→∅ (consonne zéro); GV→V+ GCF (Groupe Cons Final); V→[a]; GCF→C1+ C2; C1→[R]; C2→[k]. Par ailleurs, la "single order hypothesis" prédit que les groupes de consonnes initiaux, au contraire des groupes finals, ne peuvent que rarement être réécrits sous la forme C1C2→C2C1. Ceci peut être aisément vérifié pour le français, à partir (par ex.) des données de Rothe [28]. Mackay montre que la vitesse de prononciation est liée au degré de complexité. Enfin, nous avons montré, dans le cadre d'une étude sur l'interlangue français-néerlandais, qu'un même résultat phonique pouvait être atteint après application d'un nombre de règles phonologiques plus ou moins élevé (Bonnot et Spa [29].). Ce processus, mis en évidence chez des apprenants, peut très vraisemblablement caractériser des locuteurs unilingues. Ces stratégies diversifiées peuvent contribuer à expliquer la variabilité des schémas d'activité.

2.3. Il est difficile, sur la base des connaissances actuelles, de proposer une description des traits distinctifs qui soit fondée uniquement sur les indices recueillis lors de la segmentation du signal EMG. Il faudrait d'ailleurs redéfinir les traits, (Rossi [30]; Zerling [31]) qui, trop souvent ne correspondent pas aux données de la production. On peut aussi remarquer que les plus chauds défenseurs de la théorie de l'action évitent la plupart du temps de faire appel à cette notion. Si l'analyse du signal EMG ne permet pas encore de proposer une théorie vraiment globale, elle ne nous en informe pas moins sur beaucoup d'aspects fondamentaux de l'organisation de la parole, tels que timing, contraintes et équilibre des structures musculaires, caractère dynamique de l'encodage.

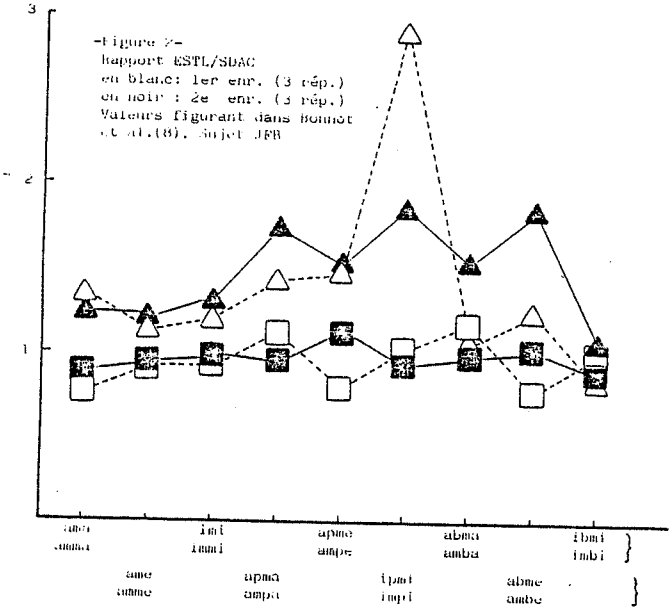
[1] Kent R. "The segmental organization of speech" The Production of Speech, Springer, 57-89, New-York, 1983.
[2] O'Dwyer N., Quinn P. et al."Procedures for verification of electrode placement in EMG studies" JSHR 24, 273-288, 1981.
[3] Bouisset S. et Maton B."Quantitative relationship between surface EMG and intramuscular EMG activity" Am.J.Phys.Medecine, 51/6, 285-295, 1972.
[4] Doran G. et Baggett H. "The genioglossus muscle"Acta Anatomica, 83,403-410,1972.
[5] Painter C. "Pitch control and larynx width in Twi"Phonetica, 33, 334-353, 1976.
[6] Künzel H. "Reproducibility of EMG and velographic measures" J.Phonetics, 6, 345-352, 1978.
[7] Bonnot J-F., Greiner G., Maton B. et Chevrie C."Etude EMG des consonnes simples, géminées,etc."Sém.Int.Labialité,Lannion, 80
[8] Bonnot J-F., Chevrie C., Greiner G. et Maton B."Activité EMG labiale et vélaire" TIPS, 12, 177-224, 1980.
[9] Bonnot J-F., Chevrie C., Greiner G., Guidet C. et Maton B."Eléments pour un modèle temporel d'encodage moteur" TIPS, 16, 1-65, 1984.
[10] Gentil M. et Boë L-J. Les lèvres et la parole, Inst.phonétique Grenoble, 1979.
[11] Delaire J., Fève J. et al."Anatomie et physiologie des muscles des lèvres" Revue de Stomatologie, 2, 93-103, 1977.
[12] Leanderson R. et Lindblom B."Muscle activation for labial speech gestures"Acta Otolaryngologica, 73, 362-373, 1972.
[13] Bonnot J-F., Chevrie C., Maton B.,

Greiner G. et Arabia-Guidet C."Corrélats électromyographiques de la voix chuchotée et de la voix criée" en préparation.
[14] Fritzell B."The velopharyngeal muscles in speech" Acta Oto-Laryng. Suppl.250,1969.
[15] Hirose H."Electromyography of the articulatory muscles"Haskins SR-SR,25/26, 73-86, 1971.
[16] Bell-Berti F."An EMG study of velopharyngeal function"JSHR,19,225-240, 1976.
[17] Fritzell B."Electromyography in the study of the velopharyngeal function"Folia Phoniatrica, 31, 93-102, 1979.
[18] Legent F., Perlemuter L. et Vandenbrouck Cl."Fosses nasales, pharynx", Cahiers d'anatomie,Masson, Paris, 1974.
[19] Bonnot J-F. et Chevrie C."De la validité de quelques modèles de coarticulation" Etudes de phonologie, phonétique et linguistique descriptive du français, vol.1, 1-35, Buske, Hambourg, 1984.
[20] Bonnot J-F., Chevrie C., Greiner G., Guidet C. et Maton B."Coarticulation anticipante et coarticulation rétentrice" 13e J.E.P., 215-216, Bruxelles, 1984.
[21] Bothorel A. "Contraintes physiologiques et indices articulatoires" Speech Communication, 2/2-3, 119-122, 1983.
[22] Moll K., Zimmermann G. et Smith A. "The study of speech production as a human neuromotor system" Dynamic aspects of speech production, 107-129, Tokyo, 1977.
[23] Tuller B. et Kelso J."The timing of articulatory gestures"JASA 76/4,1030-36,84.
[24] Kelso J., Tuller B. et Harris K."Control and coordination of movement" The Production of Speech, 137-173, Springer, New-York, 1983.
[25] Gentil M., Gracco V. et Abbs "Multiple muscle contributions to labial closure during speech" Revue d'Acoustique, Hors série 11-14, 1983.
[26] Klich R., Ireland J. et Weidner W. "Articulatory and phonological aspects of consonant substitutions in apraxia of speech" Cortex 15, 451-470, 1979.
[27] Mackay D. "Aspects of the syntax of behavior: syllable structure and speech rate" Quart.J.of Exp.Psych.26,642-57,1974.
[28] Rothe W. Phonologie des Französischen Grund. der Romanistik, E.Schmidt, 1971.
[29] Bonnot J-F. et Spa J.J. "De la structure théorique de l'interlangue à l'apprentissage du français langue étrangère" IRAL, à paraître.
[30] Rossi M. "Niveaux de l'analyse phonétique: nature et structuration des indices et des traits" Speech Communication 2/2-3, 91-106, 1983.
[31] Zerling J-P. "Phénomènes de nasalité et de nasalisation vocaliques: étude cinéradiographique pour deux locuteurs" TIPS, 16, 241-266, 1984.



-Figure 1-
Valeurs figurant dans Bonnot et al.(9).
2 sujets: JFB-VDW.
Logatomes CVCVCV

▲ : durée tot. d'activité de l'orbiculaire/durée acous. segment bilabial. ddl=86, $p < 0.05$.
● : temps de latence séparant l'entrée en activité de l'orbic. et le début du signal acous./ durée acous. du segment bilabial. ddl=87, $p < 0.001$.
⬣ : suppression de l'activité de l'élévateur/ durée acous. segment nasal. ddl=87, $p < 0.05$.
■ : temps de latence séparant la cessation d'activité de l'élév. du début du segment nasal/ durée du segment nasal. ddl=88, NS.



-Figure 2-
Rapport ESTL/SDAC
en blanc: 1er enr. (3 rép.)
en noir : 2e enr. (3 rép.)
Valeurs figurant dans Bonnot et al.(8). Sujet JFB

Durée totale de suppression d'activité de l'élévateur/ durée acoustique de la nasale (partie nasalisée pour les groupes Chanson ou Corcnas). La figure montre également que le modèle d'activité de l'enregistrement 1 est généralement reproduit lors du second enregistrement. Chaîne parlée.

****
* PROSODIE *
****

DUREE VOCALIQUE, VARIATIONS DE Fo ET PERCEPTION DE FRONTIERES

A. Nicaise'et N.Bacri"

'D.R.L.,Université Paris VII, 2 Place Jussieu, 75005 Paris

"Lab. de Psychologie, 54 Bd Raspail, 75006 Paris

SUMMARY

The influence of vowel duration and fundamental frequency (Fo) changes on the perception of a boundary was studied. Several authors have reported that vowel duration and Fo changes can influence the perception of the voicing feature of stop consonants. Studies in production have also reported on an interaction between durations and Fo mouvements. In exp.1 and 2, we established that vowel duration can cue the distinction between a ..V C≠..and a ..V≠C.. sequence (where ≠ is a syntactic boundary). In exp.3 we established that this distinction is sensitive to the difference between monotone Fo and changing Fo. This may be interpreted in terms of the influence of this difference on the perception of vowel duration. But we could not establish an influence of the amount of Fo change. In Exp.4 we established that a much smaller amount of vowel lengthening is necessary to produce the effect in case the vowel is already lengthened for other reasons.

On a observé que le contour Fo que porte une voyelle influe sur la durée perçue de cette voyelle [5], [4] : Les voyelles avec un mouvement de Fo (quelle que soit la direction et la complexité de ce mouvement) sont perçues comme plus longues que les voyelles avec un Fo monotone. Cette observation a servi de point de départ à plusieurs travaux ([2], [4]) sur l'influence de la durée vocalique et des mouvements de Fo sur la perception du voisement d'une consonne post-vocalique. On se propose ici d'étudier l'influence de ces facteurs dans un autre domaine: la perception d'une frontière.

Une suite V C occlusive peut être rompue par une frontière de mot. Par exemple dans les deux énoncés suivants:

Annick lassait les gens
Annie classait les gens

la suite /ik/ sera analysée comme /ik≠/ dans le premier exemple et comme /i≠k/ dans le deuxième. Ces deux cas sont distingués en production par les durées relatives de la voyelle et de la consonne, à côté de l'utilisation d'un autre indice qui apparaît notamment quand la frontière est aussi une frontière prosodique majeure: la présence d'une pause après la consonne dans le cas VC≠. Une étude rapide de la production de deux locuteurs donne les résultats résumés dans le tableau 1 (On a donné les valeurs extrêmes en ms).

	Annie	Annick
durée de /i/	138/149	89/118
durée de /k/	110/117	155/162

Tableau 1.

Dans le cas "Annick lassait.." une pause ou une zône bruitée de très faible intensité et de durée très variable (30 à 120 ms) suit l'explosion du /k/.

Diverses observations faites en production peuvent être prises en compte dans cette discussion, si on admet qu'un modèle de perception tienne compte des faits de production. On a notamment remarqué ([3]) qu'une consonne subséquente non-voisée, surtout si elle appartient à la même syllabe que la voyelle abrège celle-ci et empêche le mouvement de Fo d'atteindre son terme. On a en outre proposé qu'il existe, en production, des relations entre mouvements de Fo et durée des segments, que les premiers conditionnent les seconds ([6]) ou l'inverse ([1]).

Une première pré-expérience que nous ne ferons que résumer ici démontre:1) que l'échange des durées vocaliques et consonantiques à partir des deux énoncés "Annie classait les gens" et "Annick lassait les gens" permet bien de passer d'une interprétation V≠C à V C≠ et vice versa.

2) que l'allongement de V (passage de VC≠ à V≠C) est plus efficace que sa réduction. On peut expliquer ce fait par l'absence de pause après l'explosion du /k/ pour lever l'ambiguité potentielle dans le cas où on part d'"Annie classait.."

Les résultats de cette première expérience suggèrent donc que le paramètre prépondérant pour le passage de consonne finale de mot ("Annick lassait...") à consonne initiale de mot ("Annie classait...") est la durée vocalique. La modification de la durée vocalique (indice nécessaire) est-elle un indice suffisant? Bien qu'il ne s'agisse pas ici de catégorisation, le passage est-il assimilable au type de passage

qu'on obtient dans les expériences du type "perception catégorielle": Si on fait varier la durée de la voyelle, toutes choses étant égales par ailleurs, peut-on mettre en évidence l'existence d'une frontière? L'expérience 2 examine ces questions.

EXPERIENCE 2

Un enregistrement de l'énoncé "Annick lassait les gens" (durée du /i/: 102 ms, tenue du /k/: 98.5 ms, explosion du /k/ + bruit de faible intensité: 63 ms) a été digitalisé à 10 KHz et la voyelle /i/ allongée par réduplication de périodes dans la partie stable de la voyelle. Le pas d'allongement était de 11 ms. On a ainsi produit 8 stimuli dont le /i/ durait 102, 113, 124, 135, 146, 157, 168 et 179 ms. Ces stimuli ont été présentés en ordre aléatoire à 7 sujets, chaque stimulus apparaîssant deux fois dans le test, par l'intermédiaire d'un casque à un niveau jugé confortable par les sujets. Les sujets devaient décider s'ils avaient entendu "Annick lassait les gens" ou "Annie classait les gens" (choix forcé). La figure 1 donne la proportion de réponses "Annick..." en fonction de l'allongement de la voyelle /i/ (chaque point correspond à un total de 14 observations).

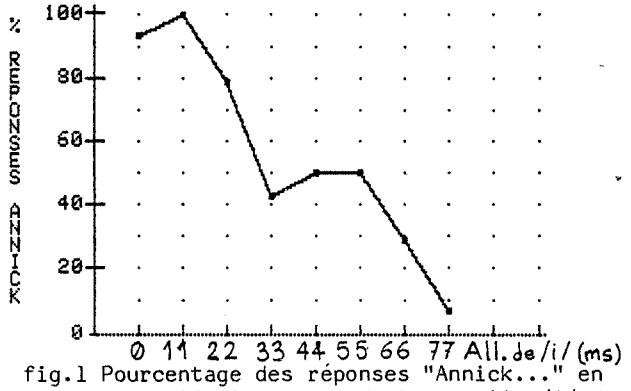


fig.1 Pourcentage des réponses "Annick..." en fonction de l'allongement de la voyelle /i/ de "Annick lassait les gens"

Ces résultats montrent: 1) qu'un allongement de la voyelle supérieur à 65% est nécessaire pour passer d'une interprétation ...V C ≠... à une interprétation ...V ≠ C...

2) que les réponses deviennent rapidement aléatoires (entre 22 et 33% d'allongement). On peut donc penser qu'il existe une large zône pour laquelle les stimuli sont ambigus. Ceci peut s'expliquer par le fait que les autres indices pertinents et notamment les caractéristiques de la consonne ne vont pas dans le même sens que l'allongement vocalique.

EXPERIENCE 3

Les travaux sur l'influence de la durée vocalique et des mouvements de Fo sur la perception du voisement montrent que non seulement la durée vocalique influe sur la perception du voisement mais que cette perception est aussi influencée par la différence Fo constant / Fo changeant sur la voyelle. Les auteurs cités font généralement l'hypothèse que l'effet de Fo est obtenu par l'intermédiaire de son influence sur la durée perçue de la voyelle. Derr et Massaro [2] mentionnent néanmoins qu'on pourrait aussi faire l'hypothèse que le mouvement de Fo est un indice direct de la perception de voisement. On note que toutes ces expériences sont effectuées sur des mots isolés: même si le mouvement de Fo descendant de Gruenenfelder et Pisoni [4] est naturel pour la forme de citation, ce mouvement de Fo n'a pas de rôle prosodique du type: contribution à la segmentation d'un énoncé en groupes prosodiques, ce qui est le cas quand les stimili sont faits à partir d'énoncés naturels. Puisque l'allongement vocalique a un rôle important dans le phénomène qui nous occupe ici (opposition ..V C ≠ , ..V ≠C..) on peut tester le rôle de la différence Fo montant / Fo constant dans la perception de cette opposition.

D'autre part, si, comme Lyberg [6], on admet qu' un mouvement de Fo plus ample entraîne un ajustement de la durée en production, et si on admet que pour les auditeurs qui décodent les énoncés proposés ces faits de production ont une influence sur la perception, la modification de durée de la voyelle /i/ de notre exemple peut être interprétée comme conditionnée par Fo et pas par la présence d'une consonne du même côté de la frontière. Dans cette hypothèse, on s'attend à ce qu'un mouvement de Fo plus ample contrarie le passage de "Annick lassait..." à "Annie classait...". Si, au contraire, les durées une fois fixées, on assiste à un ajustement des mouvements de Fo (Bruce [1]), un mouvement de Fo plus ample ou plus complexe sur la voyelle pourrait entraîner une meilleure perception de l'allongement de la voyelle comme marque de frontière et favoriser le passage de "Annick" à "Annie" ou même être sans effet si l'allongement vocalique est un indice suffisant : le mouvement de Fo plus ample serait simplement compatible avec l'allongement vocalique. Ceci permet de formuler une troisième hypothèse: il y aurait une préprogrammation des durées en fonction de la place des frontières et en fonction de la place des segments dans une structure prosodique. Les mouvements de Fo seraient partiellement autonomes. L'expérience 3 étudie ces problèmes: Existe-t-il une interaction entre durée vocalique et mouvement de Fo dans la perception d'une frontière?

Méthode:

A partir d'un enregistrement de l'énoncé "Annick lassait les gens" digitalisé à 10 KHz, une analyse LPC a été réalisée. Un premier stimulus a été resynthétisé conservant les valeurs de Fo de l'original, puis deux variantes par modification de la Fo de la séquence /ani/. Les caractéristiques de ces trois variantes sont les suivantes:

-variante A (original): Montée de Fo de 166 à 220 Hz sur le /i/ de "Annick". Durée de /i/: 89 ms.
-variante B : Montée de Fo de 166 à 280 hz . Durée de /i/: 89 ms.
-variante C : Contour Fo constant à 152 Hz sur le segment /ani/. Durée de /i/: 87 ms.

Pour ces trois variantes la tenue du /k/ était de 92 ms et l'explosion du /k/ durait 31 ms (durée totale: 123 ms). Un remplacement du bruit de friction (40 ms) suivant l'explosion du /k/ par

un silence s'est avéré nécessaire pour éviter la possibilité d'interpréter les stimuli de départ comme "Annick classait les gens". Ensuite 7 stimuli supplémentaires par variante ont été réalisés en allongeant la voyelle /i/ de "Annick" par réduplication de périodes dans la partie centrale de la voyelle. La Fo étant différente pour chaque variante, cette technique ne permet pas d'obtenir des durées vocaliques strictement identiques. Nous avons fait en sorte qu'elles le soient à 4 ms près.

Durées de la voyelle /i/ en ms:

-variante A:	89	100	111	122	133	144	155	166
-variante B:	89	98	112	121	134	143	157	166
-variante C:	87	100	114	120	134	147	153	167
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*

Ces stimuli ont été présentés en ordre aléatoire en trois blocs de 16 stimuli (un par variante, chaque stimulus apparaissait deux fois) à 28 sujets francophones. La passation du test a eu lieu à l'aide d'un casque en milieu non bruyant. Les stimuli étaient espacés de 7 s. La tâche des sujets était la même que dans l'exp.2. Les résultats sont résumés dans la fig.2.

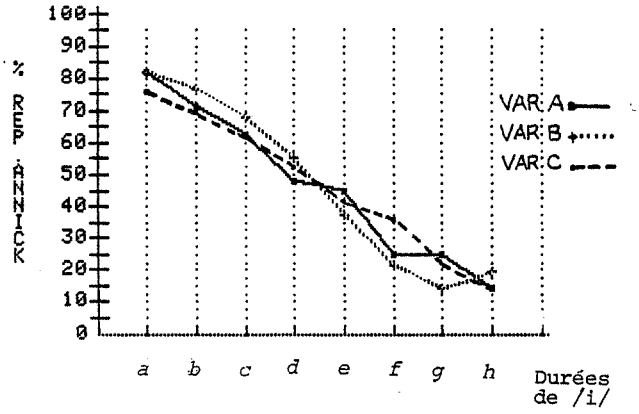


Fig.2. Proportion de réponses "Annick" pour 3 contours Fo en fonction de la durée de /i/

Analyse des résultats et discussion:
Pour les trois configurations Fo, la répartition des réponses "Annick.." varie significativement en fonction de la durée vocalique. Pour des allongements nuls ou inférieurs à 15%, les réponses "Annick.." sont significativement plus nombreuses (z compris entre 4.68 et 2.54 ; $p < .001$ et $p < .01$) que les réponses "Annie..". Lorsque l'allongement dépasse 60% (durée de V= 143/147 ms) ces dernières l'emportent de façon significative (z compris entre -2.27 et -5.21 ; $p < .02$ et $p < .001$). De 25/31% à 50/54% d'allongement, les réponses sont indéterminées. La zône où les réponses sont aléatoires semble relativement étendue mais il faut se souvenir que les allongements choisis sont proches les uns des autres. Ces résultats confirment les résultats de l'expérience 2 pour le cas où Fo n'est pas modifié , montrant une influence négligeable de la nature synthétique des stimuli.

L'effet des variations de Fo sur la distribution des réponses n'apparaît que ponctuellement pour certains triplets de stimuli. L'effet le plus caractéristique apparaît lors de la comparaison des réponses aux stimuli où l'allongement est de 61/69 et 74/76%. Bien que la différence entre les groupes de réponses n'atteigne la significativité statistique que dans le premier cas, on constate dans les deux cas un nombre plus important de "maintien" de la réponse "Annick.." lorsque le contour de montant (166 à 220 hz) devient constant.Les sujets qui pour le contour montant répondaient "Annie.." ont significativement (pour V= 143 ms, $X2= 4.9$ - $p < .02$) tendance à répondre "Annick.." dans le cas du contour monotone. Ce dernier résultat peut être interprété comme une facilitation des réponses "Annie.." lorque le contour est "modulé" et une facilitation des réponses "Annick.." lorsque le contour Fo est constant.

On retiendra de cette analyse des résultats un effet massif de la durée vocalique sur le placement de la frontière mais que cet indice ne devient suffisant que lorque l'allongement dépasse 60%. Bien que la différence que nous observons entre les contours Fo constant et Fo montant soit loin de reproduire l'effet de Fo sur la perception du voisement d'une consonne subséquente, cette différence peut être utilisée par les sujets surtout pour des valeurs "critiques". Ceci pourrait être pris comme une indication indirecte que l'effet de Fo dans la perception du voisement n'est pas obtenu par l'intermédiaire de son influence sur la perception de la durée vocalique mais qu'il s'agit d'un indice direct du voisement. On pourrait aussi proposer que la montée de Fo jouant un rôle linguistique (signal de la fin du groupe nominal sujet) n'est pas toujours prise en compte par les sujets pour la détermination de la durée perçue de la voyelle.

D'autre part, même si l'on admet qu'il existe une relation entre ampleur du mouvement de Fo et durée vocalique en production, on peut au moins dire que ces deux paramètres n'interagissent pas dans la perception de la position de la frontière: un mouvement de Fo plus ample n'empêche pas la perception de l'allongement vocalique comme signalant que la voyelle est finale. Il ne le favorise pas non plus. On peut donc faire l'hypothèse que dans la position considérée le taux de montée Fo est "libre": une forte montée est une montée simplement différente sans qu'on en attende une influence sur la durée perçue de la voyelle portant cette montée. Inversement, une voyelle allongée signale qu'elle n'est pas suivie par une consonne occlusive sourde dans la même unité (syllabe?), le taux de montée étant suffisamment autonome pour supporter de grandes variations sans interagir avec la durée.

EXPERIENCE 4

Certains contours intonatifs comme le contour "d'évidence" (montée+descente sur la syllabe qui porte le centre du contour intonatif suivie d'une mélodie plate sur les syllabes qui suivent le centre du contour intonatif) s'accompagnent d'un fort allongement de la voyelle qui porte le centre du contour intonatif. On peut proposer deux sources pour cet allongement:
- une contrainte de production: il faut plus de temps pour produire ce contour plus complexe

qu'une simple montée.
- une contrainte de haut niveau: parmi les caractéristiques linguistiques de ce contour figure l'allongement de la voyelle qui porte la montée+descente.

Le problème que nous voulons examiner est le suivant: si une voyelle est fortement allongée pour des raisons tenant à la structure prosodique de l'énoncé, quelle proportion d'allongement déclanchera le passage de "Annick lassait.." à "Annie classait.." (déplacement de la frontière)?

Si dans les mêmes conditions, une voyelle n'est pas allongée (mais allongeable?), faudra-t-il la même proportion d'allongement que dans les expériences 2 et 3 (où la frontière ne suit pas immédiatement le centre du contour intonatif) pour déplacer la frontière?

Méthode:

Deux séries de stimuli ont été confectionnées: une série à partir de l'énoncé "Annick lassait les gens" prononcé avec un contour d'évidence sur "Annick", une série à partir du même énoncé prononcé avec un contour descendant d'assertion sur le même mot. Ces deux énoncés avaient l'avantage de permettre de comparer deux cas où le centre du contour intonatif est porté par la syllabe cruciale pour le problème que nous voulons étudier. Dans l'un des cas (contour EVI) la voyelle est fortement allongée (durée de /i/= 170 ms) et porte une montée/descente de Fo (170 - 200- 185 Hz) suivie d'une mélodie plate (105 Hz) alors que dans l'autre (ASS), elle ne l'est que très peu (durée de /i/= 109 ms) et porte une descente de Fo (130-100 Hz) suivie d'une mélodie plate (85 Hz). Ces deux enregistrements ont été digitalisés à 10 KHz. Nos deux stimuli de départ comportaient une pause après le /k/ de "Annick" qui atteignait 118 ms dans un cas et 127 ms dans l'autre. Ces deux pauses ont été supprimées ou plutôt la durée de l'explosion du /k/+léger bruit de friction la suivant a été ramenée à 60 ms (durée comparable à celle des stimuli de l'expérience 3). Ensuite la voyelle de chaque stimulus de départ a été allongée dans sa partie stable en respectant la forme du contour Fo de façon à obtenir les deux séries suivantes:

durée de /i/	all.	Durée résultante	% all.
EVI 170 ms	0 ms	170 ms	0
	6 ms	176 ms	3.52
	12 ms	182 ms	7.05
	18 ms	188 ms	10.58
	24 ms	194 ms	14.11
	29 ms	199 ms	17.05
	35 ms	205 ms	20.58
ASS 109 ms	0 ms	109 ms	0
	23 ms	132 ms	21.10
	38 ms	147 ms	34.86
	54 ms	163 ms	49.54
	69 ms	178 ms	63.30
	84 ms	193 ms	77.06

Ces deux séries de stimuli ont été présentées à 18 sujets (dans les mêmes conditions que ci-dessus) en deux blocs, l'un correspondant à la série EVI et l'autre à la série ASS. Le premier bloc comportait 14 stimuli (chaque durée apparaissait deux fois) et le deuxième 12 stimuli rangés en ordre aléatoire. La tâche des sujets était la même que dans les exp.2 et 3. Les réponses sont résumées dans la fig.3.

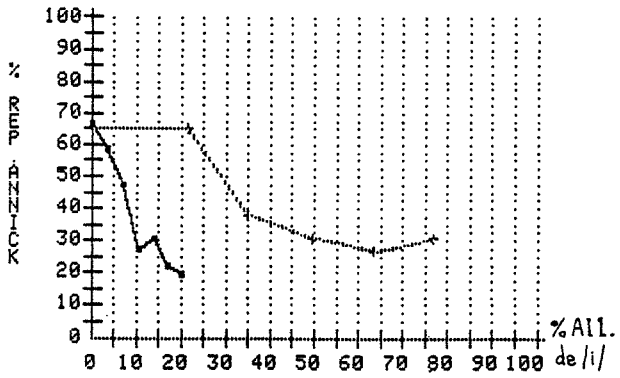


Fig.3.% réponses "Annick.." en fonction du % d'allongement de /i/ (EVI= ASS=)

Analyse des résultats et discussion

Les réponses "Annick.." sont significativement plus nombreuses que les réponses "Annie.." pour l'allongement nul mais ne dépassent pas 66%. Ceci peut être dû au fait que nous avons réduit la pause suivant le /k/ et que celle-ci aurait un rôle non négligeable, et peut-être plus important que dans les exp.2 et 3, dans la levée d'une ambiguité potentielle entre "Annie classait.." et "Annick lassait..".

Dans le cas EVI, un allongement de 10.5% suffit à faire passer les réponses à "Annie classait.."($p < .006$), alors que dans le cas ASS, il faut allonger la voyelle de 49.5% pour obtenir cet effet. Les deux types d'allongement (allongement lié au contour et allongement de voyelle finale) ne se cumulent pas. On peut donc conclure que l'interprétation de la durée vocalique n'est pas indépendante du type de contour qui l'accompagne: l'information sur la durée de la voyelle est confrontée avec l'information issue du traitement de Fo qui fait interpréter cette durée comme "allant avec" un certain contour. Reste à expliquer la très grande sensibilité perceptive des sujets à l'allongement supplémentaire qui a été introduit. On pourrait suggérer qu'on se trouve alors proche de l'allongement maximal admissible pour une voyelle dans la position considérée ou de façon plus plausible qu'il existe une "bonne" durée pour la réalisation du contour qui apparaît dans les stimuli proposés.

Références:(voir [1] et [4] pour une bibliographie)

1.Bruce G. 81,Tonal and Temporal Interplay, Working Papers 21-Lundt University (49-60)

2.Derr M.Massaro D. 80,The Contribution of Vowel Duration, Fo Contour & Fricative Duration as Cues to the /jus//juz/ Distinction, Perception & Psychophysics 27.1 (51-59)

3.DiCristo A. 78,De la Microprosodie à l'Intono-syntaxe, Thèse Doct.Etat, Aix

4.Gruenenfelder T.& Pisoni D. 80,Fo as a Cue to Postvocalic Consonantal Voicing, Perception & Psychophysics 28.6 (514-520)

5.Lehiste I.76,Influence of Fo pattern on the Perception of Duration, J.of Phonetics 4 (113-117)

6.Lyberg B. 79,Final Lengthnening-Partly a Conse--quence of Restrictions on the Speed of Fo Change, J.of Phonetics 7 (186-196)

DUREE VOCALIQUE INTRINSEQUE ET CO-INTRINSEQUE EN FRANCAIS: CONTRAINTES PHYSIOLOGIQUES ET VARIATIONS TEMPORELLES DANS DES SYLLABES C V C

D. Rostolland, C. Parant, A. Takahashi et E.Pandales

Laboratoire de Physiologie Neurosensorielle du C.N.R.S
15, rue de l'Ecole de Médecine 75270 Paris Cedex 06

SUMMARY: Lists of sentences containing nonsense syllables (C1 V C2) were read at normal speed, by 6 male speakers, in quiet and in noise. Timing of segmental average durations were studied (for both C1 and V= 5760 measures) and expressed in terms of vocalic opening (O), vocal effort or noise level (L), number (n) of features of C2 (voicing, manner, place) and individual characteristics. It was found that C1 duration (V.O.T) much decreases with (O) and (L) and that V duration increases with (O), (L) and (n). In spite of wide inter-individual differences, a certain internal cohesiveness of the component parts of the "syllabic" (C1 + V) duration was shown, in terms of (O) and (L). Timing and regulation of sequences are considered in relation to physiological constraints and aero-dynamics of speech production mechanism.

L'étude de la durée relative des segments phonémiques de la parole continue, et en particulier celle des variations intrinsèques et co-intrinsèques de la durée des voyelles, a donné lieu à de nombreux travaux dans plus de trente langues différentes (réf./1/ à /14/). Toutefois, à part la langue anglo-américaine la plus étudiée, assez peu d'investigations concernent la durée vocalique d'une langue donnée. Pour le français, nous avons trouvé seulement quelques études systématiques dans les publications de Durand /1/, Delattre /2/, Straka /15/, Rossi /16 et 17/, Di Cristo /9 et 18/ et Fredet /19/.

La durée vocalique n'est pratiquement jamais un trait distinctif en français contemporain puisque les quelques cas d'opposition cités par Delattre /2/ et Martinet /20 et 21/ tendent à disparaître (par ex. patte-pâte et mettre-maître). Cependant, l'étude de cette variable temporelle est intéressante à plus d'un titre: d'abord la durée conserve une importance perceptive pour la bonne "intelligibilité" de certains énoncés (380 et 381 où la durée du deuxième /ɛ̃/ doit être plus grande; quand, à cause des liaisons, on prononce des voyelles doubles, triples et même quadruples comme le souligne Carton /22/). Ensuite, même en dehors des exceptions ci-dessus, la durée vocalique est très variable en français au niveau acoustique comme au niveau perceptif : le /i/ de "quitte" est beaucoup plus court que le /a/ de "rare" toutes choses égales par ailleurs (même locuteur, même niveau de voix, etc.). Enfin il est clair que la variable temporelle est essentielle dans la production de la parole qui ne peut que se "dérouler dans le temps" et que la question fondamentale est celle des mécanismes de régulation du langage parlé dit "courant". Parler c'est "faire signe" en rendant audibles des gestes invisibles (ou presque, cf. l'influence de la lecture labiale).

La durée d'une voyelle est une grandeur parfois difficile à mesurer et souvent très difficile à interpréter car les variations temporelles sont régies, comme les variations fréquentielles et énergétiques, par des facteurs issus de domaines d'analyse variés : acoustique (bruit ambiant, réverbération), physiologiques (habitudes articulatoires, niveau et débit) et linguistiques (nombre et place des syllabes dans le mot, phonétique combinatoire, place de l'accentuation). En français, toute voyelle accentuée est longue devant les consonnes /v z ʒ R, vR / et les voyelles accentuées suivantes / a o ø, nasales/ sont longues quelle que soit la consonne subséquente. La durée d'une voyelle est liée aussi à son ouverture, à son timbre et à sa "tension". Les voyelles ouvertes sont plus ouvertes (et les voyelles fermées plus fermées) lorsqu'elles sont longues que lorsqu'elles sont brèves (/1,22/). Les voyelles sont d'autant plus tendues qu'elles sont fermées et/ou longues mais, en français, la différence entre phonèmes tendus et phonèmes relâchés est très faibles bien que l'énergie musculaire employée pour la mise en place des organes phonateurs et le maintien de l'articulation pendant une certaine durée soit importante (/1,2,22,31,32,33/).

La présente étude s'inscrit dans le

cadre plus général de l'influence du bruit ambiant (et par conséquent du niveau d'émission de la parole) sur les caractéristiques acoustiques et phonétiques de la voix humaine. Elle fait suite aux études que nous avons commencées dans ce domaine dès 1974. (voir réf. /12,14,23 à 30/).

PROCEDURE EXPERIMENTALE

Des séries de phrases contenant des logatomes /ClVC2/ ont été prononcées à un rythme moyen, mais imposé, dans le silence puis dans un bruit d'intensité croissante. Ces monosyllabes sont placées en position accentuée dans la phrase porteuse suivante: "il a dit ... trois fois". On a choisi pour Cl la consonne /k/, pour V les voyelles /a, i/ et pour C2 les 8 consonnes / pkbg , f/vʒ/ de façon à pouvoir étudier systématiquement l'influence du voisement, du mode et du lieu d'articulation de la consonne finale.

L'enregistrement a été fait dans une cabine audiométrique de 18 m3 (t= 0,2 s.) avec 6 sujets masculins de la région parisienne. Le corpus est prononcé à raison d'une phrase porteuse toutes les 4 s., à trois niveaux sonores: voix normale (N1), voix forte (N2) et voix criée (N3). Ces niveaux de parole correspondent à des niveaux de bruit que le sujet entend au moyen d'écouteurs (bruits "roses" de 50dBA, 80dBC et 85 dBC). Les sujets avaient pour consigne de "lire les phrases le plus régulièrement possible et suffisamment haut pour qu'on puisse les entendre". Au niveau N3 la consigne était de "crier le plus fort possible en mettant l'effort maximum sur la syllabe CVC". (en général, les sujets ont prononcé la phrase porteuse en voix très forte et la syllabe en voix criée).

Les 6 locuteurs ont répété, aux 3 niveaux, 10 fois la même phrase pour chacune des 2 voyelles placées devant l'une des 8 consonnes C2 données.(2880 échantillons pour Cl et autant pour V). A partir du signal de parole non filtré, on a mesuré les 5760 durées phonémiques sur des tracés logarithmiques, sonagraphiques et oscillographiques. La comparaison des trois techniques permet de lever les ambiguïtés dans les cas de segmentations difficiles. La segmentation du signal de parole est depuis longtemps, et continue à être, un problème majeur de l'analyse temporelle de ce signal extrêmement complexe (/34 à 38/).

RESULTATS

Les durées vocaliques que nous avons mesurées varient dans de larges limites en fonction des sujets, des niveaux de voix, des consonnes finales et des voyelles elles mêmes. Les minimums correspondent à /i/ de /kip/ chez les sujets S2, 4 et 5 au niveau N1 (moyennes de 10 valeurs = 65ms, 66ms et 67,5ms) et les maximums à /a/ de /kaʒ/ chez S3 à N1, N2 et N3 (372ms, 364ms et 371,5). Entre les situations extrêmes, les durées moyennes sont dans un rapport de 5,7 : 1 , et en prenant les valeurs absolues de 8,7.

Les durées de Cl varient beaucoup en fonction des sujets, des niveaux de voix et surtout de la voyelle centrale. Le minimum est de 8,5ms (/a/ à N3 chez S4) et le maximum de 86,5ms (/i/ à N1 chez S3). Malgré ces écarts importants -pour Cl comme pour V- on observe des régularités et des compensations remarquables dans plusieurs domaines.

Variations de V

La durée de V dépend d'abord de l'ouverture vocalique. Au niveau 1, /a/ est toujours (48/48 cas) plus long que /i/, chez tous les sujets et pour toutes les C2 : a/i = 42% (tous sujets et toutes consonnes confondus). Suivant les sujets, les pourcentages vont de 37% (S4) à 47% (S2) et suivant les C2, de 28% /ʒ/ à 50% /b et g/. Les % moyens sont peu différents pour /pk bgf∫/, de 46 à 50%, mais inférieurs pour /v/ : 35% et surtout pour /ʒ/ : 28% seulement. Aux niveaux 2 et 3, /a/ est plus long (40/48 cas) ou égal à /i/ (8 cas). C'est chez le même sujet S4 que l'on observe une égalité des durées moyennes de a et i. A N2 et N3, on a presque les mêmes pourcentages sur l'ensemble : a/i = 17,5% et 17,9% (2,4 fois moins qu'à N1). Sans S4 on aurait a/i = 20% à N2 et 21% à N3, soit quand même deux fois moins qu'à N1.

La durée de V dépend ensuite de la nature de C2 : influence du voisement, du mode et du lieu. Au niveau 1, les consonnes /pk/ sont toujours les plus abrégeantes et /vʒ/ les plus allongeantes. On a p<k dans 11/12 cas et v<ʒ dans 12/12 cas. Les voyelles /a/ et /i/ les plus brèves (devant p) durent 126 et 86ms; les plus longues (devant ʒ), 257 et 201ms. Ainsi, la "fricative sonore palatale" et la "plosive sourde labiale" sont dans le rapport de 2,0 : 1 avec /a/ et de 2,3 : 1 avec /i/. Pour les durées intermédiaires, les consonnes /bgf∫/ ont à peu près le même effet d'allongement sur la voyelle précédente, bien qu'elles fassent partie de deux groupes distincts (plosives sonores et fricatives sourdes). On note que les durées dues à /bf/ d'une part et /g∫/ d'autre part sont presques égales, quel que soit la voyelle (les effets du voisement et du mode s'équilibrent, aussi bien pour les labiales que les palatales). L'influence globale du voisement, à N1, est de 35% pour /a/ et de 43% pour /i/. Celle du mode est comparable : 33% pour /a/ et 44% pour /i/ mais celle du lieu est nettement plus faible : 14% et 16% respectivement. Au niveau 2, les voyelles /a/ et /i/ les plus brèves durent 153 et 132ms; les plus longues 252 et 221ms. Les "rapports d'effet" de ʒ/p sont de 1,65 pour a et de 1,68 pour i, plus faibles qu'à N1. Au niveau 3, les voyelles les plus brèves durent 194 et 166ms; les plus longues 280 et 241ms. Les rapports ʒ/p sont plus faibles qu'à N2 : 1,44 et 1,45 respectivement.

C2	N1	N2	N3	N1	N2	N3
/p/	126	153	194	86	132	166
/k/	138	167	203	93	142	170
/b/	165	189	221	110	160	190
/g/	186	206	236	124	174	201
/f/	165	184	224	112	157	187
/ʃ/	183	200	238	123	162	196
/v/	214	219	253	158	188	218
/z/	257	252	280	201	221	241
M :	179	196	231	126	167	196

Tableau 1: Durées vocaliques en ms, pour /a/ (à gauche) et /i/ (à droite) dans des mots C1 V C2. N1=voix parlée, N2=voix forte et N3=voix criée. Moyennes de 6 locuteurs.

Ainsi, dans l'ensemble, le passage de N1 à N3 augmente les durées vocaliques mais réduit les différences d'effets des consonnes subséquentes. Concernant les 8 C2, un autre résultat a été trouvé à N1 (et il semble exister aussi à N2 et N3). En effet, le classement des C2 par ordre d'allongement de /a/ ou de /i/ est différent de l'ordre indiqué au Tableau 1, où nous avions pensé que "les plosives sont (toujours) plus abrégeantes que les fricatives". L'autre hypothèse: "les sourdes sont (toujours) plus abrégeantes que les sonores" se serait avérée également inexacte. En réalité -c'est du moins ce que montrent les présents résultats- si l'on part de la consonne /p/, la plus abrégeante et la moins "marquée", on obtient dans l'ordre: la consonne /k/ marquée par le lieu (le plus faible des 3 traits), puis /fb/ marquées par le mode ou le voisement, puis /ʃg/ marquées en plus par le lieu, puis /v/ par le mode et le voisement et enfin la consonne /z/ par la somme des 3 traits. Une hiérarchie des effets de C2 apparaît donc : "la durée d'une voyelle augmente avec le nombre de traits de la consonne subséquente" proposition peut-être moins vague que celle de Delattre /2/ en terme de force d'articulation consonantique anticipée ?

Une troisième source importante de variation provient des caractéristiques vocales individuelles des 6 sujets. Bien que tous aient prononcé les séries de phrases à raison d'une toutes les 4 s. les rythmes d'élocution, les habitudes articulatoires, n'étaient évidemment pas les mêmes.

	/a/	/i/	/a/	/i/	/a/	/i/
S1	227	157	217	196	201	188
S2	135	92	186	120	299	218
S3	258	182	318	274	317	248
S4	141	103	132	132	184	183
S5	146	101	152	133	210	185
S6	168	121	172	146	176	155
M :	176	126	196	167	231	196

Tableau 2: Valeurs individuelles des durées vocaliques en ms, à N1 (gauche), N2 (au milieu) et N3 (à droite).

Le Tableau 2 montre les moyennes obtenues, toutes consonnes confondues, pour les 6 sujets et les 2 Voyelles. Les différences inter-individuelles sont comprises entre 60% et 141% (S3/S6 à N3 pour /i/ et S3/S4 à N2 pour /a/). A N1, les différences (91 et 98%) sont intermédiaires à celles de N3 (60 et 80%) et de N2 (128 et 141%).

Variations de C1

Les mesures des durées de C1 (VOT de /k/) devant /a/ et /i/ ont donné les résultats du Tableau 3. La durée de C1 dépend de l'ouverture vocalique, du niveau d'effort vocal et, bien sûr, des sujets. Tous sujets confondus, C1 dure 2,30 fois plus devant la voyelle fermée, à N1 (2,27 et 1,66 à N2 et N3). La durée de C1 diminue entre N1 et N2: de -24% pour /a/ et -25% pour /i/. Entre N2 et N3 les diminutions sont de -16% et -38%. Au niveau 3, les différences de VOT tendent vers 0 et le VOT lui-même tend vers 20 ms. Dans l'ensemble, les différences inter sujets aux 3 niveaux sont plus grandes avec /a/ (131, 176 et 187%) qu'avec /i/ (73, 99 et 151%). La durée maximum de C1 est chez S3 avec /i/ à N1 (86,5ms), le minimum étant chez S4 avec /a/ à N3 (8,5ms). Ainsi, les valeurs extrêmes sont dans le rapport de 10,2:1 alors que pour les voyelles on a seulement 3,5:1.

	N1	N2	N3	N1	N2	N3
S1	40	23	23	79	45	32
S2	29	33	23	72	75	28
S3	36	22	18	86	65	54
S4	17	11	8	52	38	21
S5	25	22	19	50	43	25
S6	29	23	21	69	39	29
M :	30	23	19	68	51	32

Tableau 3: Durées de C1 (VOT) devant /a/ à gauche, et devant /i/ à droite. Noter la diminution du VOT en fonction de l'intensité sonore et l'augmentation avec la fermeture de la voyelle.

Stabilité de (C1 + V)

Au niveau N1, les résultats ci-dessus montrent que les durées de C1 et V sont corrélées négativement et que la durée (C1+V) est pratiquement invariante chez un sujet donné. Le VOT(i) dure, en effet, au moins 2 fois plus que le VOT(a), et la voyelle /i/ est plus brève dans des proportions telles que la somme (C1+V) reflète une compensation presque exacte chez S2,S4,S6 et une légère "surcompensation" chez S1, S3 et S5. L'effet de compensation, de régulation temporelle, est bien visible à N2 sauf chez le sujet S4. A N3, on observe une tendance à l'égalisation des durées "syllabiques" chez S1, S5 et S6. Chez S2 et S4 il n'y a pas égalisation, pour deux raisons différentes: chez S2 les VOT sont égaux mais on a /a/ > /i/, et chez S4 les voyelles sont égales et le VOT(i) vaut 2,5 fois le VOT(a).

Chez S3 enfin, on note une surcompensation: a > i et le VOT(i) vaut 2,9 fois le VOT(a). L'ensemble des résultats montre aussi les grandes différences inter-individuelles de (C1+V), malgré l'effet de compensation. Par ailleurs, le "timing" est fortement influencé par le niveau d'émission de la voix : certaines différences de durées, qui étaient négligeables à N1, apparaissent nettement à N2. On trouve aussi une invariance paradoxale de (C1+V) en fonction de l'intensité sonore. Cependant, cette "double invariance" ne se vérifie pas pour les mêmes moyennes : l'invariance de la durée syllabique en fonction de l'intensité n'existe que pour des moyennes inter-sujets, tandis que l'invariance en fonction de l'ouverture vocalique existe, elle, pour des moyennes intra-sujet.

CONCLUSION

Cette étude montre d'abord l'étendue des variations de durées vocaliques et consonantiques avec "seulement" 6 locuteurs et pour 3 niveaux de voix: les durées de C1 sont comprises entre 8ms et 86ms et celles de V (toutes consonnes C2 confondues) entre 92ms et 318ms. Même en voix parlée normale, les variations restent grandes: C1 est compris entre 17ms et 86ms et V entre 92ms et 258ms. Notons que ces variations correspondent à un protocole très limité et que dans la parole courante elles seraient encore plus grandes. Les résultats montrent ensuite la relation entre la durée de V et ce que l'on pourrait appeler la "complexité" articulatoire de C2: plus C2 est marquée par les 3 traits (voisement, mode, lieu) plus la durée vocalique augmente. Enfin l'étude de la durée relative des segments phonémiques, en fonction de l'ouverture vocalique et du niveau sonore de la voix, a montré la relative stabilité de (C1+V) chez un sujet donné. Celle-ci tendrait à prouver que la parole est programmée, sous certaines conditions, en termes d'unités dont la taille serait comprise entre le phonème et la syllabe /39/. Tout se passe comme si le début de la partie audible de C1 (impulsion initiale du bruit d'explosion de la plosive sourde) déclenchait les mécanismes physiologiques nécessaires à l'articulation de la fin de la voyelle et que la durée de celle-ci dépendait, en grande partie, de contraintes articulatoires et/ou aérodynamiques. Si des variations temporelles compensatoires n'existaient pas à cette échelle (C1+V), sans doute la parole "courante" ne le serait plus guère.. paraissant irrégulière pour celui qui parle et celui qui écoute.

Références bibliographiques

/1/ M.Durand,"Voyelles longues et voyelles brèves", Ed. Klincksieck, Paris 1946.
/2/ P.Delattre,"Studies in French and Comparative Phonetics" Ed. Mouton, 1966.
/3/ I.Lehiste,"Suprasegmentals", The M.I.T Press. Cambridge (Mass.), 1970. Ch. 2.
/4/ M.Chen,"Vowel length variation as a function of.."Phonetica,Vol.22.3p.129,1970.
/5/ S.Nooteboom,"Production and perception of vowel.."Philips Res.Lab.N°5, 1972.
/6/ G.Fant,"Auditory analysis and perception of speech",Ses.V. Academic Press, 1975.
/7/ A.Di Cristo,"70 ans de recherches en prosodie",Ed.Univer. de Provence, 1975.
/8/ N.Lass,"Contemporary issues in exp.phonetics", Part 3, Academic Press, 1976.
/9/ A.Di Cristo,"De la microprosodie à l'intonosyntaxe",Thèse Univ.de Provence1978.
/10/F.R.A.N.C.I.S,"Recherche bibliographique sur la durée vocalique",CDSH,Paris,1981.
/11/K.Kohler,"Temporal aspects of speech production.."Phonetica,Vol.38.1-3,1981.
/12/D.Rostolland , "Acoustic Features of Shouted Voice",Acustica V.50.2,p118,1982.
/13/P.MacNeilage,"The production of speech" Part 2, Ed.Springer-Verlag,New Y.,1983.
/14/D.Rostolland,"The influence of voice SL on the duration.."10thICPS,Utrecht,1983.
/15/G.Straka,"Durée et timbre vocalique" Zeitschrift Phon. N°12, p.276-300, 1959.
/16/M.Rossi,"Le seuil différentiel de durée" in Pap.Mem.P.Delattre,Mouton,p.435,1972.
/17/M.Rossi,"L'intonation, de l'acoustique à la sémantique",Klincksieck,Paris,1981.
/18/A.Di Cristo,"La durée intrinsèque des voyelles du Fr.",TIPA N°7 p211,Aix,1980.
/19/F.Fredet,"Etude des caractères intrinsèques des phon..",Thèse Paris III, 1980.
/20/A.Martinet,"La phonologie synchronique et diachro." Trav.ILUP N°6 p.41-58,1938.
/21/A.Martinet,"La linguistique synchronique" Coll.Le linguiste, P.U.F, Paris, 1965.
/22/F.Carton,"Introduction à la phonétique du français", Ed. Bordas, Paris, 1974.
/23/D.Rostolland,"Influence de l'intensité sonore de la voix..",5°JEP,p29 Paris1974.
/24/Rostolland,"Physical analysis of shouted voice",The 8th ICA p240,Londres, 1974.
/25/Rostolland,"Structure acoustique et perception..",15e A.N.P.A, Paris, 1977.
/26/Rostolland,"Influence de l'effort vocal sur les caract.physiques.."GALF, 1978.
/27/Rostolland,"Etude de l'audition de la parole..",Thèse d'Etat, Paris,281p.1979.
/28/Rostolland,"Influence du niv.sonore de la voix sur la durée.."11e.ICA,p275,Paris,83.
/29/E.Pandales,"Variations de durée de la voyelle/a/..",Mémoire Maîtr.50p,Paris 1983.
/30/A.Takahashi,"Relations entre durée et intensité..",Mémoire Maîtr.Paris3,93p;1984.
/31/P.Rousselot,"Principes de phonétique expérimentale",Ed.Didier,Paris, 1925.
/32/R.Jakobson,"Tenseness and Laxness, in Preliminaries.."M.I.T Press,Cambrid.1962.
/33/P.Fouché,"Etat actuel du phonétisme français",in Carton 1974, p.43, 1936.
/34/G.Peterson,"Duration of syllable nuclei in English",JASA N°32-6, p.693-703, 1960.
/35/N.Umeda,"Vowel duration in American English", J.A.S.A. Vol.58-2, p434, 1975.
/36/D.Dutta Majumder,"Some studies on Acoustic Phonetic..",Acustica V.41-2, 1978.
/37/J.Goudailler,"Le passage voyelle-consonne", TIPP Vol.3, p.39-54, Paris, 1980.
/38/G.Mercier,"La segmentation en syllabes et en phonèmes",11e ICA, Toulouse, 1983.
/39/D.Rostolland,5e C.AFCET, Grenoble, 1985.

DIALOGUE HOMME/MACHINE ET ASPECTS TEMPORELS DES STRATEGIES DE LOCUTEURS DE TYPE PHONETIQUE, SYNTAXIQUE ET SEMANTIQUE

G. Caelen-Haumont

Laboratoire C.E.R.F.I.A., Unité Associée 824 du C.N.R.S.,
Université P. Sabatier, 118 Route de Narbonne,
31062 Toulouse Cedex

ABSTRACT

In a preceding phase of our work, within the scope of man/machine dialog in oral quasi-natural language, we adressed the question of defining the speech strategies displayed in a simulated situation, by 10 speakers (5 females, 5 males), as these are given increasingly specific instructions as to message intelligibility and as to kind of receptor (human or machine). As a first step, these strategies were studied from the point of view of time relations at the syntactic level.

The present phase of our work follows two different lines of interest. On the one hand, using a simplified model for signifieds, we perform an analysis of variability among speakers, at the semantic level. On the other hand, the question of a speaker's own choices, made to increase the intelligibility of his/her message, is adressed. In this respect, it is interesting to note, through analysis, the set of priorities given by a speaker to either one or several levels, among the phonetic, the syntactic and the semantic levels.

INTRODUCTION

Dans le dialogue Homme/machine, à composante orale, quelles stratégies choisit l'homme pour clarifier son message et l'adapter aux performances des machines?

L'aspect ergonomique a été peu traité jusqu'à présent dans les études prosodiques. S'agissant du domaine temporel, les analyses sur le français ont porté essentiellement sur les variables [4], sur les facteurs paralinguistiques [8], sur la variabilité du rythme [6] [7], sur la durée et distribution des pauses et des groupes rythmiques [9].

Dans nos études précédentes, nous avons envisagé le problème des relations entre l'homme et la machine sous l'aspect syntaxique et phonétique [1] [2] [3] en tentant de montrer l'aptitude du locuteur à s'adapter naturellement à des consignes de lecture de plus en plus strictes, la dernière spécifiant que l'auditeur est une machine. L'originalité du travail présent repose sur l'introduction d'un paramètre de nature sémantique et sur l'aspect dynamique du choix des fonctions grammaticales que le locuteur opère d'une consigne à une autre.

1. CORPUS, FICHIERS ET DONNEES EXPERIMENTALES

1.1. Le corpus et les consignes

L'étude présente s'inscrivant dans une suite, on a utilisé le même corpus que précédemment [2] [3], composé d'un texte de 45 mots et d'une durée moyenne de 29 secondes. Le texte est un résumé d'un article paru dans la revue Sciences et Vie [10] :

"D'éminents biologistes et d'éminents zoologistes américains ont créé pour des vers géants, un nouveau phyllum dans l'actuelle classification des nombreuses espèces vivantes. Ces longs vers prospèrent sur le plancher marin des zones sous-marines profondes. Des sources thermales chaudes y maintiennent une température moyenne élevée."

Le terme "phyllum" a été explicité avant l'enregistrement. Les 30 lectures différentes de ce texte réalisées par 10 locuteurs (5 masculins, 5 féminins) selon 3 consignes (1/ lecture naturelle et intelligible 2/ lecture très intelligible 3/ lecture très très intelligible pour un ordinateur) se sont effectuées dans une pièce calme, sur magnétophone Radiola N 4420.

1.2. Les niveaux de structuration syntaxique

Un modèle d'analyse simplifiée, utilisé en psycholinguistique [5] et légèrement modifié pour répondre à notre expérimentation, nous a permis de décomposer la phrase en 4 niveaux, selon une perspective grammaticale descendante*, des structures profondes vers les structures superficielles :

- niveau 1 : celui de la phrase
- niveau 2 : celui des constituants majeurs
- niveau 3 : celui des constituants sans subordination ni coordination
- niveau 4 : infrasyntagmatique.

*l'ordre de numérotation des niveaux est inverse: il augmente en fonction d'un découpage plus fin. Ce modèle d'analyse syntaxique simplifiée ne prend pas en compte la nature des syntagmes, ni la fonction, ni l'emplacement dans la structure.

1.3. Les niveaux de structuration sémantique

Nous avons défini sur un plan théorique, un modèle simplifié d'analyse sémantique original, fondé sur une étude de la complexité des signifiés. Nous avons appliqué ce modèle sur l'échantillon restreint de vocabulaire que constitue notre texte et avons jugé, de manière arbitraire, que 4 niveaux étaient suffisants pour rendre compte de son degré de complexité :

- niveau 1 : vocabulaire concret usuel (ex : "ver"),
- niveau 2 : vocabulaire concret courant (ex : "plancher marin"),
- niveau 3 : vocabulaire abstrait usuel ou courant (ex : "classification),
- niveau 4 : vocabulaire concret ou abstrait de spécialité (ex : "phyllum").

L'étude n'a considéré que les groupes de mots délimités par deux pauses, qu'ils soient syntaxiques ou non. Pour l'ensemble des locuteurs et des 3 consignes, ils sont au nombre de 473 et forment donc 473 observations.

Nous devions décider également du niveau de complexité à affecter à un groupe de mots et avions la possibilité de choisir entre deux solutions : elles consistaient à attribuer à ce groupe soit le niveau du dernier mot lexical situé avant la pause, soit un niveau moyen calculé sur les niveaux respectifs des mots qui constituent ce groupe.

Nous avons opté dans un premier temps pour la deuxième solution qui offre l'avantage de respecter la cohésion du groupe; la différence de toutes façons, est peu importante car généralement les groupes forment des sortes d'îlots de complexité dans lesquels les éléments lexicaux ont un degré de complexité très proche. Dans les cas où le groupe se réduit à un mot outil (déterminant ou préposition...), nous avons convenu que le niveau du mot lexical auquel il se rattache syntaxiquement, détermine le sien.

1.4. Fichiers et analyse systématique

Nous utilisons les mêmes fichiers que dans les études antérieures [2] [3], augmentés cependant du paramètre de nature sémantique SM que nous venons de décrire (cf 1.3.).

Dans un bref rappel de la constitution de nos fichiers, nous précisons que les 30 classes correspondent aux réalisations des 3 consignes pour chacun des 10 locuteurs, et comportent en tout 473 groupes devant pause. Ces données déterminent 2 fichiers, le fichier de l'ensemble des groupes (STAC8P : 473 observations et 30 classes de locuteur/consigne) et le fichier des moyennes des parametres par classe (STM08P : 30 observations et 30 classes). Ces observations, quel que soit le fichier, sont analysées par l'intermédiaire de 8 paramètres :

- NI : niveaux syntagmatiques,
- SM : niveaux sémantiques de complexité,
- DP : durée de pause,
- XP : nombre de pauses (pour le fichier moyennes seulement),
- P1 : durée phonémique moyenne (tous phonèmes confondus) dans le groupe devant pause,
- P5 : durée phonémique moyenne (tous phonèmes confondus) en syllabe finale de groupe devant pause,
- NP : nombre de phonèmes.

Seule l'analyse statistique des corrélations a été utilisée dans un premier temps pour cette étude*.

2. RESULTATS

2.1. Présentation formelle

2.1.1. Aperçu grammatical des corrélations

On considère traditionnellement que la grammaire est composée de 3 parties : la morpho-phonologie (recouvrant le domaine de la phonétique), la syntaxe et la sémantique. Nous avons repris cette structuration pour notre étude, et avons assigné à chaque corrélation des paramètres, selon leur nature, une fonction grammaticale. Le tableau 1 ci-dessous en rend compte :

	NI	SM	DP	P1	P5	NP
NI	/	S	S	S	S	S
SM	S	/	Sé	Sé	Sé	Sé
DP	S	Sé	/	P	P	P
P1	S	Sé	P	/	P	P
P5	S	Sé	P	P	/	P
NP	S	Sé	P	P	P	/

TABLEAU n°1 : FONCTIONS GRAMMATICALES DES CORRELATIONS PARAMETRIQUES.

Légende :
S : fonction syntaxique,
Sé : fonction sémantique,
P : fonction phonétique.

2.1.2. Stratégies et perspective globale

Avant de commenter les résultats par locuteur et par consigne, nous considérerons les faits sous l'aspect de regroupements des classes, qu'il s'agisse des corrélations calculées globalement sur les 30 classes, ou sur les 10 que compte chaque consigne. Sauf exception, l'ensemble de l'étude ne s'interessera qu'aux coefficients de corrélation supérieurs ou égaux à 0.50.

Considérant le fichier des 473 observations

*Les logiciels nécessaires à la création des fichiers, leur exploitation par les méthodes statistiques, ont été réalisés et implémentés par J. CAELEN sur MINC-DECLAB 23.

(fichier STAC8P), on note pour la totalité des 30 classes ou chacune des 10 classes, une constante: les coefficients supérieurs ou égaux à 0.50 font référence soit à la fonction syntaxique (DP/NI cf tableau 1) soit à la fonction phonétique (P1/XP, P1/P5, P1/NP). La fonction sémantique n'apparait pas nettement au niveau du regroupement des classes, mais son coefficient de valeur moyenne en consigne 2 (P5/SM : -0.40) laisse présager qu'elle est très dépendante de la consigne et/ou du locuteur.

Le commentaire porte maintenant sur chacun des locuteurs du même fichier, sans distinction des consignes. Le tableau 2 présente une vue d'ensemble des tendances des individus et fournit un premier aperçu de leurs stratégies.

Le nombre de corrélations par fonction grammaticale a été converti en pourcentages : un seul symbole "o" correspond à 20%.

LOCUTEURS		FONCTIONS		
		Phonétique	Syntaxique	Sémantique
FEMININS	UR	oo	o	oo
	PE	oooo	ooo	oooo
	FO	ooo	o	o
	SE	ooo	o	o
	RI	ooo	oo	o
MASCULINS	PA	ooo	oo	o
	CA	oooo	oo	
	PI	oooo	ooo	ooooo
	FA	oo	o	oo
	BU	oo	o	ooo

TABLEAU N°2 : POURCENTAGES DU NOMBRE DE CORRELATIONS UTILISEES PAR LE LOCUTEUR PAR FONCTION GRAMMATICALE.

Légende : o par 20%.

2.1.3. Stratégies et consignes

Ces résultats sont intéressants dans la mesure où ils informent sur la tendance dominante des locuteurs, mais ils ne renseignent pas sur l'aspect dynamique des stratégies adoptées par le locuteur. Les tableaux 3 et 4 en comblant cette lacune, offre une description détaillée des changements de tactique des individus.

Si l'on résume les tendances (cf tableaux n°3 et 4), il apparaît :

1- que les paramètres, quelle que soit la consigne,font constamment référence à la fonction syntaxique.

TABLEAUX N°3 ET 4 : POURCENTAGES DU NOMBRE DE CORRELATIONS UTILISEES PAR LE LOCUTEUR PAR FONCTION GRAMMATICALE ET SELON LES 3 CONSIGNES.

Locuteurs	Consigne	FONCTIONS		
		Phonétique	Syntaxique	Sémantique
UR	1		o	o
	2	o	o	o
	3	o	o	
PE	1	ooo	oo	oo
	2	oooo	ooo	oo
	3		oo	oo
FO	1	o	o	
	2	oo	o	o
	3	ooo	o	
SE	1	o	o	o
	2	oo	o	
	3		o	
RI	1	oo	oo	o
	2	o	o	
	3	oo	o	

TABLEAU N°3 : LOCUTEURS FEMININS

Légende : o par 20%.

Locuteurs	Consigne	FONCTIONS		
		Phonétique	Syntaxique	Sémantique
PA	1		o	o
	2	oo	oo	o
	3	ooo	o	
CA	1	ooo	oo	
	2	o	o	
	3	ooo	o	
PI	1	oo	ooo	oooo
	2	o	oo	o
	3	ooo	o	o
FA	1	o	o	oo
	2	o	o	
	3	o	o	
BU	1	o	o	o
	2	oo	o	oo
	3	o	o	oo

TABLEAU N°4 : LOCUTEURS MASCULINS

Légende : o par 20%.

2- que la fonction phonétique, et plus encore la fonction sémantique sont dépendantes à la fois du locuteur et de la consigne.

3- que les locuteurs recourent de moins en moins fréquemment à la fonction sémantique, plus les exigences d'intelligibilité sont strictes, de la consigne 1 (8 locuteurs sur 10), à la consigne 2 (6 locuteurs) et à la consigne 3 (3 locuteurs).

2.2. Locuteurs et stratégies

Il est interessant de tenter de définir le type de stratégie (phonétique, syntaxique ou sémantique) adoptée par l'individu selon les consignes.

En première analyse, nous pouvons nous contenter de fonder notre critère de décision sur le nombre de corrélations réalisées par le locuteur dans chacune des fonctions grammaticales. Il peut exister des stratégies mixtes, unissant 2 ou 3 fonctions selon que le locuteur opte ou non pour une plus grande structuration de son message. Le tableau n°5 ci-dessous présente une répartition des locuteurs par type de stratégies possibles et par consigne.

TYPES DE STRATEGIES	CONSIGNES 1	2	3
Phonétique	PE,CA	PE,FO,SE	FO,RI,PA, CA,PI
Syntaxique		PI	SE
Sémantique	PI,FA		BU
Phonétique Syntaxique		RI,PA CA,FA	UR,FA
Phonétique Sémantique	FO,RI	BU	
Syntaxique Sémantique	UR,PA		PE
Phonétique Syntaxique Sémantique	SE,BU	UR	

TABLEAU N°5 : REPARTITION DES LOCUTEURS PAR TYPE DE STRATEGIES ET PAR CONSIGNE.

Les conclusions que l'on peut tirer de cette classification sont les suivantes:

- il existe une grande variabilité dans le choix des stratégies : toutes les combinaisons possibles sont exploitées. Cette variabilité existe aux deux niveaux, intralocuteur et interlocuteur.

- Cette variabilité manifeste une bonne adaptabilité des locuteurs aux consignes. En particulier, tous les locuteurs semblent percevoir une différence de compréhension entre l'homme et la machine et développer à cet effet des stratégies spécifiques.

- Au sein d'une dispersion stable de la consigne 1 à la consigne 3, une tendance générale s'établit vers une simplification et une unification des stratégies.

- Cette unification a pour corrélaire un regroupement des locuteurs autour des stratégies phonétique et syntaxique, ou simplement phonétique.

- Le maximum d'intelligibilité en vue de la reconnaissance du message par la machine semble donc être obtenu pour 8 locuteurs sur 10 en négligeant le niveau sémantique.

- La clarté du message est essentiellement assurée par le niveau phonétique (7 locuteurs sur 10).

2.3. Contenu des corrélations

Nous avons vu que la fonction syntaxique était constamment utilisée par le locuteur, qu'il s'agisse de vouloir rendre son message plus intelligible ou mieux adapté à une machine. Cette corrélation qui ne souffre aucune exception est celle qui lie durée de pause et niveaux de structuration syntagmatique (NI/DP toujours inférieur à -0.71 sauf une classe : -0.64). Les pauses sont en effet d'autant plus longues que la structure est profonde.

En règle générale, la fonction grammaticale de la durée de la pause semble être plus naturellement syntaxique (10 locuteurs) que phonétique (7 locuteurs) ou sémantique (3 locuteurs). Inversement, les durées phonémiques moyennes du groupe ou de sa syllabe finale, ainsi que le nombre de ses phonèmes, semblent être caractérisés par des fonctions plus souvent phonétiques (10 locuteurs) que sémantiques (respectivement 5-6-4 locuteurs) ou syntaxiques (2-4-1 locuteurs).

Une remarque interessante concerne par ailleurs les niveaux de complexité sémantique. En effet dans tous les cas (une fois sur trois) où les durées moyennes phonémiques (P1 et P5) ont un coefficient de corrélation supérieur ou égal à 0.50, ces durées moyennes sont d'autant plus longues (ralentissement du débit) que la complexité sémantique du groupe ou du mot est moindre, les groupes sémantiquement les plus complexes n'étant pas systématiquement les plus longs (une fois sur six seulement) ni intégrés dans une structure plus profonde (une fois sur quinze).

CONCLUSION

Nous centrerons nos conclusions sur le fichier STM08P (moyennes de classes). Observant l'effet de la consigne sur les corrélations, nous constatons que la fonction sémantique, bien attestée en consigne 1, s'effondre en consignes 2 et 3, tandis que les corrélations des autres fonctions restent stables.

Si l'on ne retient que les composantes syntaxique et sémantique des fonctions grammaticales de la consigne 1, on en tire les remarques suivantes :

1- le nombre de phonèmes dans le groupe est d'autant plus grand que la structure est profonde (NI/NP : -0.94) et que la complexité sémantique est moindre (NI/SM : -0.58), ce qui traduit une tendance à regrouper structurellement les mots simples.

2- Le nombre de pauses est d'autant plus grand que les structures sont superficielles (XP/NI : 0.88) et que la complexité sémantique s'accroît (XP/SM : 0.55), ce qui signifie que spontanément le locuteur pense faciliter l'interprétation de son message en isolant les mots les plus complexes par des pauses.

Cette stratégie est confirmée par la corrélation s'établissant entre structures profondes et complexité sémantique (NI/SM : 0.62).

3- la durée phonémique du groupe est d'autant plus longue que les structures sont plus superficielles (P1/NP : 0.64): ce phénomène de ralentissement du débit, déjà constaté dans lés études antérieures [3], affecte les mots isolés ou les mots courts.

Si la durée phonémique de la syllabe finale de groupe pour sa part n'accuse qu'un coefficient de corrélation très moyen (P5/SM : -0.33), ce coefficient recouvre en fait des taux supérieurs ou égaux à 0.50 pour 7 locuteurs sur 10.

REFERENCES BIBLIOGRAPHIQUES

[1] G. Caelen-Haumont, G. Perennou, "Utilisation de la prosodie pour la reconnaissance de la parole continue dictée", Actes du Séminaire Prosodie et RAP, GALF-CNRS, Aix-en-Provence, pp 25-57, octobre 1982.

[2] G. Caelen-Haumont,"Rythme et intelligibilité de la parole dictée en français", Actes du 4ème Colloque Langage et Significations, Albi, II, pp 233-245, juillet 1983.

[3] G. Caelen-Haumont, G. Perennou, "Stratégies du locuteur dans le dialogue Homme/machine, Aspects prosodiques", Actes du Symposium Franco-Soviétique, Poutchino, URSS, 1984, à paraître.

[4] J. Grosjean, A. Deschamps, "Analyse des variables temporelles du français spontané", Phonética, vol. 26, 3 , pp 129-156, 1972.

[5] A.R. Lecours, G. Dourdain, J-L. Nespoulous, F. Lhermite, " Vocabulaire de la neuro-linguistique", in "L'Aphasie", Presses Universitaires de Montréal, Flammarion, pp 59-83, 1979.

[6] V. Lucci, "Etude phonostylistique du rythme et de la variabilité de la langue en français parlé et en français lu", Bulletin de l'Institut Phonétique de Grenoble, 2, 139-161, 1973.

[7] V. Lucci, "Rythme et longueur du message parlé. La conversation.", Bulletin de l'Institut de Phonétique de Grenoble, 3, 139-152, 1974.

[8] A. Malécot, R. Johnston, P.A. Kizziar, "Syllabic rate utterance length in French", Phonetica, 26, pp 235-251, 1972.

[9] M. Saint-Bonnet, L-J. Boe, "Les pauses et les groupes rythmiques : leur durée et distribution en fonction de la vitesse d'élocution", Actes des 8èmes JEP, GALF-CNRS, Aix-en-Provence, pp 337-343, mai 1977.

[10] "Les vers géants des Galapagos", Sciences et Vie, juin 1979.

NOUVELLE APPROCHE DANS LE MODELE DE PREDICTION DE LA DUREE SEGMENTALE

K. Bartkova

Centre National d'Etudes des Télécommunications, TSS/RCP
route de Trégastel, 22 300 Lannion

ABSTRACT

The purpose of this study is the prediction of the segmental duration in French continuous speech. In a previous study (Sorin and Bartkova, 1984) we defined a prediction model with fixed coefficients used to modify the inherent duration of vowels and consonants. The value of the coefficients depends on the position of the segment in the word and the sentence. The present study use the same kind of model, but each coefficient is variable in the limits corresponding to the minimum and maximum duration of the segment. This new model allows the simulation of different speaking rates between 3.6 and 7.6 syll/sec and still gives a good prediction for the "neutral" articulation rate. The prediction of segmental duration can be based either on a chosen speaking rate or on the duration of a chosen segment.

I - INTRODUCTION

Dans la parole, le contrôle des phénomènes temporels est essentiel pour préserver l'intelligibilité et le naturel. Aujourd'hui, la connaissance trop limitée des mécanismes de commandes articulatoires ne peut fournir d'informations suffisantes pour prédir les durées intrinsèques et co-intrinsèques des sons. Ainsi les modèles des durées segmentales existants sont, aujourd'hui, des modèles "phénoménologiques", utilisant des mesures faites sur le signal acoustique segmenté.

Le besoin d'un modèle prédictif de la durée segmentale (même si, dans un premier temps, ce n'est qu'un modèle "phénoménologique") s'avère urgent pour la synthèse de la parole et pourrait être très utile pour la reconnaissance de la parole.

En effet, le manque de fluidité de la synthèse de la parole est dû, pour une grande partie, à un traitement inadéquat du rythme et de la durée segmentale. Ceci a été observé en particulier pour la synthèse par diphones du français.

La synthèse articulatoire et la synthèse par formants nécessitent, encore plus impérativement que la synthèse par diphones, la prédiction de la durée segmentale et du rythme de la parole.

II - UN PREMIER MODELE DE LA PREDICTION DE LA DUREE SEGMENTALE

Lors d'une étude antérieure (1,2) sur le "style" de la voix naturelle, en vue de son utilisation dans le système de synthèse du CNET, un modèle linéaire de prédiction de la durée segmentale dans la parole continue en français a été proposé. Ce premier modèle comportait certaines analogies avec le modèle prédictif de la durée en anglais (3). L'originalité de ce modèle résidait dans la tentative d'introduction de quatre vitesses d'élocution en simulant l'élocution de quatre locuteurs de référence. Le modèle peut être résumé par la formule suivante :

$$DUR = DI * ki * kc$$

où

- DI = durée intrinsèque mesurée dans les logatomes, inclus dans les phrases porteuses;
- ki = coefficient introduisant des contraintes liées à l'emplacement du son à l'intérieur de la chaîne sonore;
- kc = coefficient exprimant, pour une voyelle dans une position accentuée, l'influence de la consonne suivante. Dans les autres cas kc=1.

Le défaut essentiel de ce premier modèle réside dans sa "rigidité". A cause des coefficients fixes, utilisés dans le modèle, on perd la possibilité de simuler l'accélération ou le ralentissement de la vitesse d'élocution dans la parole ainsi que la possibilité d'allonger ou de racourcir la durée des sons en fonction de la

longuer des groupes de souffle (la durée des sons étant inversement proportionnelle à la longueur du groupe de souffle).

En outre, la différenciation entre les quatre vitesses d'articulation dans le modèle n'a pas donné les résultats espérés. Pour ces raisons on n'a maintenu que les deux styles extrêmes (lent et rapide) et l'étude des deux autres (moins lent et moins rapide) n'a pas été poursuivie.

III - ELABORATION D'UN NOUVEAU MODELE

Pour pallier les défauts du premier modèle, mentionnés ci-dessus, l'étude d'un nouveau modèle de la durée segmentale a été engagée. Le but espéré consistait à mieux tenir compte des rapports existants entre la longueur des groupes de souffle et la durée des sons (nécéssaire surtout pour la prédiction de la durée segmentale dans les phrases isolées courtes) et atteindre un champ de variation des vitesses d'articulation plus étendu. En effet, malgré l'effort pour simuler deux vitesses d'élocution à partir de deux locuteurs de référence, seule était atteinte une fourchette de 5.1 à 5.5 syll/sec pour le style "lent" et une fourchette allant de 5.6 à 5.9 syll/sec pour le style "rapide" dans les mêmes conditions.

Cette nouvelle étude a conservé certains résultats, obtenus lors de l'étude du premier modèle, comme la durée intrinsèque des sons (DI) et les coefficients exprimant l'influence de la consonne suivante sur la durée de la voyelle en position accentuée (kc).

III.1 Corpus utilisé

Pour déterminer un nouveau jeu de coefficients exprimant les contraintes que l'emplacement du son exerce sur la durée segmentale on a étudé six textes continus de 100 mots ou plus et 30 phrases isolées contenant entre 7 et 10 mots chacune.

Le corpus a été enregistré dans l'élocution "normale" d'un locuteur masculin (le locuteur "lent" du modèle précédent). Le choix du locuteur a été déterminé par ses idéosyncraties individuelles : en cas de changement de vitesse d'articulation, ce locuteur allongeait ou abrégeait systématiquement et ses durées vocaliques et ses durées consonantiques. Le locuteur en question présentait des caractéristiques intéressantes aussi par l'étendue importante de sa plage de vitesses d'articulation allant de 3.2 syll/sec jusqu'à 7.2 syll/sec.

III.2 Règles de positionnement des coefficients

Alors que le modèle précédent ne distinguait que 11 positions de coefficients pour les voyelles et 9 positions pour les consonnes, ce nouveau modèle en distingue respectivement 13 et 10.

L'application des coefficients (ki) est conditionnée :

- par la nature du mot (mot lexical, mot grammatical);
- par la position de la syllabe dans le mot lexical (première syllabe, syllabe médiane, syllabe finale);

Pour les voyelles en syllabe finale on distingue de plus la présence ou l'absence d'une frontière syntaxique ou d'une frontière de phrase, et en cas de pause on prend en compte la longueur de celle-ci.

Pour les consonnes on distingue la présence ou l'absence d'un agrégat consonantique et on tient compte de la structure syllabique au voisinage des pauses.

III.3 Nouveaux éléments du modèle

A) Durée minimale - durée maximale

Pour obtenir une plus grande souplesse dans le modèle de la prédiction automatique de la durée des sons et pour atteindre une plus grande variation de la vitesse d'articulation, on a recherché, lors de l'analyse du corpus, des bornes (durées extrêmes) pour chaque position citées ci-dessus :

$$D_{min} \leqslant D \leqslant D_{max}$$

On a pu observer, dans le corpus étudié, une dispersion de la durée segmentale plus ou moins importante selon la localisation des sons à l'intérieur de la chaîne sonore. Ainsi, par exemple, le voisinage d'une pause entraînait une grande variation de la durée des sons (ex: fricative (+voisée) en syllabe fermée, suivie d'une pause, variait de 109% de sa DI) alors que les mots grammaticaux et les syllabes médianes des mots lexicaux comportaient des sons d'une durée beaucoup plus constante (ex: fricative (-voisée) en mot grammatical variait de 0.5% de sa DI). D'autre part, la dispersion de la durée était correlée aux caractéristiques acoustiques et articulatoires des sons : les voyelles, les fricatives voisées, les liquides et les semi-voyelles variaient beaucoup par rapport à leur durée intrinsèque, alors que les nasales et les occlusives se sont montrées les plus stables dans leurs durées.

B) Coefficients variables

La recherche des durées extrêmes a permis d'obtenir une plage de manoeuvre pour les coefficients (ki) qui prennent ici des valeurs variables entre deux bornes :

$$ki_{min} \leqslant ki \leqslant ki_{max}$$

IV - UTILISATION DU NOUVEAU MODELE

On peut envisager trois façons d'utiliser ce modèle linéaire de la prédiction de la durée segmentale :

a) en choisissant au départ la durée d'un segment, on estime la durée des autres segments par interpolation :

$$ki = ki_{min} + (kj-kj_{min}) * \frac{(ki_{max}-ki_{min})}{(kj_{max}-kj_{min})}$$

où ki est le coefficient lié au segment courant, et kj le coefficient de segment précédent;

puis:

$$D = DI * ki * kc$$

avec DI = durée intrinsèque, kc = coefficient de l'influence consonantique sur les voyelles en position accentuée.

b) en choisissant au départ une vitesse entre les deux valeurs extrêmes (de 3.6 syll/sec à 7.6 syll/sec), on fait correspondre à cette vitesse une durée segmentale dans une position donnée à partir des corrélations calculées. Ensuite, on calcule les durées des autres segments à partir de cette durée en appliquant la méthode décrite ci-dessus;

c) en choisissant une vitesse d'articulation pour estimer la durée des consonnes et une autre pour estimer la durée des voyelles, on arrive, en utilisant ce nouveau modèle, à "mixer" les styles d'élocution différents. Ainsi, en choisissant la durée des consonnes dans une vitesse d'articulation très lente et celle des voyelles dans une vitesse d'articulation rapide, on arrive à "fabriquer" un style "consonantique" et en inversant on obtient un style de caractère "fortement vocalique".

V - VALIDATION DU MODELE

Les premiers tests du modèle ont été réalisés sur une dizaine de phrases isolées. Quand on utilise le modèle pour une prédiction de durées segmentales en choisissant au départ la durée d'un segment on obtient un écart moyen qui ne dépasse pas ±15 ms entre la durée prédite et la durée mesurée. Dans la prédiction de la durée segmentale à partir de la vitesse d'articulation, l'écart entre la vitesse mesurée et la vitesse prédite était inférieur ou égal à 0.5 syll/sec.

VI - APPORTS DU NOUVEAU MODELE

Même si le ralentissement ou l'accélération du débit d'élocution sont fortement influencés par le nombre et la durée des pauses (4) l'étude sur les possibilités de variation de la vitesse d'articulation n'est pas sans intérêt. En effet, ce nouveau modèle permet :

- de maintenir un rapport juste dans l'assignation automatique des durées des consonnes et des voyelles pour un style "neutre";
- de faire varier la vitesse d'articulation entre deux vitesses extrêmes qui provoque l'impression d'accélération ou de ralentissement;
- de simuler différents "styles" d'élocution en utilisant un "mixage" de différentes vitesses d'articulation pour différents types de sons.

VII - CONCLUSION

Ce nouveau modèle prédictif de la durée segmentale, même s'il devient par ces améliorations plus souple et plus approprié pour exprimer les différentes vitesses et styles d'élocution, n'est toujours pas définitif ni complet. Pour le compléter, il sera nécéssaire de rechercher une durée "neutre" d'un segment en fonction de la durée du groupe de souffle dans lequel il se trouve. A partir de cette durée on pourra, par la suite, effectuer la simulation de l'accélération ou du ralentissement de la vitesse d'articulation et on pourra aussi prévoir l'emplacement et la durée des pauses.

Remerciements

Je tiens à remercier C. Sorin et F. Emerard du CNET de Lannion, ainsi que l'équipe de l'Institut de Phonétique de Grenoble, particulièrement J.-L. Boë et C. Abry, pour leurs critiques constructives de ces modèles prédictifs de la durée segmentale.

Bibliographie

(1) K. Bartkova, "Modèle prédictif des durées phonéthiques en français pour 4 vitesses d'élocution", Rapport de stage, Document interne CNET, 25 pages, 1984.

(2) C. Sorin, K. Bartkova, "Synthèse de plusieurs styles d'élocution : Invariance et Variantes prosodiques", 13e journée du GALF, pp 221-222, Bruxelles, 1984.

(3) D. Klatt, "Synthesis by rule of segmental durations in English sentences", in Frontiers in Speech Communication Research, edited by B.Lindblom and S.Ohman (Academic, New York), 1979.

(4) T. Crystal, A. House, "Segmental durations in connected speech signals : Preliminary results", J. Acoust. Soc. Am. 72, pp 705-716, 1982.

LA PROSODIE DANS LE SYSTEME SYNTHEX

A. Aggoun, F. Emerard, C. Sorin et M. Stella

CNET - Route de Trégastel - 22301 LANNION

ABSTRACT

In this paper, we discuss the SYNTHEX system developped at C.N.E.T. The design of SYNTHEX is an example of the use of Artificial Intelligence techniques in order to study various problems related to speech synthesis from a written text. SYNTHEX is a tool used to formalize prosodic knowledge related to the problem of speech synthesis and to improve this knowledge.

1. INTRODUCTION

Actuellement, il existe de nombreux systèmes de synthèse de la parole [4,6,8] fonctionnant chacun pour une ou plusieurs langues. La prosodie entre dans une large part dans l'agrément de la parole synthétique. Pour traiter la prosodie, on assiste depuis quelques temps à la création des systèmes de développement des règles prosodiques.

Au CNET, les études actuelles dans ce domaine visent à diversifier d'une part les styles de voix synthétiques en Français (style "lecture" ou style "commercial") et d'autre part les langues (Allemand, Italien, Anglais), en utilisant la technique de synthèse par diphones. Il est donc nécessaire de pouvoir modifier un certain nombre de règles "segmentales" et "prosodiques". Disposer de la possibilité d'obtenir plusieurs styles de voix, voire plusieurs voix à partir d'un seul dictionnaire de diphones serait un atout important pour la mise en oeuvre de la synthèse dans diverses applications.

Le système SYNTHEX (Système SYNTHèse EXpert) utilise des techniques d'intelligence artificielle dans le domaine de la synthèse de la parole. Il permet en particulier de formaliser, modifier et améliorer les règles prosodiques nécessaires pour réaliser la synthèse de la parole.

2. LE SYSTEME DE SYNTHESE DE PAROLE

Le système de synthèse de la parole du CNET se compose : d'un module de transcription orthographique-phonétique, d'un module prosodique qui applique un certain nombre de règles [5,7] pour calculer la prosodie de la phrase (la mélodie, la durée), d'un dictionnaire de diphones [5], d'un synthétiseur qui met en oeuvre des techniques de traitement de signal [8] pour produire un signal acoustique de parole.

3. LE SYSTEME SYNTHEX

L'objectif du système est de permettre en particulier une meilleure formalisation du module prosodique d'un système général de synthèse vocale.

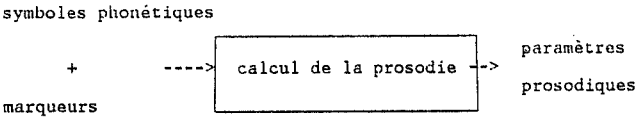


Figure 1. Module prosodique

Le système proposé [1,2] accepte en entrée une suite de symboles phonétiques et de marqueurs prosodiques, il délivre en sortie les paramètres prosodiques : la mélodie (la fréquence fondamentale) l'intensité et le rythme (les durées des segments et des pauses).

SYNTHEX a été développé pour les besoins suivants :

- formaliser les connaissances prosodiques ;
- tester et améliorer facilement ces connaissances (modification des règles, ajout de nouvelles règles) ;
- exprimer les règles prosodiques dans un formalisme simple et inspiré du domaine d'application.

SYNTHEX comprend trois modules : une interface, des bases de connaissances et un mécanisme d'exploitation (moteur d'inférence).

L'interface : elle permet l'acquisition des données et des connaissances sur le domaine sous différentes formes.

Les bases de connaissances : il existe trois types de bases :

la base de faits : elle constitue l'espace de travail. Au début, on y trouve les faits initiaux, les phrases à synthétiser. Ces faits sont utilisés pour évaluer les prémisses des règles.

les bases de connaissances du domaine : elles contiennent des règles relatives au domaine, elles sont introduites par l'expert : ce sont des connaissances prosodiques.

les dictionnaires : ils contiennent des données relatives au domaine appelées les données d'applications (ex : le dictionnaire des diphones).

le mécanisme de contrôle : c'est le processus de déduction du système. Il contient le mécanisme de résolution de problèmes [1] et l'applique aux bases de connaissances pour résoudre le problème posé par l'utilisateur.

SYNTHEX est un système basé sur les règles de production. Une règle de production est de la forme :

si (un ensemble de prémisses)
alors (un ensemble d'actions).

3.1. L'expression des connaissances

Pour faciliter le transfert de la connaissance de l'expert vers le système, nous avons développé un langage d'expression de connaissances adapté aux besoins des experts et du domaine d'application.

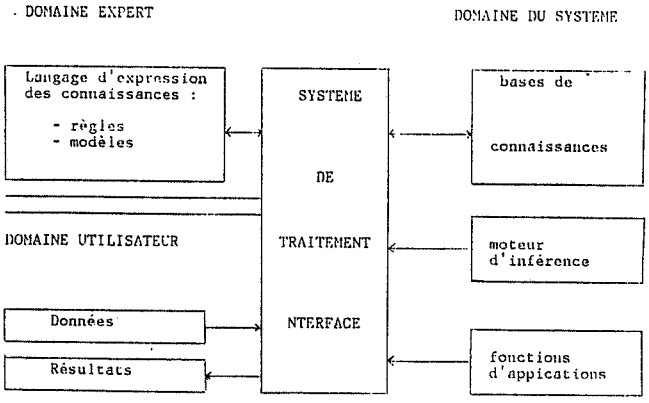


Figure 2. Architecture générale de SYNTHEX

Dans le système SYNTHEX, on distingue d'une part "les modèles syntaxiques" qui permettent de structurer une phrase à synthétiser sous forme d'arbre, et d'autre part "les règles" qui calculent les paramètres prosodiques.

LES MODELES SYNTAXIQUES :

Ces modèles comprennent essentiellement les structures syntaxiques, les définitions et enfin les descriptions toutes fournies par l'utilisateur. A partir de la phrase exprimée sous forme phonétique et enrichie de marqueurs, ces modèles doivent permettre de construire un arbre "syntaxique" sur lequel seront appliquées les différentes règles prosodiques.

Les structures syntaxiques :

exemple : (1) (syntaxe : (phrase mot))

Une phrase est une suite de mots.

(2) (syntaxe : (phrase marqueur))

Une phrase comprend aussi des marqueurs.

Dans SYNTHEX la combinaison des deux structures (1) et (2) sera interprétée comme : une phrase est une suite d'entités "mots" séparés par des entités "marqueurs".

(3) (syntaxe : (mot syllabe phonème trame))

La règle (3) signifie : un "mot" est une suite de "syllabe", une "syllabe" est une suite de "phonème", un "phonème" est une suite de "trame".

Les descriptions : Ce modèle décrit en terme d'attributs chaque objet déclaré dans les structures syntaxiques.

exemples :
(1) (description : phrase (genre))

L'objet phrase est caractérisé par son attribut "genre", le genre peut être par exemple : interrogatif, exclamatif ou affirmatif.

(2) (description : mot (nombre-de-syllabes))

(3) (description : phonème (valeur))

La valeur d'un phonème appartient à l'ensemble des phonèmes d'une langue (ex : le phonème B, le phonème a, etc...).

(4) (description : trame (durée pitch))

L'attribut durée représente le paramètre prosodique durée associé à une trame d'un phonème. Le pitch représente la valeur de la fréquence fondamentale.

(5) (description : marqueur (valeur pause))

La pause est calculée en fonction de la valeur du marqueur et du nombre de mots entre deux marqueurs ; elle représente la période (pause de respiration).

les définitions : Les définitions permettent de décrire comment produire les noeuds fils à partir d'un noeud père lors de la génération de l'arbre syntaxique associé à une phrase.

```
exemple : (définition : mot
           domaine    : PHONEMES
           réécriture : syllabiser)
```

Dans cet exemple, PHONEMES est l'ensemble des phonèmes d'une langue. Les valeurs renvoyées à l'étage "syllabe" sont réécrites en une autre suite, grâce à la fonction "syllabiser", qui décompose la suite de phonèmes d'un mot en syllabes.

L'interprétation de ces modèles consiste à générer un arbre syntaxique (ou structure arborescente) en fonction des connaissances déduites des modèles syntaxiques et de la phrase en entrée à synthétiser. La phrase introduite est une liste de phonèmes et de marqueurs.

```
exemple : l eu " " d r o m a d a i r "S" b w a
          "=" d eu " " l o ".")
```

Dans cet exemple, la suite de symboles (phonèmes + marqueurs) représente la valeur de l'objet "phrase" (racine de l'arbre). A partir des modèles ci-dessus, il y a génération des descendants de l'objet racine pour le second niveau. Deux types d'objet sont générés :

- objets de type "mot" : (l eu), (d r o m a d ai r), (b w a), (d eu) et (l o)
- objets de type "marqueur" : " ", "S", "=" et ".".

Ensuite, il y a application des règles de décomposition syllabique. La suite "(d r o m a d ai r)" se décompose en trois objets de type "syllabe" dont les instances respectives sont "(d r o)", "(m a)" et "(d ai r)". Le même processus est répété pour tous les niveaux jusqu'à la production complète de l'arbre. L'ensemble des objets générés constitue la base de faits sur laquelle sont appliquées les règles prosodiques.

LES REGLES PROSODIQUES

Les "règles prosodiques" tiennent compte des déclarations des modèles syntaxiques et sont exprimées sous forme de règles de productions.

```
syntaxe  : (problème : IDENTIFICATEUR
             si       [ - PREMISSE  ] n , *
             alors    [ - ACTION    ] n , * )
```

où : IDENTIFICATEUR : chaîne de caractères identifiant un problème.

Les propositions sont séparées par "-" qui signifie "et si" quand celles-ci expriment des prémisses (si) et signifie "et puis" quand elles expriment des actions.

L'une des originalités du système SYNTHEX est l'utilisation d'un sous-ensemble du langage naturel (Français) comme langage d'expression des connaissances. Les PREMISSES et les ACTIONS des règles prosodiques sont exprimées dans ce langage quasi-naturel, c'est-à-dire sous forme de propositions respectant une grammaire restrictive du Français augmentée de variables. Syntaxiquement, une variable se présente comme un identificateur précédé du caractère "*".

Le composant "règles prosodiques" du système est organisé en bases de connaissances, chacune d'elles regroupant un ensemble de problèmes. Un problème est constitué des règles qui contribuent à la solution d'une même difficulté. Les règles d'un même problème ont le même nom.

EXEMPLES DE REGLES

```
(problème : durée-initiale
 si
    - il existe un mot *m
 alors
    - attribuer à la durée de toutes les trames du
mot *m la valeur 22.1)
```

commentaires : Cette règle affecte à l'attribut "durée" de toutes les trames d'un phonème de tout mot d'une phrase la valeur 22.1ms.

```
 (problème : durée-consonne
  si
     - la valeur du premier phonème *p1 d'un mot
       *m est de nature consonne
     - la valeur du phonème adjacent droit au
       phonème *p1 dans le mot *m est de nature
       voyelle
  alors
     - écrire ("règle : durées des trames du
       premier phonème")
     - multiplier les durées de toutes les trames
       du phonème *p1 par la valeur 1.15
     - imprimer la valeur du phonème *p1
     - imprimer les durées des trames du phonème
       *p1).
```

Commentaires : Si le premier phonème d'un mot possède la caractéristique "consonne" et si son successeur dans le même mot possède la caractéristique "voyelle", alors appliquer les actions suivantes dans l'ordre d'apparition : imprimer le texte en paramètre de la primitive "écrire", puis multiplier la valeur de l'attribut "durée" de toutes les trames du premier phonème par la valeur 1.15.

(problème : pause
si
- la valeur d'un marqueur *x appartient à la liste ("S" "+" " ")
alors
- attribuer à la pause du marqueur *x la valeur 65ms)

(problème : pause
si
- la valeur d'un marqueur *x appartient à la liste ("." "?")
alors
- attribuer à la pause du marqueur *x la valeur 400ms)

commentaires : Le problème "pause" se compose de deux règles "pause:1" et "pause:2", permet de calculer la pause (période) en fonction de la valeur du marqueur.

(problème : prosodie

- exécuter durée-initiale
- exécuter durée-consonne
- exécuter pause)

commentaires : Exemple de spécification d'une demande d'application des problèmes "durée-initiale", "durée-consonne" et "pause".

3.2. Domaine de l'utilisateur

A partir du moment où l'expertise est suffisante, on peut utiliser le système SYNTHEX pour calculer les paramètres prosodiques d'une phrase. Le système fournit également des facilités pour le développement des règles prosodiques. L'utilisateur dispose d'un éditeur interactif spécifique du domaine. Les modèles syntaxiques et les règles prosodiques sont structurés en bases de connaissances. L'utilisateur peut - interactivement - définir, mémoriser, éditer, exécuter et modifier les règles.

4. CONCLUSION

Cette étude montre l'intérêt de l'utilisation des techniques de l'intelligence artificielle dans le domaine de la synthèse de la parole. L'outil développé sur ces bases permet de formaliser les connaissances prosodiques et de tester facilement de nouvelles connaissances pour améliorer le naturel de la parole synthétique.

Le système est écrit en LISLOG (3). Le temps d'exécution de l'ensemble des règles introduites actuellement dans le système est d'environ une minute pour une phrase d'une dizaine de mots. Le système a été réalisé sur le système Multics.

5. BIBLIOGRAPHIE

[1] A. Aggoun, F. Emerard, C. Sorin, M. Stella, "Prosodic Knowledge in the Rule-Based SYNTHEX Expert System for Speech Synthesis", NATO advanced Studies Institute New Systems and Architectures for Automatic Speech Recognition and Synthesis, Bonas (France), July 2-15,1984.

[2] A. Aggoun "SYNTHEX" : un système de traitement des connaissances prosodiques pour la synthèse de parole", Thèse Docteur-ingénieur, Université de Rennes, 1985.

[3] S. Bourgault, M. Dincbas, J.P. Le Pape, "LISLOG - Programmation en Prolog en Environnement LISP", 4ème congrès AFCET, Reconnaissance des Formes et Intelligence Artificielle, Paris, 1984, pp. 275-289.

[4] R. Carlson, B. Granstrom, S. Hunnicut, "A Multi-Language Text-to-Speech Module", IEEE, ICASSP 82, pp. 1604-1607.

[5] F. Emerard, "Synthèse par Diphones et Traitement de la Prosodie", Thèse troisième cycle, Université de Grenoble, 1977.

[6] S. Hertz, "From Text to Speech with SRS", JASA, vol. 72, no 4, 1982, pp. 1155-1170.

[7] C. Sorin, M. Stella, A. Aggoun, K. Bartkova, "Règles Prosodiques et Synthèse de Parole MULTI-STYLE", Symposium Franco-soviétique sur le dialogue Homme-Machine, Punchino, sep.84.

[8] M. Stella, "Synthèse de parole", L'Echo des Recherches" No. 115, trimestre 1, 1984, pp. 21-32.

LES TONS DU CHINOIS DE PEKIN, LEUR COMPORTEMENT EN PAROLE CONTINUE

P. Hallé

LIMSI BP 30, 91406 ORSAY

ABSTRACT

After a brief description of the tonal behaviour in Beijing Dialect (further BD), a contrastive method for tone recognition is proposed, which uses tone-to-tone contrasts in addition to a tone's intrinsic features in order to perform tone identification.

I. LES TONS EN CITATION

Chacun des 4 tons du chinois de Pékin est traditionnellement décrit par un contour caractéristique Fo, contour "idéal" pris par la syllabe intonée prononcée en isolation, celui du ton "en citation".

La description qui a connu le plus de succès est celle dite des 5 niveaux de Chao Y.R. [1].

Il s'agit d'une description "perceptuelle". La plage de variation de pitch du locuteur, estimée par l'auditeur (un phonéticien spécialisé) est découpée en 4 intervalles perceptuellement équivalents. Un ton est noté par une séquence de niveaux traduisant la forme de son contour:
55 ton haut plat, 51 ton haut tombant, 35 ton haut montant, 214 ton bas incurvé.

Des études acoustiques récentes de Howie [4] confirment l'allure générale des tons tels que décrits par "les 5 niveaux".

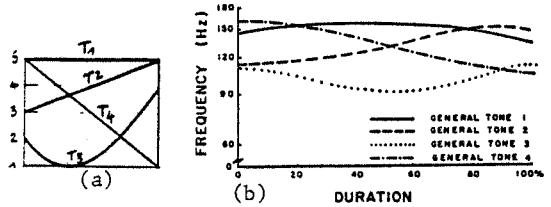


Figure 1

(a) les 5 niveaux de Chao :
ton T1 = 55, T2 = 35, T3 = 214, T4 = 51
(b) courbes moyennes obtenues par Howie

Les 5 niveaux sont couramment utilisés pour décrire la plupart des langues et dialectes à tons de l'Asie. Mis à part quelques divergences, il existe un assez bon accord entre les notations des différents phonéticiens ayant enquété sur un même dialecte (enquêtes portant sur des syllabes ou des mots en citation).

II. LES TONS EN PAROLE CONTINUE

Les contours Fo des tons en parole continue s'éloignent considérablement de leurs modèles canoniques en citation, comme le montre la fig. 2

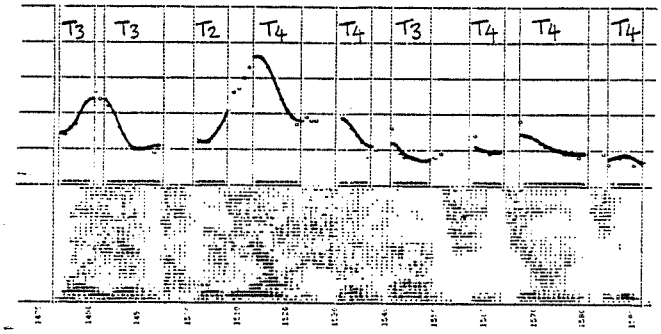


Figure 2
"wǒ yě tóng yàng shè fǎ qù jiù jì"

Les raisons de ces divergences sont principalement :
- **l'intonation:** en particulier, déclivité générale sur chaque groupe de respiration et décroissance du "range" plateau - baseline (voir Vayssière [7])
- **l'accentuation:** plus une syllabe est accentuée, plus son pitch est élevé (dans le cas des tons hauts), et aussi, plus son contour est marqué, i.e. se rapproche du modèle en citation. D'où divers degrés de dégradation allant jusqu'au ton dit neutre, ou ton zéro, des syllabes atones.
- **les sandhis:** on peut expliquer certaines "transformations" des contours tonaux comme des phénomènes de sandhi lorsqu'elles apparaissent de

façon systématique dans un contexte déterminé. Le sandhi le plus connu en BD est celui du ton T3 (cf. la fig. 2 : le premier T3 devient montant).

- la coarticulation entre tons: Des études récentes (Kratochvil à paraître, Sagart & Hallé à paraître) montrent que certains tons ont un effet systématique sur le ton suivant ou, plus rarement, sur le ton précédent. En particulier, le ton T2 relève significativement l'onset (Fo en début de syllabe) du ton suivant **quel qu'il soit.**

Les contours Fo des tons peuvent s'expliquer par un modèle commande-réponse analogue à celui proposé par Fujisaki [2] (où, plus précisément, ln(Fo) est modélisé par la réponse d'un systéme linéaire du second ordre à des commandes échelon).

Ce modèle est en cours d'élaboration dans le cadre d'une ATP (Sagart, Hallé & Boysson-Bardies)

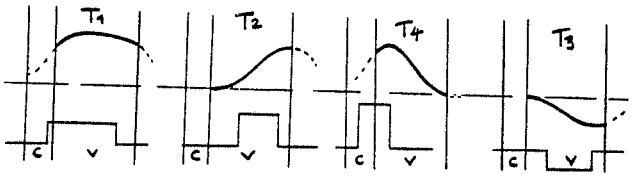


Figure 3
- les tons hauts et leur commande positive
- le ton bas et sa commande négative

L'un des objectifs de ce modèle est de rendre compte de la dépersonnalisation progressive des tons avec la diminution de leur degré d'accentuation, jusqu'au cas extrême du ton zéro, expliqué par une commande d'amplitude nulle, i.e. une absence de commande tonale. Le ton zéro se réalise ainsi comme le prolongement dynamique du ton précédent

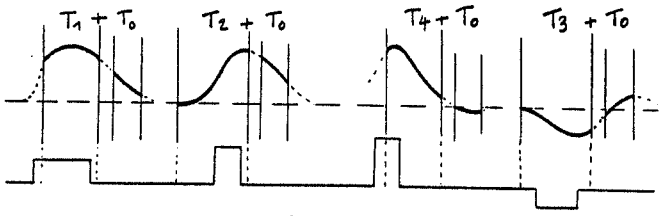


Figure 4
Comment le modèle explique les réalisations du ton zéro en fonction du ton précédent, réalisations généralement observées.

Un autre objectif du modèle est d'expliquer la part mécanique de certains effets de contexte.

Au niveau de la perception, on peut faire l'hypothèse que les tons doivent être identifiés non seulement par leurs propriétés intrinsèques, mais aussi par les contrastes qui existent entre tons successifs. Par exemple, dans une étude non publiée, Sagart a avancé l'hypothèse de la spécificité des différentiels mélodiques entre tons successifs.

La perception contrastive semble aussi jouer dans le domaine spectral (identification de "voyelles" à spectre plat de Summerfield [6])

La suite de cet article décrit une tentative de mettre en oeuvre l'idée que les contrastes aident à l'identification.

III. EXPERIENCE DE RECONNAISSANCE "CONTRASTIVE"

1. Principe

1.1 propriétés intrinsèques des tons

Suivant les expériences de Lublinskaja et al [5],nous pensons que pour décrire un contour Fo, il suffit, plutôt que d'utiliser la séquence entière des valeurs mesurées, de retenir 2 ou 3 paramètres schématisant ses propriétés. Nous proposons ici une valeur Fo, le Fo maximum normalisé sur la section principale du contour (voir § III.2) et la pente relative, au centre de cette portion, soit Pc = (dFo/dt)(1/Fo). La répartition des tons dans le plan Fo x Pc (corpus de 126 syllabes) montrent des recouvrements importants malgré la normalisation.

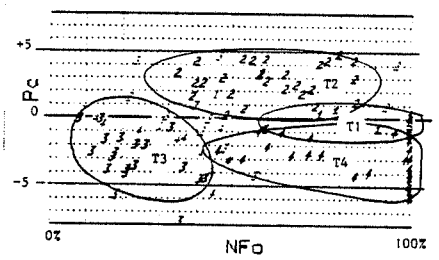


Figure 5
Répartition des 4 tons dans le plan Fo x Pc

Pour la reconnaissance, on cherchera à évaluer $P(s:Ti) = f_i(Fo,Pc)$, probabilité - mais nous préfèrerons parler de **plausibilité** - pour que la syllabe s de paramètres Fo,Pc soit au ton Ti.

1.2 Contrastes dans un disyllabe

Les contrastes tonaux entre les 2 syllabes s_{i-1} et s_i d'un disyllabe sont décrits, selon les paramètres intrinsèques choisis, par :

$$\Delta Fo = Fo(s_i) - Fo(s_{i-1})$$
et $$\Delta Pc = Pc(s_i) - Pc(s_{i-1})$$

On cherchera à déterminer les plausibilités :

$P(s_{i-1}:Tj\ ,s_i:Tk) = g_{jk}(\Delta Fo,\Delta Pc)$

Plausibilité que dans le disyllabe $s_{i-1}s_i$, s_{i-1} soit au ton Tj et s_i au ton Tk, le contraste étant ΔFo,ΔPc. D'où l'on tire :

$$P(s_{i-1}:Tj) = \sum_k g_{jk}(\Delta Fo,\Delta Pc) = g_{aj}(\Delta Fo,\Delta Pc)$$

$$P(s_i:Tk) = \sum_j g_{jk}(\Delta Fo,\Delta Pc) = g_{pk}(\Delta Fo,\Delta Pc)$$

1.3 Les 3 éléments de décision pour une syllabe

Soit la séquence de syllabes $...s_{i-1}s_is_{i+1}...$ s_i y apparaît successivement dans les ditones d_1 et d_2. On pourra donc déterminer :

a) $P_0(s_i:Tj) = f_j(Fo,Pc)$
(plausibilités "intrinsèques")

b) $P_1(s_i:Tj) = g_{pj}(\Delta Fo_1,\Delta Pc_1)$
(plausibilités "contrastives" pour s_i en tant que second élément du ditone d_1)

c) $P_2(s_i:Tj) = g_{aj}(\Delta Fo_2,\Delta Pc_2)$
(idem pour s_i premier élément de d_2)

On en déduit les plausibilités globales résultant de ces 3 "expériences" sur s_i :

$$P(s_i:Tj) = P_0.P_1.P_2$$

On décidera que le ton de s_i est celui qui maximise P.

2. Mise en oeuvre

2.1 Extraction des contours tonaux

Les segments porteurs de tons sont extraits automatiquement par un algorithme de détection des noyaux vocaliques, décrit ailleurs [3] utilisant les fonctions de "loudness" et d'instabilité spectrale.

Le pitch est estimé par un algorithme basé sur les propriétés du "cepstre lissé" [3].

Les contours "bruts" sont ensuite lissés (lissage médiane 3 points suivi d'un lissage Hanning 3 pts) pour pallier aux effets de quantification et aux irrégularités locales de Fo.

2.2 Extraction des paramètres intrinsèques

De ces contours lissés est extraite la **section principale**, i.e. la plus longue portion entre deux extréma locaux successifs. C'est cette section que nous supposons être significative. Elle correspond à la portion ascendante d'un ton T2 ou descendante d'un ton T4 ou T3. Dans le cas d'un ton (sans doute T1) dont le Fo maximum serait centré sur le noyau vocalique, on choisit une section principale centrée sur ce maximum.

Le paramètre Pc est la pente relative prise au centre de la section principale, Fo est le Fo maximum dans cette section (on pourrait aussi choisir le Fo moyen).

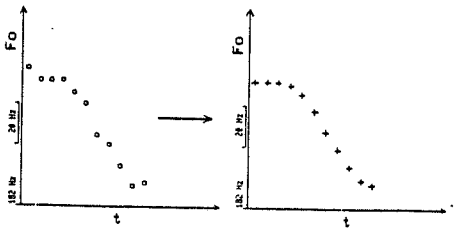


Figure 6
Mise en forme du contour tonal

2.3 Normalisation dans un groupe de respiration

Pour chaque groupe de respiration sont déterminés baseline et plateau [7] :

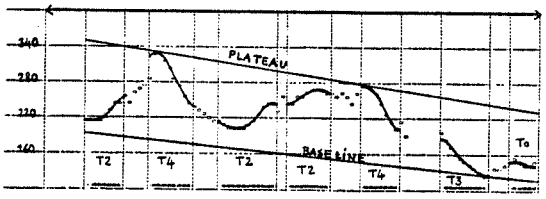


Figure 7
Baseline et Plateau

Les valeurs Fo sont ensuite normalisées en fonction de la hauteur de ces 2 courbes :

NFo = (Fo - Baseline)/(Plateau - Baseline)

Cette "normalisation", comme on le voit, n'est pas sans rapport avec la notation de Chao.

2.4 Estimation des plausibilités

Les fonctions f_j et g_{jk} (d'où les g_{aj} et les g_{pk}) sont approximées à partir des distributions constatées pour les données recueillies sur un corpus initial (texte d'environ 300 syllabes, lu rapidement par une locutrice de Pékin, LSY).

Pour chaque paramètre, la distribution des tons (ou des ditones) est schématisée par les fréquences d'occurence dans un petit nombre de régions (ici 4) partitionnant son domaine de variation. Par exemple, pour le paramètre Fo, la fréquence du ton i est, dans la région k, $f^{Fo}_{k,i}$. De même, $f^{Pc}_{k',i}$ est la fréquence du ton i dans la région k' pour le paramètre Pc. Si Fo appartient à

la région k et Pc à la région k', nous approximons alors $f_i(Fo,Pc)$ par $f^{Fo}_{k,i}.f^{Pc}_{k',i}$

La même méthode est utilisée pour les paramètres contrastifs ΔFo et ΔPc et les 16 ditones possibles.

3. Expérimentation

Nous avons testé un corpus de 126 syllabes prononcées par LSY. 123 noyaux ont été correctement détectés et l'expérience de reconnaissance a donc porté sur les 123 tons correspondants.

(a)

TON RECONNU \ TON LEXICAL	1	2	3	4	ø
1	12	5		4	1
2		31	1		
3		5	21	5	3
4			1	30	4
ø	12	41	23	39	8

(b)

TON RECONNU \ TON LEXICAL	1	2	3	4	ø
1	10	5		5	
2	1	29	2		1
3	1	7	20	13	3
4			1	21	4

Table I

Matrices de confusion obtenues
(a) reconnaissance contrastive
(b) reconnaissance à partir des seules propriétés intrinsèques (Fo et Pc)

Les résultats pour le ton neutre vont dans le sens des prédictions (voir fig.4) : reconnu T4 après T1 ou T2, T3 après T4.

DISCUSSION

Les divergences les plus notables correspondent aux sandhis répertoriés. En particulier au sandhi du double 3ème ton traditionnellement considéré comme paradigmatique: T3 T3 -> T2 T3 (T2 substitué à T3). T2 souvent pris pour T1 confirme le sandhi - moins connu - du 2ème ton plat.

L'utilisation des contrastes permet cependant dans nombre de cas de de reconnaitre des tons dont les propriétés intrinsèques sont fort éloignées de la "norme".

Les améliorations prévues portent sur l'estimation des plausibilités, devenant si possible évolutive au cours de la reconnaissance.

REFERENCES

[1] Chao, Y.R. (1930). A system of tone letters. Le Maître phonétique 45, 24-47.

[2] Fujisaki (1984). Dynamic characteristics of voice Fo in speech an singing. In MacNeilage (ed.), The Production of Speech. Springer Verlag.

[3] Hallé (1985). Segmentation syllabique et reconnaissance des tons du chinois en parole continue. Thèse de 3ème cycle, Paris XI.

[4] Howie (1976). Acoustical Studies of Mandarin Vowels and Tones. Cambridge University Press.

[5] Lublinskaja, Mikiel & Sheikina (1972). Scaling of psychological distances between sounds with changing Fo. Voprosy teorii i metodov issledovanija vosprijatija rechevych signalov, no 3. Leningrad.

[6] Summerfield, Haggard, Foster & Gray (1984). Perceiving vowels from uniform spectra: phonetic exploration of an auditory aftereffect. Perception & Psychophysics, 27 (4), 203-213.

[7] Vayssière (1980). Search for language-independent prosodic features. 1st International Congress on the Perception of Speech. Florence, Décembre 1980.

```
*********************************************
*      S E S S I O N   A F F I C H E E      *
*********************************************
```

ASSIA, UN EDITEUR "INTELLIGENT" POUR LA MANIPULATION ET L'ANALYSE DU SIGNAL VOCAL

GONG Yifan et Jean-Paul HATON

Equipe "Reconnaissance des Formes et Intelligence Artificielle"

Centre de Recherche en Informatique de Nancy
B.P. 239 54506 Vandoeuvre-les-Nancy Cedex

A general purpose integrated signal and data processing system is defined on the basis of the abstraction of three interactively related activities: - processing of signals and large data base which is viewed as information retrieval and signal modification, - development and simulation of new algorithms and - visualisation of the results under different parametrizations. Particularly designed for acoustic-phonetic editing and analysis in speech recognition research, the system provides an environment for signal processing and the simulation of signal processing algorithms described by combination of operations. Signals together with signal processing routines, including graphic terminal display facilities, are represented by defining a set of abstract objects associated with specified operations. The basic principle of the system is that the operators are implemented as modular programs having compatible input-output data structures and new operators can be constructed either from existing operators or from user-specified programs, thus simplifying system maintenance, modification, augmentation and improvement. The general system design is described followed by examples of practical use.

I. INTRODUCTION

L'étude d'un signal se fonde à la fois sur le phénomène physique et la théorie. Par conséquent, la recherche dans le domaine de traitement du signal est un processus ayant deux aspects liés. Le premier aspect part de phénomènes concrets tel que la la paramétrisation de la parole. Il s'agit alors de traiter le signal d'un grand nombre de manières différentes. L'objectif de ces traitements est d'extaire de l'information transportée par le signal ou de modifier le signal lui-même. Le deuxième aspect part de constructions abstraites telles que le développement et l'implantation de nouveaux algorithmes. Cet aspect consiste à définir et réaliser des opérations sur les phénomènes. On étudie la structure interne des traitements et on définit de nouveaux signaux à l'aide de ces traitements. Par ailleurs, il est nécessaire d'établir la liaison entre l'utilisateur et le système. Ceci constitue un troisième aspect important, la visualisation des phénomènes.

Depuis ces dernières années, la recherche sur l'environnement logiciel intégré pour faciliter l'étude du signal a conduit à plusieurs systèmes liés aux trois aspects précédents. ILS (Interactive Laboratory System) [1] a été conçu plus particulièrement pour l'analyse, la synthèse et la transmission de la parole. Il peut être considéré comme une collection de programmes de traitement du signal et l'organisation du système et les structures de données ne permettent pas facilement d'enchaîner et de recombiner des programmes. Un autre système plus formel, ISP (Integrated Signal Processing System) [2], est fondé sur la création et la manipulation des "objets signal". L'une des caractéristiques importantes de ISP est que c'est un système initialement conçu pour traiter des sigaux individualisés et les signaux à manipuler sont réferencés par une pile. Par conséquent le système est relativement limité pour le traitement des grands corpus.

Partant de l'abstraction des trois aspects décrits plus haut, nous avons défini et construit un systeme pour **l'Analyse de Signal et la Simulation Interactive d'Algorithmes** (ASSIA), système logiciel destiné au traitement numérique et statistique du signal et des données et au développement d'algorithmes de traitement du signal. Sa conception est orientée vers la reconnaissance acoustico-phonétique de la parole. L'objectif est d'avoir un environnement bien conçu, ayant une certaine abstraction, puissant, facile à modifier, ayant la possibilité d'évoluer et de construire de nouveaux algorithmes et enfin facile à manipuler. L'idée de base du système consiste à définir des types de signaux, objets abstraits, et des opérations indépendantes associées. Le système permet de construire de nouvelles opérations en enchaînant des opérateurs primitifs et donc de modifier et d'augmenter le système. De plus, le système peut manipuler des bases de données de signal, ce qui constitue un outil puissant pour traiter de grand corpus de parole.

Nous décrivons dans cet article les points essentiels concernant la définition, l'organisation et l'application de ce système. Nous supposons le lecteur familiser avec les visualisations graphiques de signaux et nous attachons ainsi aux concepts en laissant de côté l'aspect "imagier".

II. DEFINITION DES TYPES ABSTRAITS LIES AU SIGNAL

1. Type signal

Dans un grand nombre de problèmes concernant le traitement du signal et la reconnaissance de formes, l'hypothèse de la stationarité est souvent faite. Cette hypothèse conduit à la notion d'analyse à court terme. Sous cette hypothèse, un segment de signal est isolé et traité en tant que tel. Ce segment correspond à l'observation du signal sur une fenêtre de largeur fixée. Le signal fenêtré contient la plupart des informations sur le signal original et constitue une representation du signal original. Dans le système ASSIA, on modèlise cette représentation par l'introduction du type "signal" de la forme:

signal : [0..N) |----> R.

Le type signal est un objet abstrait directement manipulé par un certain groupe d'opérateurs dans le système. Il est défini par les opérations qui lui sont associées. Le signal vocal temporel, son spectre, le contour de pitch, etc., sont des exemples du type signal.

L'ensemble des types ayant les mêmes propriétés graphiques constituent une classe. A chaque classe est associée une image avec un format correspondant. Une image constitue un intermédiaire important de la communication entre l'utilisateur et le système. L'utilisateur a la possibilité de définir une zone, une durée, un événement ou un point dans la séquence du signal par une marque ou un curseur à l'aide d'un moyen de communication tel que menu, souris ou clavier. Il existe des images pour le signal temporel et le profile spectral (1-dim), le spectre-temps (2-dim) et le spectrogramme (3-dim).

L'ensemble des signaux créés par la même source et de même type réside sur un support: l'univers. Un univers est une reproduction fidèle de la réalité. Physiquement, il est l'enregistrement du signal. Pour un signal de dimension un il est complètement défini par le nom de l'enregistrement, la position initiale, la longeur de la séquence et la valeur de chaque echantillon.

2. Opérateurs primitifs

La définition complète des types nécessite de préciser des opérations qu'on peut effectuer. Le traitement d'un signal par un opérateur est consideré dans notre système comme une extraction de l'information du signal ou une modification du signal. Nous avons consideré qu'un système doit être non seulement conçu pour son application présente mais également pour des applications futures. Or, étant donné l'accroissement de nombre d'algorithmes développés pour l'analyse du signal et de la parole en particulier, au cours des dernières années, aucun système figé ne pout correctement remplir cette condition. Le système doit donc posséder la possibilité de construire de nouveaux opérateurs pour s'adapter à l'application demandée [3].

Un autre problème pour l'utilisateur est que, pour un nouveau système, il faut apprendre à comprendre l'organisation, le fonctionnement, les définition et les commandes du système. Le temps nécessaire pour se familiariser avec un système et exploiter toutes ses possibilités est souvent considérable. Dans notre système nous avons cherché le plus possible à utiliser des notions naturelles et faciles à accepter.

Il est souhaitable, d'autre part, qu'un grand nombre d'opérateurs dans le système puissent être réutilisés directement et de façon transparente. L'utilisateur a donc à la fois à sa disposition des outils puissants préexistants et la possibilité de compléter et d'améliorer ces outils selon ses applications spécifiques.

Un opérateur réalisant une certaine opération peut se décomposer en une séquence de sous-opérateurs. Chaque opération possède des caractéristiques relativement simples et bien définies et un format commun d'échange d'information avec l'exterieur. Cette notion a déjà été partiellement utilisée pour des applications [4]. On sait par exemple qu'une analyse cepstrale classique pour obtenir un spectre lissé peut se décomposer en: -acquisition des données brutes, -fenêtrage temporel, -transformée de Fourier, -calcul de log-norm pour chaque composante, -transformée de Fourier inverse, -fenêtrage cepstral et -transformée de Fourier. Ainsi, disposant d'une batterie d'opérateur de base, il devient possible de définir et de réaliser de nouveaux algorithmes de traitement du signal avec un gain de temps important. Cette organisation permet aussi aux utilisateurs non-informaticiens de manipuler aisément les opérateurs du système.

La plupart des algorithmes de traitement du signal peuvent être formulés à l'aide de la composition de calculs matriciels ou de traitements successifs, ce qui permet leur décomposition en étapes de réalisation représentées par des opérateurs. Cela fournit la possibilité de tester facilement un algorithme avant de chercher à reduire son temps de calcul. C'est une propriété très utile lors du développement d'algorithmes, car c'est souvent la performance de traitement qui est recherchée dans un premier temps et la préoccupation sur la vitesse d'exécution n'a de sens que si cette performance est satisfaisante.

Le plus souvent, le resultat d'une opération est encore de type signal. Ceci permet d'enchaîner les opérateurs pour définir des nouveaux types de signal. Le système ASSIA, posséde un grand nombre d'opérateurs pour l'analyse de la parole, il ne doit cependant pas être considéré simplement comme une collection de programmes, mais comme un noyau qui fournit une base de travail pour la construction et le développement de nouveaux algorithmes.

III. IMPLANTATION

1. Structures de programmation

Le système est implanté sur mini-ordinateur SM90 sous système d'exploitation Unix. Unix fournit un moyen très souple de communication entre les processus. Il réalise le schéma de transmission de l'information du producteur vers le consommateur par le mécanisme de "tube" (pipe-line). La redirection de l'entrée-sortie d'un programme permet facilement de diriger le flux de données sans passer par des fichiers intermédiaires. Un autre avantage de Unix est de pouvoir disposer d'un standard uniforme d'entrée-sortie permettant de fournir aux opérateurs des arguments, source ou destination de l'information, sans explicitement les préciser. On a aussi la possibilité de lancer en paralèle plusieurs processus. Tous ces caractéristiques ont facilité la réalisation du système ASSIA.

Dans Unix, Shell est une interface essentielle entre l'utilisateur et le système d'exploitation. Il est non seulement un interprète de commande mais aussi un langage de programmation. Ceci permet d'écrire des procédures de commande et d'enchaîner des différents programmes pour des applications importantes. Shell est également utilisé comme interprète du système ASSIA. Il sert principalement à l'analyse syntaxique des commandes, la préparation des différents arguments de description de l'opération à effectuer, le lancement d'un ou plusieurs programmes correspondant aux opérations demandées par l'utilisateur, la lecture et l'écriture sur l'entrée ou la sortie standard pour recevoir ou émettre de l'information complémentaire concernant les programmes d'opérations et la réalisation des nouveaux opérateurs à partir de ceux qui existent déjà dans le système.

ASSIA dispose d'un ensemble de variables définies globalement et représentant l'environnement du système. Ces variables sont une source d'information accessible par le Shell et par les programmes d'opération et sont modifiables par utilisateur à l'aide des opérateurs spéciaux. Un opérateur peut être une variable. Un exemple de variables sont la longeur de fenêtre d'analyse et le type de fenêtre utilisée, par exemple "window_length = 256" (variable scalaire) et "window = hamming" (variable d'opérateur).

Les opérateurs sont implantés physiquement par des programmes. Ils sont traités par Unix comme des "Unix processus". Il n'y a pas de limitation sur le langage de programmation, le système actuel étant écrit en C et en Pascal. La structuration stricte et les possibilités de

diagnostic de Pascal facilitent le développement du système tandis que l'efficacité de C permet d'obtenir des exécutions rapides. L'utilisation de C permet aussi de disposer des primitives de base pour la gestion de l'écran "bit-map". Un programme d'opération contient essentiellement quatre parties separées:
- La partie physiquement liée au système, effectuant l'interprétation des commandes descriptives fournies comme des paramètres de programme,
- La partie assurant les entrées-sortiées, généralement commune pour toutes les opérateurs,
- La partie fonctionnelle qui réalise l'opération demandée et
- La partie décrivant la syntaxe de commande et le mode d'emploi de l'opérateur consideré, faisant partie du système d'aide à l'utilisation.

Les programmes d'opération (P.O) sont définis recursivement par une syntaxe présentée sous forme de Backus-Naur sur la figure 1.

```
P.O ::= NULL | COMMANDE SHELL |
        OPERATEUR | PROCEDURE SHELL.
COMMANDE SHELL ::= grep | cat |... <expression>
PROCEDURE SHELL ::= FONCTION(PROCEDURE SHELL |
                    COMMANDE SHELL | OPERATEUR ).
OPERATEUR ::= P.O.P.1 | P.O.P.2 | .... | P.O.P.n <déscriptions>.
P.O.P.i ::= i-ième programme d'opération primitive.
```

Fig. 1 Définition d'un programme d'opération

Nous distinguons cinq niveaux de programmation dans ASSIA. Ce sont: le noyau du système d'exploitation, les bibliothèques, les programmes d'opérations primitives, les opérateurs de construction en procédure de Shell qui réalisent des opérations composées fréquemment utilisées, et l'utilisateur. En ce qui concerne les bibliothèques, nous disposons de la bibliothèque de Unix et celle des primitives de la gestion du "bit-map". Une bibliothèque propre au système ASSIA a été construite; elle contient trois sous-bibliothèques: le traitement et l'analyse de signal, le calcul vectoriel et matriciel et le graphique "bit-map".

Différents types de signal sont implantés sous une structure commune. Ils sont compatibles au sens de leur représentation physique. Cette réalisation est avantageuse car elle permet la coexistence dans le système de plusieurs représentations des données: virgule fixe, virgule flottante et signal-réel, signal-complexe. La dimension des données peut aussi être une variable. Une considération importante pour la représentation de données est de trouver un compromis entre la rapidité d'échange d'information entre les différents processus, la compatibilité de format et la visualisation des données. Diverses couches sont implantées dans le système: le tableau, la suite d'octets (structure commune pour communication entre programme, par "Unix pipe" notamment), la suite de nombres entiers (représentant l'univers) et la représentation en ASCII liant le système et l'utilisateur pour introduire manuellement et visualiser des données.

2. Interface avec l'utilisateur

Le poste de travail du système ASSIA comprend un écran de haute resolution, de type "bit-map", associé à un clavier, une souris pour le positionnement du curseur et la sélection des opérateurs et des descriptions associées et une entrée-sortie vocale. Au niveau de l'interface avec l'utilisateur, le système doit résoudre les problèmes suivants : indiquer en clair à l'utilisateur ce que peut faire le système a un instant donné, donner à l'utilisateur la possibilité de choisir une action proposée par le système, récupérer l'information fournie par l'utilisateur et maintenir le contact direct avec le système d'exploitation.

La communication se fait par trois moyens: la souris avec le menu, la souris avec l'image du signal et le clavier alpha-numérique. Une flèche sur l'écran dont le mouvement correspond au déplacement de la souris permet à l'utilisateur de pointer ou de se positionner sur l'écran. Trois boutons sur la souris peuvent être utilisées pour déclencher un petit ensemble de commandes dont la définition change selon ce qui a été tracé à un instant donné. La fonction de base de la souris est d'indiquer une position sur l'écran. En plus de la sélection de menu, la souris permet également d'indiquer aisément un point d'une courbe. L'ensemble menu plus souris constitue un moyen rapide et puissant pour le contrôle interactif du système. Après avoir reçu une séquence de commandes, le système posséde les informations sur ce que l'utilisateur va pouvoir demander. Ces informations contextuelles peuvent être utilisées pour reduire la manipulation. Dans un état donné, le système peut donc proposer tous les choix possibles par un menu fonction du contexte affiché sur l'écran. L'utilisateur peut faire son choix à l'aide de la souris. Cette méthode de communication évite l'introduction de fausses commandes et les erreurs syntaxiques puisque seuls les choix valides peuvent apparaître sur le menu. Quelques heuristiques concernant l'utilisation de menus sont adoptées dans le système: - Des menus longs sont décomposés en plusieurs morceaux organisés en arborescence. - Les symboles n'existant pas sur clavier (tels que les symboles phonétiques) sont affichés sur le menu pour être sélectionés. - Un menu peut être placé à tout entroit de l'écran.

Un système d'aide à l'utilisation constitue une partie importante d'ASSIA. Son rôle essentiel est d'expliquer l'utilisation de chaque opérateur. Il fournit en particulier des messages sur la description syntaxique de commande, les entrées-sorties (message court) et le fonctionnement exterieur de l'opérateur, la limitation de l'application et de l'implantation de cet opérateur (message long). Ces messages apparaissent lorsque la commande ne correspond pas à la syntaxe attendue ou encore à la demande de l'utilisateur.

3. Présentation générale des opérateurs

Les opérateurs agissent sur les objets "source" et (éventuellement) provoquent des objets "destinateur". La plupart d'entre eux possédent un résultat du type signal et la sémantique de ces opérations sont des paramétrisations différentes. Les sources et les destinateurs peuvent être indiqués dans l'argument. La syntaxe de commande est de la forme:

OPERATEUR <expression>

Où "expression" désigne une description de l'opération, une liste de paramètres ou un mot-clé du langage. Les opérateurs disponibles concernent:
- la compression de données et l'extraction de paramètres: codage par prédiction linéaire, mélodie, passages par zéro, énergie, fonction d'aire du conduit vocal, valeur maximale et minimale, fréquences des formants, etc.
- les opérations matricielles et arithmétiques: addition, soustraction et mutiplication de deux vecteurs, multiplication matricielle, solution du système linéare, calcul de vecteurs propres, dimension de vecteur, etc.
- les transformations orthogonales: transformée de Fourier discrète (DFT), analyse cepstrale, transformée de Karhunen-Loeve, etc.
- la modification du signal: filtrage (convolution) par filtres de différentes caractéristiques, sous-échantillonnage, addition de bruit au signal, etc.
- la génération de signaux: signal sinus, bruit blanc,

conception de filtres à reponse finie ou infinie, etc.
- l'estimation spectrale: modèle AR, ARMA, etc.
- l'analyse statistique: moyenne, corrélation de deux signaux, analyse en composantes principales, analyse de "clustering", etc.
- la visualisation: signal temporel, signal positif, profil de spectre, spectrogramme en 12 niveaux de gris, contour de mélodie, etc.
- la conversion de représentation: univers->signal, signal->univers, octets->ASCII, ASCII->octets, acquisition et restitution de parole, etc.
- la modification de variables globales.
- l'aide à l'utilisateur.
- la manipulation du "bit-map": menu, curseur, souris et affichage.
- l'édition du signal vocal. A chaque univers représentant le signal, une liste de quintuplets {N1,N2, transc, contexte, description} est associée pour indiquer la transcription phonétique et la position correspondante. Un groupe d'opérations permettent d'effectuer les manipulations courantes: étiquetage phonétique de segments, insertion et suppression des zones de signal, segmentation et concaténation de signaux, recherche d'un signal particulier dans une base de donnée, recherche de toutes les occurrences d'un phonème, etc.

4. Construction des nouveaux opérateurs

A titre d'exemple simple, nous présentons la réalisation d'un algorithme d'estimation spectrale à l'aide d'un modèle ARMA [5] (hparma) synthétisé à partir des opérateurs primitifs. Cet algorithme est décrit sur la figure 2.

```
hparma [p][q] ::=
  $window_length|$window|mautoc $1 $2|tee automat.m|mat -c > p.v
  cat  p.v automat.m| zero > z.v
  cat  p.v z.v  |  dfthparma

mautoc [p][q] ::= matautoc $1  $2 >  autoc.m
                  autoc  $1     > autoc.v
                  cat  autoc.m  autoc.v
```

Fig. 2 Définition d'un estimateur spectral ARMA

Sur cette figure, "window" est une variable d'opérateur prédéfinie. "mat -c" indique que l'opérateur "mat" a travaillé sur l'objet transmis par "Unix pipe" (le même que "automat.m") tandis que l'objet créé "p.v" a pour semantique les coéfficients des pôles du modèle. En fait, l'opérateur "mat" lui-même se compose d'un certain nombre d'opérateurs. L'option "-c" signale l'intervention de l'opérateur "mat.cholesky", donnant la solution d'un système d'équation par la méthode de Cholesky. On peut visualiser le résultat en appliquant:

```
u_s [block][sample][length]<univ|hparma [p][q]|image [y1][y2] as [mode]
```

"u_s" est un opérateur qui convertit le signal de l'univers et qui a comme source de données l'entreé standard. Y1 et y2 précisent la position sur l'écran, "as" est un mot-clé de la syntaxe et "mode" un mode graphique. Il est également facile de définir des opérateurs formant des boucles contenant un ou plusieurs opérateurs. L'exemple de la figure 3 qui consiste à appliquer l'opérateur "hparma" à tous les univers dans un sous-ensemble de base de donnée illustre cet aspect. Dans cet exemple, l'opérateur "forall [arg1] in [arg2] do [operator] ..." applique le traitement "operator" sur toutes les occurences de l'événement "arg1" dans l'univers "arg2", ce qui est une opération très fréquente en reconnaissance acoustico-phonétique de la parole, citons par exemple la création de références, le test de résultats de classification, la réglage de seuils ou l'évaluation d'algorithmes.

```
with [phoneme] do [operator][expression] ::=
  for univers in *
  do
  forall  $1 in  $univers do $3 $4 $5 $6 $7
  done
```

Fig. 3 Exemple d'application d'un opérateur sur toutes les occurrences d'un phonème

IV. CONCLUSION

Dans la conception du système ASSIA, nous avons tenté de résoudre le problème suivant: Etant donné de nombreux algorithmes de traitement du signal, des grands corpus de données à manipuler, comment organiser un système logiciel efficace pour la recherche? Nous proposons de considérer l'étude du signal comme un processus à trois aspects liés. En introduisant la notion de type d'objet signal, défini en terme des différentes opérations applicables, nous avons réalisé avec un modèle conceptuel unique un système à la fois souple et général pour l'analyse et traitement de signaux ainsi que pour la simulation de algorithmes de traitement. Le système fournit à l'utilisateur non seulement un moyen puissant pour le calcul et la visualisation, mais aussi la possibilité de construire de nouveax algorithmes à partir d'opérateurs primitifs existants, ce qui permet de suivre l'évolution des techniques dans le domaine de traitement du signal et/ou de la communication parlée.

Parmi les systèmes de visualisation et traitement le signal, le système ASSIA se caractérise par une grande généralité dans sa construction et dans les objets qu'il manipule, son orientation vers la modèlisation et la reconnaissance acoustico-phonétique de la parole, notamment vers l'évaluation et l'utilisation de traits acoustico-phonétiques. De plus, le nombre des opérateurs dont ASSIA peut disposer est potentiellement illimité: l'utilisateur peut facilement construire lui-même des opérateurs algorithmiques en utilisant des opérateurs primitifs du système. Le système est ainsi "vivant" et sa capacité s'augmente au fur et à mesure de son utilisation. Le système permet à l'utilisateur d'essayer et de tester rapidement diverses variantes d'un algorithme. On dispose ainsi d'un langage de très haut niveau très facilement accessible et utilisant au mieux les possibilités du système Unix dans lequel il est intégré.

BIBLIOGRAPHIE:

[1] "Summary of ILS Program", Signal Technology, Inc.
[2] G. E. Kopec, "The Integrated Signal Processing Systeme ISP", IEEE Trans. ASSP, vol.ASSP.32, pp842-851, Aug.1984.
[3] H. Hanusa, "Tools and Techniques for the Monitoring Interactive Graphics Dialogues", Int. J. Man-Machine Studies 1983(19), pp.163-180.
[4] D. H. Johnson, "Signal Processing Software Tools", Proc. IEEE ICASSP., 1984, 8.6.
[5] J. A. Cadzow, "High Performance Spectrum Estimation - A New Method", IEEE Trans. ASSP, Vol.ASSP.28, pp.524-529, Oct.1980.

UN INVERSEUR DE SPECTRE INSTANTANE

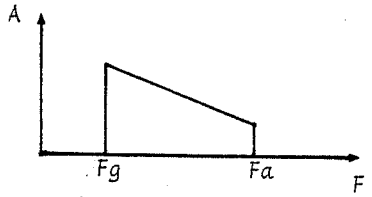
B.Teston

Institut de Phonétique L.A.261 C.N.R.S.
29 Av R.Schuman AIX en PROVENCE 13621

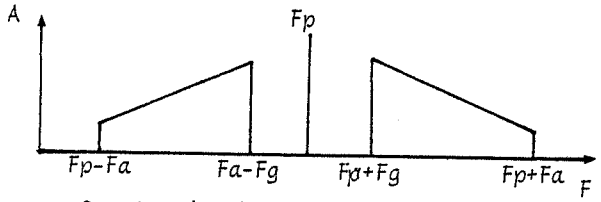
AN INSTANTANEOUS SPEECH SPECTRUM ROTATOR.
The speech spectrum rotator is made like a didactic device for example:prosodic countours acquisition aid in foreign language studies,phoniatric education,and like stimulus generator for speech perception experiments on prosodic parameters. The spectral rotation destroy the semantic and syntactic content without altering the prosodic structures.For this application the apparatus must work instantaneously.The rotated spectrum is obtained by isolation of the lower side band after a modulation of speech signal with a carrier frequency.Accurated phase-shift networks and multipliers give to the machine good signal range and linearity.It is also possible to shift the inverted or no inverted spectrum by a certain amount and to balance it.So the apparatus can make a great many processes on speech and music signal.

INTRODUCTION.
Les sciences du langage accordent une grande importance à la connaissance et à l'analyse des faits prosodiques dont les paramètres peuvent être définis au sens large par:L'intensité(l'accent),la hauteur(la mélodie),et la durée(le rythme,la pause,et le débit).Pour cela,l'opération qui consiste à supprimer l'information sémantique et syntaxique du langage parlé ,pour n'en conserver que l'information prosodique,a depuis longtemps,attirer l'attention des chercheurs dans deux domaines d'application en particulier.D'une part,dans la didactique des langues(acquisition des contours prosodiques),et l'éducation ou la rééducation phoniatrique(traitements de certaines aphasies et trouble de l'élocution prosodique).D'autre part,dans l'étude des mécanismes de perception et d'intégration des différents paramètres prosodiques de la parole.
Cependant,brouiller suffisemment le signal de parole pour en détruire l'intelligibilité et conserver l'essentiel de son information prosodique n'est pas une opération facile si on la veut satisfaisante.
Dans le premier domaine d'application,on utilise pour ce faire,un énergique filtrage passe-bas du signal de parole,qui ne laisse subsister que la partie grave du spectre contenant la fréquence fondamentale et le premier formant.Le résultat en est un son très sourd et très bruité car une grande partie de l'énergie du signal disparait dans le filtrage.
Sa fréquence de coupure étant fixe,l'éfficacité du filtrage est tres variable selon le type de voyelle et la dynamique de variation de la fréquence fondamentale.Le mauvais résultat acoustiquede cette méthode semble être la raison d'une éfficacitté pédagogique douteuse et peu reconnue.Pour cela son utilisation ne s'est pas généralisée,bien que certains auteurs de méthodes pédagogiques la recommande.
L'inversion spectrale,semble par contre,être un procédé de choix pour ce type d'application.

PRINCIPE DE L'INVERSION SPECTRALE.
Comme son nom l'indique,cette opération consiste à inverser le spectre d'un objet sonore sur l'axe des fréquences,c'est à dire que les fréquences graves (Fg) se retrouvent à la place des fréquences aigues (Fa) et réciproquement.On la réalise en modulant une fréquence porteuse Fp (son pur) par le signal sonore On obtient ainsi un spectre complexe composé par la fréquence porteuse Fp centrée entre deux bandes latérales dont la plus basse est constituée par le spectre inversé du signal de modulation.



Spectre du signal avant modulation.



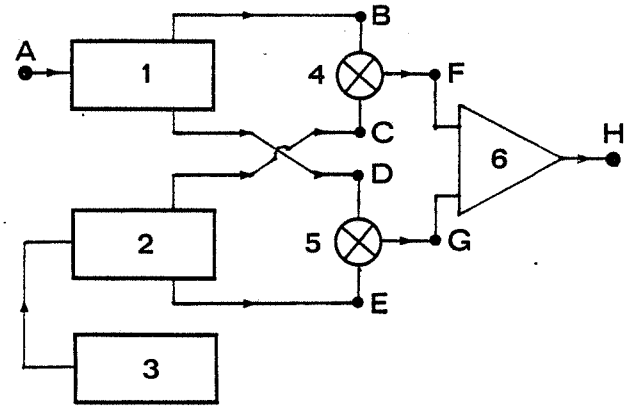
Spectre du signal après modulation.

On peut isoler cette bande latérale basse au moyen de différents procédés qui éliminent la porteuse Fp et la bande latérale haute(Fp+Fg,Fp+Fa),un filtrage passe-bas par exemple.La position du spectre inversé sur l'échelle des fréquences dépend de la valeur de la fréquence porteuse Fp.Le résultat auditif d'une telle opération sur des signaux de parole est une inintelligibilité quasi absolue de leurs contenus sémantique et syntaxique.Au plan prosodique,tous les paramètres liés à la durée(pause,rythme,débit) sont parfaitement conservés.L'intensité acoustique objec-

tive est également conservée,car les niveaux d'énergie des composantes fréquentielles du signal inversé sont identiques à ceux du signal initial (contrairement aux effets du filtrage passe-bas mentionné précédemment).Cependant,la position des composantes sur l'échelle des fréquences étant différente,la sonie est affectée par l'inversion spectrale,mais cet inconvénient peut être rattrapé facilement.Quant à la mélodie,elle semble trés perturbée par l'inversion spectrale.En effet,la périodicité du signal inversé est généralement détruite car ses composantes représentent une série non harmonique et ne sont pas des multiples de la fréquence fondamentale.Cependant il a été montré que la sensation de hauteur mélodique d'une série non harmonique est donnée par la différence entre ses composantes qui reste constante malgré l'inversion.La sensation de hauteur est également donnée par la modulation d'amplitude du signal inversé au rythme de la fréquence fondamentale. Il résulte de ces considérations ,selon Blesser(1), que la hauteur mélodique d'un signal n'est pas affectée par l'inversion spectrale.

REALISATION DE L'INVERSEUR DE SPECTRE.

Le principe de l'inversion de spectre étant connu, il est nécessaire de supprimer la bande latérale haute ainsi que la fréquence porteuse.Nous avons mentionné précédemment le filtrage passe-bas pour réaliser cette opération.C'est la méthode traditionnelle,utilisée entre autres par Blesser(1).Mais,cette méthode rend nécessaire l'utilisation d'un filtre caractérisé par une trés forte chute par octave,si l'on veut efficacement supprimer la porteuse.Ceci des distorsions dues au déphasage non homogène d'un tel filtre et une difficulté particulière à ajuster sa fréquence de coupure en fonction de la valeur de la fréquence porteuse Fp.Une autre méthode décritte par Van Nes (2),consiste à supprimer la bande latérale supérieure en la déphasant de 180° avant sommation.Le principe en est donné dans la figure suivante.



1et2-Réseau de déphasage constant à 90°.
3-Oscillateur à fréquence variable générateur de Fp.
4et5-Multiplieur à quatre quadrants.
6-Sommateur ou soustracteur.
A-Entrée du signal A(t).B-cos At.C-cos Pt.
D-sin At.E- sin Pt.
F- 1/2cos(P-A)t+1/2cos(P+A)t.
G- 1/2cos(P-A)t-1/2cos(P+A)t.
H- Signal de sortie.Si 6 est sommateur:
F+G=cos(P-A)t=Bande latérale inférieure:Le signal de sortie a un spectre inversé.

Si 6 est soustracteur:
F-G=cos(P+A)t=Bande latérale superieure:Le signal de sortie a un spectre décalé.

Il est ainsi possible de faire glisser sur l'axe des fréquences les spectres inversés (bande latérale inférieure) ou non (bande latérale supérieure) en agissant uniquement sur la valeur de la fréquence porteuse Fp.Cette méthode séduisante d'inversion spectrale necessitte cependant dans sa mise en oeuvre des soins particuliers pour obtenir des résultats satisfaisants.Tout d'abord il est necessaire,afin d' avoir une bonne suppression de la bande latérale superieure,d'utiliser des circuits de déphasage linéaires sur toute la bande passante de l'appareil.Ces circuits passe-tout,d'un déphasage constant de 90° sont constuits pour avoir une erreur totale de + ou - 0,6° entre 70 Hz et 13 kHz.Ils sont réalisés selon une méthode décritte par Williams (3) en utilisant les valeurs normalisées de Bedrossian (4).Les multiplieurs doivent avoir également une trés bonne précision.Nous utilisons des multiplieurs intégrés à variation de transconductance AD534L dont l'erreur totale (pleine échelle) est de l'ordre de 0,25% et la non linéarité inferieure à 0,2%.Ils sont cablés dans la configuration d'amplificateur controlé en tension préconisée par le constructeur (5).Nous obtenons ainsi une dynamique superieure à 60 dB et une distorsion totale inferieure à 1% dans la bande passante choisie.Ces caractéristiques autorisent l'utilisation de l'appareil pour la fabrication de stimulus dans le cadre d'experiences de perception sur les paramètres prosodiques.Pour une utilisation pratique de l'inverseur de spectre,il est necessaire de lui adjoindre quelques circuits auxilliaires.Tout d'abord,un oscillateur qui génère la fréquence porteuse.IL est constitué par un générateur de fonction intégré XR2206 choisi pour sa bonne linéarité et son faible taux de distorsion harmonique en sortie sinus.Ensuite,un filtre de pondération particulier qui permet,aprés l'inversion spectrale,de compenser le gain d'énergie dans les hautes fréquences et la perte dans les basses.Ce filtre est réalisé au moyen d'un circuit inspiré des correcteurs de tonalité utilisés sur les amplificateurs électro-acoustiques.Il permet de corriger la sonie du signal inversé par rapport à celle du signal d'entrée.Enfin,un filtre passe-bas permet d'atténuer le signal d'entrée dans les aigus,en particulier pour les signaux de parole,en fonction de la fréquence porteuse Fp.C'est un filtre de Butterworth d'ordre 4 Dont la fréquence de coupure est variable de 4 à 12 kHz.L'appareil fonctionne en temps réel l'inversion de spectre étant instantanée.Nous devons signaler que l'inversion spectrale peut être réalisée facilement au moyen de techniques numériques sur calculateur. Aprés conversion analogique-numérique,il suffit d'inverser la valeur d'un échantillon sur deux (de le multiplier par -1).Cette opération est similaire à la multiplication du signal par une fréquence égale à la moitiée de la fréquence d'échantillonnage (Fe/2).Cette derniére est supprimée en sortie ,ainsi que la bande latérale superieure,par le filtre de reconstitution aprés conversion numérique-analogique.Cette technique necessitte un investissement important,n' est pas d'un fonctionnement instantané,et n'a pas la souplesse du système que nous préconisons.Cependant, elle est plus performante au plan de la dynamique et de la linéarité a condition de disposer d'une résolution de conversion suffisante.

UTILISATION DE L'INVERSEUR DE SPECTRE.

Avant toute manipulation,il est necessaire de définir la bande passante du signal que l'on désire inverser.Celle çi dépend du type d'utilisation.Dans une application didactique,par exemple pour l'acquisition des contours prosodiques d'une langue, on limitte la bande de fréquence à 4 kHz ce qui permet d'atténuer les composantes aigus des consonnes fricatives.La fréquence porteuse Fp est ajustée à une valeur qui dépend de la dynamique de base du locuteur (voix d'homme ou de femme) aux environ de 5 kHz.Cet ajustement peut être éffectué à l'oreille par le manipulateur aprés un trés court entrainement.Il en est de même pour la pondération en fonction de la sonie.Pour la fabrication de bandes de test,une maitrise plus précise des paramètres acoustiques du signal inversé peut être nécessaire,on peut alors les controller avec des moyens d'analyse appropriés.

Bien que peu familiarisé avec ces dernières techniques,nous mentionnons la possibilité de réaliser sur des signaux de musique différents traitements tels que des glissements periodiques de spectre, inversé ou non,au moyen d'une entrée de modulation de la fréquence porteuse,ou,par injection sur l'entrée d'une partie du signal de sortie.Ces effets acoustiques peuvent être plus ou moins plaisants, parfois curieux mais interessants au plan de l'esthétique musicale.

(1) B.Blesser,"Speech Perception under Conditions of Spectral Transformation:1.Phonetic Characteristics",J.S.H.R.Vol 15,5-41,1972;
(2) A.C.Van Nes,"A Speech Spectrum Rotator",I.P.O. Eindhoven,Annual Progress Report,N° 10,83-85,1975.
(3) A.B.Williams,ELECTRONIC FILTER DESIGN HANDBOOK, Chapter 7,1-44,Mc Graw Hill,New York,1981.
(4) S.D.Bedrossian,"Normalized Design of 90° Phase-Difference Networks",I.R.E.Transactions on Circuits Theory,Vol CT-7,128-136,June 1960.
(5) D.H.Sheingold,MULTIPLIER APPLICATION GUIDE,Edit by,Analog Devices Inc,Norwood,40,1978.

INTRODUCTION DU COUPLAGE SOURCE-CONDUIT VOCAL DANS UN SYNTHETISEUR A FORMANTS

Yan Ming CHENG, Bernard GUERIN

INSTITUT DE LA COMMUNICATION PARLEE (UA 368 au CNRS)
I.N.P.G. - 46 avenue Félix Viallet
38031 GRENOBLE CEDEX

ABSTRACT

The effects of source-vocal tract coupling are not taken into account in the classic formant synthesizers. In fact, the naturalness of the synthetic speech may well be improved if this coupling is simulated. The coupling has been modelized according to two methods : (1) inclusion in the vocal source properties, (2) inclusion in the vocal tract acoustic properties. The effects of the coupling are parameterized and included into the command of the parallel formant synthesizer. The improvement is juged by perception tests, and inverse filtering is used to check the effects on an equivalent source. For this study, we have been led to modify the classic linear prediction analysis in order to obtain command parameters for a formant synthesizer with local dynamic variation.

1. Introduction

Une chaine complète de production de la parole contient les poumons, le conduit subglottique, la glotte et le conduit vocal. La glotte, |1|, module le débit la traversant dans le cas de la production des voyelles. Les conduits subglottique et vocal peuvent être considérés comme des systèmes linéaires et invariables tant que des vibrations sonores à l'intérieur sont stationnaires. Ainsi, la chaine totale est un système linéaire et variable temporellement, en négligeant les termes de second ordre et à condition de prendre un système à deux échelles de temps. En effet il est difficile de résoudre directement les équations acoustiques qui décrivent un tel système; on les résoud pratiquement soit par modélisation de la chaine en obtenant une solution numérique, soit par le découplage à la glotte en obtenant une solution empirique. La seconde méthodologie entraine une étude du couplage source-conduit vocal. Le couplage signifie que, premièrement la forme des conduits peuvent modifier l'onde glottique |1|, et deuxièmement, un système découplé étant un système passif excité par une source, l'ouverture glottique modifie la fonction transfert du système passif. La plupart des études du couplage relève de la première approche. Mais leurs résultats expliquent dificilement ce qui s'est passé pour des signaux réels. Enfin, l'application du couplage fait défaut. Ce papier consiste à essayer de représenter les deux approches du couplage dans les cas théorique et pratique. Du fait du nombre limité de pages, nous ne traiterons pas le processus théorique et ne présenterons que les résultats finaux.

Dans le paragraphe 2, nous rendrons compte des phénomènes de couplage et dans le paragraphe 3, on décrira leur introduction dans un synthétiseur à formants à structure parallèle.

2. Les Phénomènes du Couplage Source-Conduit Vocal

Les phénomènes de la production de la parole au niveau de la glotte, avec les effets du couplage, peuvent être représentés selon la figure 1. A l'aide de ce modèle nous pouvons classer les principaux effets du couplage en quatres phénomènes: (1) la dissymétrie de l'onde glottique due à l'inertance des conduits, (2) le "zéro" du spectre de l'onde glottique, coïncidant avec la fréquence du premier formant du conduit vocal, (3) La dépendance de l'ouverture glottique, (4) la superposition d'une ondulation sur l'onde glottique. On va les présenter un par un et étudier leurs effets dans le modèle à deux masses |2|, qui est considéré comme prenant en compte le couplage.

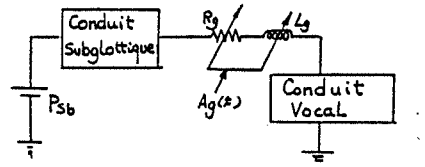


Fig. 1: Modèle simplifié de la source vocale en couplage avec le conduit vocal

La dissymétrie de l'onde glottique, due à l'inertance des conduits, est étudiée dans |3|. Son modèle peut être retrouvé à partir de la figure 1 en exprimant les inertances de la glotte, des conduits subglottique et vocal par Lt, inertance d'ensemble, et en ignorant les résistances et compliances des conduits subglottique et vocal. Sa solution analytique montre que le sommet de l'onde glottique se déplace vers la droite tant que Lt est croissant. Aussi, les voyelles avant doivent être plus dissymétriques que voyelles arrières, car dans le premier cas, Lt est plus élevé que dans le second. On peut faire remarquer que la régle de variation de dissymétrie propoée par |3| est quantitativement beaucoup plus importante que celle issue du modèle à deux masses.

L'existance du "zéro" dans le spectre de l'onde glottique est clairement indiqueé dans |4| |5|. Nous pouvons étudier son effet simplement à partir du modèle de la figure 1. Nous remplaçons le conduit subglottique par une inertance et le

conduit vocal par un circuit parallèle RLC, qui présente le comportement du premier formant du conduit vocal. A l'aide d'une présentation différente un tel circuit, nous obtenons un processus équivalent avec une erreur négligeable, sauf sur un petit laps de temps quand la glotte est fermée. On peut reproduire l'effet du zéro en filtrant la sortie du modèle proposée dans |3| par un antiformant du type de $L_1C_1p^2+L_1p/R_1+1$ (où L_1, C_1 et R_1 sont respectivement l'inertance, la compliance et la résistance; p=jw). Si on considère que la représentation paramétrique de l'onde glottique proposée par Fant |6| peut bien décrire la sortie de modèle proposé par |3|, la solution analytique du processus équivalent montre une diminution de la dissymétrie de l'onde glottique lors de l'insertion de l'antiformant. La diminution est inversement proportionnelle à la fréquence du premier formant. Les résultats d'une simulation numérique sont montrées la figure 2. Sur cette figure on voit aussi une impulsion indésirable au moment de la fermeture de la glotte. Celle-ci est due à la discontinueté à la fermeture et notre representation approximative du zéro.

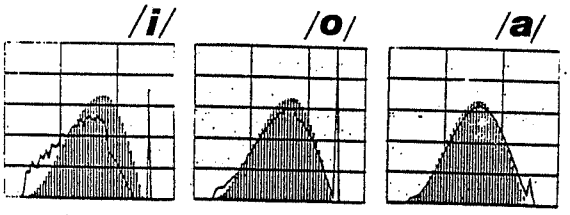


Fig. 2: Effet d'un antiformant sur l'onde de débit. En représentation ombrée, on a représenté le signal issue du modèle de FANT et en trait plein le signal obtenu en tenant compte du zéro.

Il n'y a pas de méthode théorique simple pour évaluer la dépendance de l'ouverture de l'onde glottique avec la charge acoustique. Le modèle à deux masses signale une faible influence sur elle |2|. Il apparaît que, plus basse est la fréquence du premier formant plus quand est quotient de l'ouverture de l'onde glottique. Cette faible influence est probablement due à l'interaction très faible entre les caractéristiques des conduits et le comportement de la vibration mécanique. On peut éssayer de l'évaluer à l'aide "filtre inverse". Notre "filtrage inverse", décrit au paragraphe 3, montre que l'effet du couplage sur l'ouverture glottique est identique à celle du modèle à deux masses.

La conception de la superposition, évoquée dans |7| est définie par l'existence d'une petite ondulation superposée sur la forme de l'onde glottique, grâce à un reste oscillatoire dans le conduit vocal. Nous pouvons donner une expression très élémentaire de la superposition à l'aide du modèle de la figure 1, en supposant que l'ondulation est additive à la forme principale de l'onde glottique. Dans notre déduction, nous négligeons le conduit subglottique et l'inertance glottique, et nous caractérisons le conduit vocal comme indiqué plus haut. Enfin nous supposons que la fréquence de l'oscillation libre du conduit vocal est égale à un nombre entier de fois de la fréquence foundamentale de la source vocale. L'expression est: (Ug_{-super}: ondulation superposée)

$$Ug_{-super}(t)=(P_0\ w/Rg\ \hat{w})e^{-at}\cos(\hat{w}t+\tan^{-1}(a/w))$$

P_0 : L'amplitude du reste oscillatoire

$$a=(R_1+Rg)/(2RgR_1C_1);\ w=1/\sqrt{L_1C_1}\ ;$$

$$\hat{w}=\sqrt{1/(L_1C_1)-((R_1+Rg)/(2RgR_1C_1))^2}$$

Le couplage source-conduit vocal considéré sous la première approche a été brièvement exposé ci-dessus. L'étude selon la deuxième approche dépend fortement de la façon de découpler. Nous adopterons le découplage proposé figure 3, qui implique un système temporellement variable excité par une source du courant constant. L'avantage d'une telle structure est de distinguer facilement les effets de la source et du système par l'analyse des signaux réels. La source est considèré comme fournissant de l'énergie aux formants, les effets du système seront considèrés par les variations de fréquence et bande passante des formants. Son désavantage est l'absence de preuve mathématique de la validité d'un tel découplage. Nous ne pouvons justifier de sa validité que par la qualité des sons synthétisés.

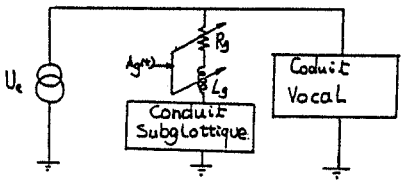


Fig. 3: Modèle de la source vocale à structure parallèle équivalente celui de la figure 1.

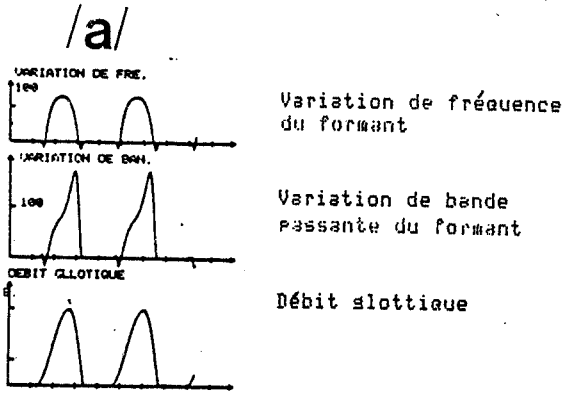


Fig. 4: L'aire d'ouveture de la glotte étant connue, on a calculé la variation de fréquence du premier formant, de la bande passante au cours d'une période en utilisant la modèlisation donnée figure 3.

A partir de la figure 3, en considèrant la présence du conduit subglottique comme une inertance et la présence du conduit vocal comme dans l'étude précédente, nous étudions théoriquement les variations des fréquence et

bande passante du formant, donc le couplage selon la deuxième approche. Un résultat typique d'une telle étude est donné figure 4 pour la voyelle /a/. Il montre que le système couplé peut être représenté approximativement comme un système où les fréquences et bande passantes des formants du conduit vocal sont variable au cours d'une période de l'onde de débit. Nous signalons aussi que la variation de la bande passante va être double et celle de la fréquence va être moitié approximativement, si l'effet du conduit subglottique n'est pas pris en compte. Pour verifier cette étude théorique, nous analysons les signaux réels par le méthode de covariance de la prédiction linéaire avec une très petite fenêtre, de l'ordre de 3 ms, pour une voix d'homme. Les fréquences et bande passantes des formants sont les racines du pôlynome. Une série de résultats typiques est montrée à la figure 5. Les tendances sont les même dans les figures 4 et 5. Les amplitudes des variations sont plus grandes dans la figure 5 que dans la figure 4. On peut en tirer la conclusions suivantes: (1) Les variations, à l'intérieur d'une période des fréquence et bande passante des formants sont d'autant plus faible que l'ordre des formants croît. (2) les variations de fréquence de formant sont plus grandes dans les voyelles arrières que dans les voyelles avants.

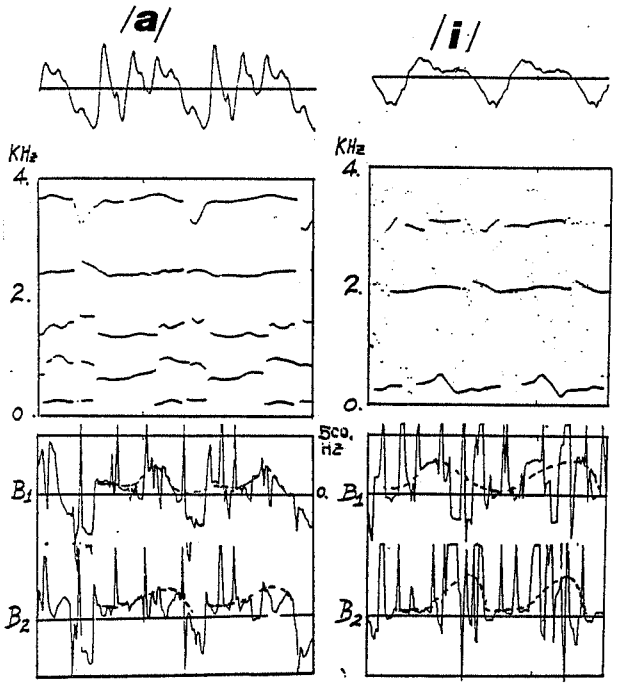


Fig. 5: Résultats de l'analyse LPC et covariance avec une fenêtre très courte dans le cas des voyelle /a/ et /i/

Aprés l'application de la prédiction linéaire sur une très courte fenêtre, la question se pose de savoir quel genre d'information serait fournie par l'étude sur une fenêtre longue. Pour évaluer cette information, nous étudions des signaux réels en faisant varier la longueur de la fenêtre. Nous trouvons que les variations de fréquence et bande passante sont moindres quand la longueur de la fenêtre croît. Nous pouvons donc dire que l'information apportée par l'analyse avec une fenêtre longue décrit la moyenne des variations des caractéristiques de la fonction de transfert. Les résultats peuvent être confortés mathématiquement en ultilissant l'une décomposition de la variation temporelle d'une système selon |8|. Cette proposition nous fournit une stratégie du commande d'une synthétiseur à formants pouvant inclure les effets de couplage source-conduit vocal.

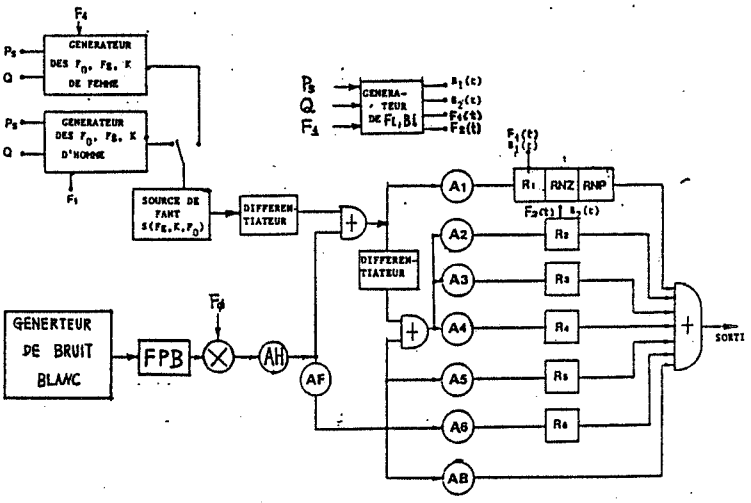


Fig. 6: Structure du synthétiseur à formant.

3. Introduction du Couplage dans un Synthétieur à Formants et Test perceptif du Couplage

Comme nous le décrivions au paragraphe précédent, notre façon de découpler à la glotte n'est pas strictement validée. Nous pouvons la justifier expérimentalement par l'amélioration des sons synthétisés. Des manipulations sont effectuées avec un synthétiseur à formants parallèles, proposé à la figure 6. La source d'excitation produit une onde glottique selon la proposition de Fant, |6|, et est commandée par les paramètres Ps Q (proposées par |4|) et Fl pour obtenir une interaction avec les caractéristiques de source. Le couplage selon la deuxième approche est présenté par les variations des caractéristiques des premier et deuxième formants du conduit vocal. Les amplitudes de variation sont établies à partir des résultats fournis par l'analyse par prédiction linéaire avec une fenêtre longue et ceux de la simulation du conduit vocal à glotte fermée. Le test consiste à demander à l'auditeur de choisir le son le plus naturel dans une paire de sons synthétisés. Seize personnes participent aux tests. La discrimination entre les sons de synthèse avec variations et ceux sans variations des bande passantes du conduit vocal apparait au tableau 1. Il montre une amélioration de la qualité des sons pour toutes les voyelles lorsque le couplage est introduit par

variation de bande passante des formants. La faible discrimination pour /a/ est due à ce que nous avons utilisé dans le cas des bande passantes fixes, une valeur plus adaptée que dans le cas des autres voyelles. Cela veut dire que dans certain cas, on peut trouver des bande passantes effectives |9|, qui diminue la discrimination. Dans un second test, on modifie en plus, la dissymétrie (Q.D.) de l'onde glottique. La discrimination observée est donnée au tableau 2. Nous codons "1" lorsque le coefficient de dissymétrie croit avec la fréquence du premier formant du conduit vocal, comme dans le cas du modèle à deux masses, et nous codons "2" la variation inverse de "1" pour la dissymétrie. Une "dissymétrie désirable", qui est croissante avec |F1-400| (400 Hz est environ la fréquence du premier formant de /o/), peut être obtenue à partir du tableau 2. On peut expliquer ce résultat, sur des signaux réels. Pour les voyelles arrières, à la fermeture de la glotte, on observe une variation très rapide et importante du signal. Jusqu'à présent, on a conservé un quotient d'ouverture glottique (Q.O.) fixe. Or celui-ci est est variable avec la charge. Si on introduit cette nouvelle caractéristique, avec les variation de Q.D. du modèle à deux masses et celle de Q.O., on obtient le tableau 3: nous avons choisi la "dissymétrie désirable" parmi "1" et "2" et nous avons codé "4" comme "1" avec variation de l'ouverture glottique (la quantité de la variation est celle donnée par le modèle à deux masses |2|). Le tableau 3 montre que le naturel des sons est encore meilleur avec la "dissymétrie désirable". En concervant les types de variation de Q.O. et Q.D. fournis par le modèle à deux masses, pour obtenir les mêmes résultats aux tests perceptifs que ceux utilisant une source ayant une variation de Q.D. avec |F1-400|, il faudra exagérer les amplitudes de variation de Q.O. Cette conclusion est conforté par l'analyse par filtrage inverse. Quelque résultats de la sortie du filtrage inverse sont données à la figure 7 (le "Filtrage inverse" est contitué ici de 4 antiformants connectes en cascade, les premier et deuxiéme antiformants ont des variations de fréquences et de bande passantes). On observe clairement une croissance de l'ouverture et la dissymétrie de l'onde glottique avec la décroissance de fréquence du premier formant. Comme nous utilisons une structure parallèle pour le synthétiseur à formants, les effets sur un filtre ne sont pas pris en compte par les autres, contrairement à ce qu'il se passe dans une structure série plus proche de réalité. Pour compenser cette absence, on peut augmenter légèrement l'amplitude des formants d'ordre supérieur. Compte tenu de cette modification et avec des variations de Q.O. et Q.D. conforme à celle du modèle à deux masses, mais pour des amplitudes plus grande de Q.O. (codé par "n"), le résultats du tableau 4 montre que la qualité est la même que celle obtenue avec la "variation désirable" de Q.D. seul.

4.CONCLUSION

Nous pouvons donc conclure que la méthode de découplage de la figure 3, où l'excitation est prise comme l'onde glottique et les caractéristiques du conduit vocal sont variables, est optimal au sens de la synthèse des sons de haute qualité. Pour obtenir la meilleur qualité avec la moindre complixité, il suffira de faire varier le quotient dissymétrie comme l'expression |F1-400|. Ceci n'est valable que pour des opération de synthèse et non pour l'étude de l'onde glottique. Avec un synthètiseur à formant ayant la structure de la figure 6, nous arrivons à synthétiser les consonnes plosives, fricatives et nasales, et des voyelles nasalisés d'une voix d'homme avec la haute qualité en contexte CVCV.

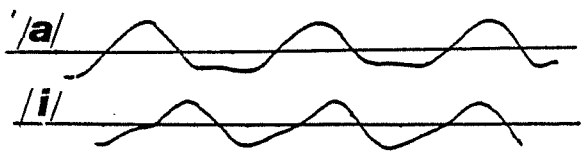


Fig. 7: Onde de débit obtenue par "filtrage inverse".

Tableau 1:

Discrimination	/a/	/i/	/u/	/o/
B.P. Fixe	33	6	0	5
B.P. Variable	67	96	100	95

Tableau 2:

Discrimination	/CaCa/	/CoCo/	/CiCi/
"1"	13.2	50	71
"2"	86,8	50	29

Tableau 3:

Discrimination	/CaCa/	/CoCo/	/CiCi/
"1"			61,1
"2"	71,8	50	
"4"	28,2	50	39,9

Tableau 4:

Discrimination	/CaCa/	/CoCo/	/CiCi/
"1"			50,9
"2"	49,8	46,67	
"n"	50,2	53,33	49,1

|1| K. Ishizaka et J.L. Flanagan (1972),"Synthesis of voiced sounds from a two mass-model of the vocal cords" B.S.T.J., 51, pp. 1233-1268.

|2| Al-Ansari (1981) "Etude des interactions source-cavité subglottique et source-conduit vocal" Thése de Doct.-Ing. INPG.

|3| M. Rothenberg (1981) "An interactive model for the voice source" STL- QPSR, 4/1981, pp. 1-17.

|4| B. Guérin (1978) "Contribution aux recherches sur la production de la parole; Etude du fonctionnement de la source vocale-simulation d'un modèle" Thése de Doct. d'Etat, INPG.

|5| R. Carré (1981) "Vocal source-tract coupling, effets in the vowel spectrum" 4th F.A.S.E. symposium, Venezia, April.

|6| G. Fant (1979) "Glottal source and excitation analysis" STL-QPSR, 3-4/1979, pp. 31-53.

|7| G. Fant et T.V.Ananthpadmanabha (1982) "Truncation and superposition", STR-QPSR 1/1982 pp. 1-17.

|8| T. Claasen et W. Mecklenbräuker (1982) "On stationary linear time-varing systems" IEEE CAS Vol 29 No. 3 March pp. 169-184.

|9| G. Fant et J. Liljencrants (1979) "Perception of vowels with truncation intraperiod decay envelopes" STL-QPSR 1/1979, pp. 79-84.

MODULE TEMPS REEL DE SOURCE VOCALE ELABOREE POUR SYNTHETISEUR A FORMANTS

Ngoc QUACH TUAN, Bernard GUERIN

Institut de la Communication parlée (UA 368 au C.N.R.S)
I.N.P de Grenoble, 46 Avenue de Félix Viallet
38000 GRENOBLE Cédex

.ABSTRACT

The modelisation of the two-mass model of the vocal cord permit to know the evolution of frequency and temporal parameters of the shape of glottal volume velocity wave stemming from the vocal cord. The linear and non-linear approximations give us the relations between these shape's caracteristics and the control parameters of the source : Ps et Q

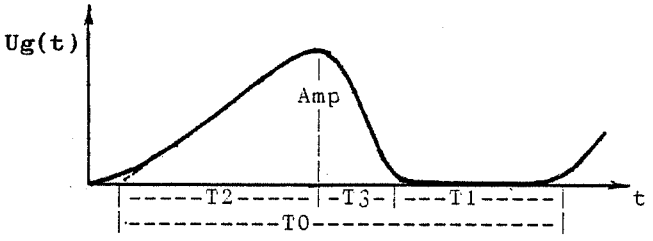
The signal of the glottal volume velocity is modeled by a set of two sinusoidal arcs. A material realisation using a micro-processor monochip INTEL 8031 supplies this signal in real time at sampling frequency of 10 KHz. This source can replace a classic one in the formant synthesizer. A research of integration feasibility of this circuit is on the way.

1.INTRODUCTION

La qualité de la parole de synthèse dépend beaucoup de la forme du signal d'excitation. Auparavant une source impulsionnelle classique suivie de filtres était utilisée dans le synthétiseur à formants pour simuler la source vocale d'une manière approximative. Parmi les différents modèles de la source vocale qui ont été proposées, le modèle à deux masses semble réaliser un bon compromis validité/complexité. La simulation du modèle à deux masses de la source vocale a permis de caractériser la forme de l'onde de débit. Dans le cadre de la synthèse à formants, nous utilisons ces résultats pour élaborer une source fournissant une onde de débit de forme convenable, en temps réel. Cette source de débit glottique a été implantée autour d'un micro-processeur monopuce INTEL 8031. L'utilisation de cette source dans un synthétiseur à formants est intéressante car elle permettra d'introduire unecommande de type physiologique des paramètres de la source de débit, ceci n'étant pas possible dans les modèles proposés jusqu'à présent. Ces paramètres sont au nombre de deux : la pression subglottique Ps et un facteur de masse-tension Q. Ils seront identiques à ceux utilisés dans le cas de la simulation complète de la source vocale à deux masses.

2. MODELE A DEUX MASSES DE LA SOURCE VOCALE. MODELE DE L'ONDE DE DEBIT GLOTTIQUE PROPOSE PAR G.FANT .

L'utilisation de modèle à deux masses dévelopé d'abord par ISHIZAKA et FLANAGAN |1| et puis B.GUERIN |2| a montré la bonne



Q.O. = (T2+T3)/T0 et Q.D. = T2/T3

Figure 1 : Caractérisation de la forme de l'onde de débit .

adéquation du comportement de ce modèle avec les données obtenues sur la parole naturelle. Nous savons que l'onde de débit, issue de la source vocale, pouvait être caractérisée par sa frequence fondamentale F0 (ou bien la période T0) , son quotient d'ouverture Q.O., son quotient de dissymétrie Q.D., son amplitude de crête Amp et sa puissance P. La forme générale de l'onde de débit est donnée sur la figure 1.

Il a été montré (|2|,|3|) que la commande de ce modèle pouvait se ramener à deux paramètres indépendants : la pression subglottique Ps et un facteur Q déterminant

les masses et les tensions des cordes vocales. Dans les conditions normales de phonation, Ps se varie entre 2 et 20 cm H_2O et le facteur Q se varie entre 1 et 3. Les simulations de ce modèle ont permis de connaître l'évolution des paramètres fréquentiels et temporels de la forme de l'onde de débit glottique. Ces relations sont illustrées sur la figure 2 (page 3). Les approximations linéaires et non-linéaires vont être décrites et discutées .

La forme de l'onde de débit peut être représentée par une association de fonctions sinus comme le propose G.FANT |4| (Figure 3).

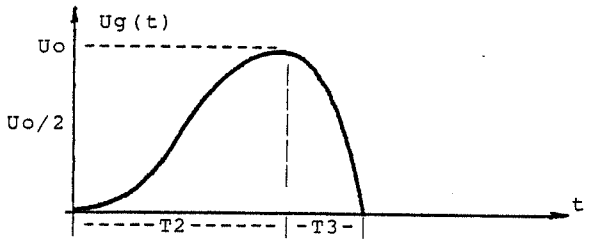


Figure 3: Description de la forme de l'onde d'après G.FANT.

Avec : PI=3,14

$$Ug(t) = \begin{cases} = \frac{1}{2}.Uo.\left[1 - \cos\frac{PI}{T2}.t\right] & t \quad T2 \\ = Uo.\left[k.\cos\frac{PI}{T2}.t - k + 1\right] & t \quad T3 \end{cases}$$

Ce type d'onde de débit peut être utilisé pour exciter un synthétiseur à formants car la représentation de l'onde avec la forme exacte, comme celle de sortie du modèle à deux masses, n'est pas nécessaire. De ce fait, cette forme va être déterminée à partir des relations décrites sur la figure 2.

3.REALISATION D'UNE SOURCE VOCALE DE DEBIT FONTIONNANT EN TEMPS REEL.

Pour calculer les caractéristiques de l'onde de débit glottique à partir des résultats simulés sur le modèle à deux masses avec les deux paramètres de commandes, Ps et Q, nous avons utilisé les approximations linéaires et non-linéaires suivantes |5|,|6|:

$$FO(Hz) = 0,33.Ps + 42,3.Q - 3,8$$

$$Uo(m^3/s) = 10^{-3}.\left|\frac{0,145}{Q} + 0,039\right|.|Ps-980|$$

$$Q.O. = Q.O.(Ps)_{|Q=1} + 0,031.(Q-1)$$

$$Q.D. = Q.D.(Ps)_{|Q=1} + 0,15.(Ps-1960).(Q-1)$$

où Ps est en Pascal,

et $Q.O.(Ps)_{|Q=1}$ et $Q.D.(Ps)_{|Q=1}$

représentent les courbes de Q.O. et Q.D. corespondant à Q=1. Une approximation linéaire de Q.O. et Q.D. aurait été plus simple mais cela entraine une différence non négligeable entre les résultats ainsi calculés et ceux fournis par le modèle à deux masses. L'approximation non-linéaire semble complexe mais en réalité, elle est très simple à réaliser : les courbes $Q.O.(Ps)_{|Q=1}$ et $Q.D.(Ps)_{|Q=1}$ sont tabulées en mémoire. Cela nous emmène à calculer T0, Q.O. et Q.D. directement à partir de Ps et Q, on en déduit alors :

$$T3 = \frac{T0.Q.O.}{1 + Q.D.}$$

$$T2 = T0.Q.O. - T3$$

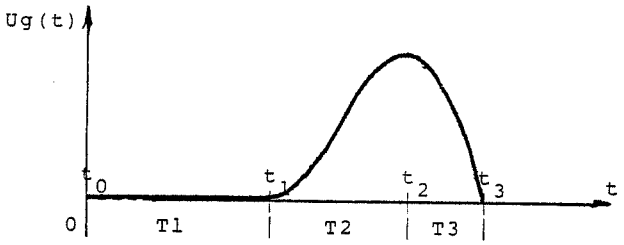


Figure 4: Définition des paramètres Ti, t_i

La forme de l'onde de débit glottique est donnée par |6|:

$$Ug(t) = \begin{cases} = 0 & \text{quand} \quad t \quad T1 \\ = Uo.\dfrac{1 - \cos\frac{PI}{T2}.(t - t_1)}{2} & \text{quand} \quad t \quad T2 \\ = Uo.\left[1 - \dfrac{1 - \cos\frac{PI}{T2}(t - t_2)}{1 - \cos(\frac{PI}{T2}.T3)}\right] & \text{quand} \quad t \quad T3 \end{cases}$$

soit

$$\Theta(x) = \frac{1 - \cos\frac{PI}{T2}.x}{2}$$

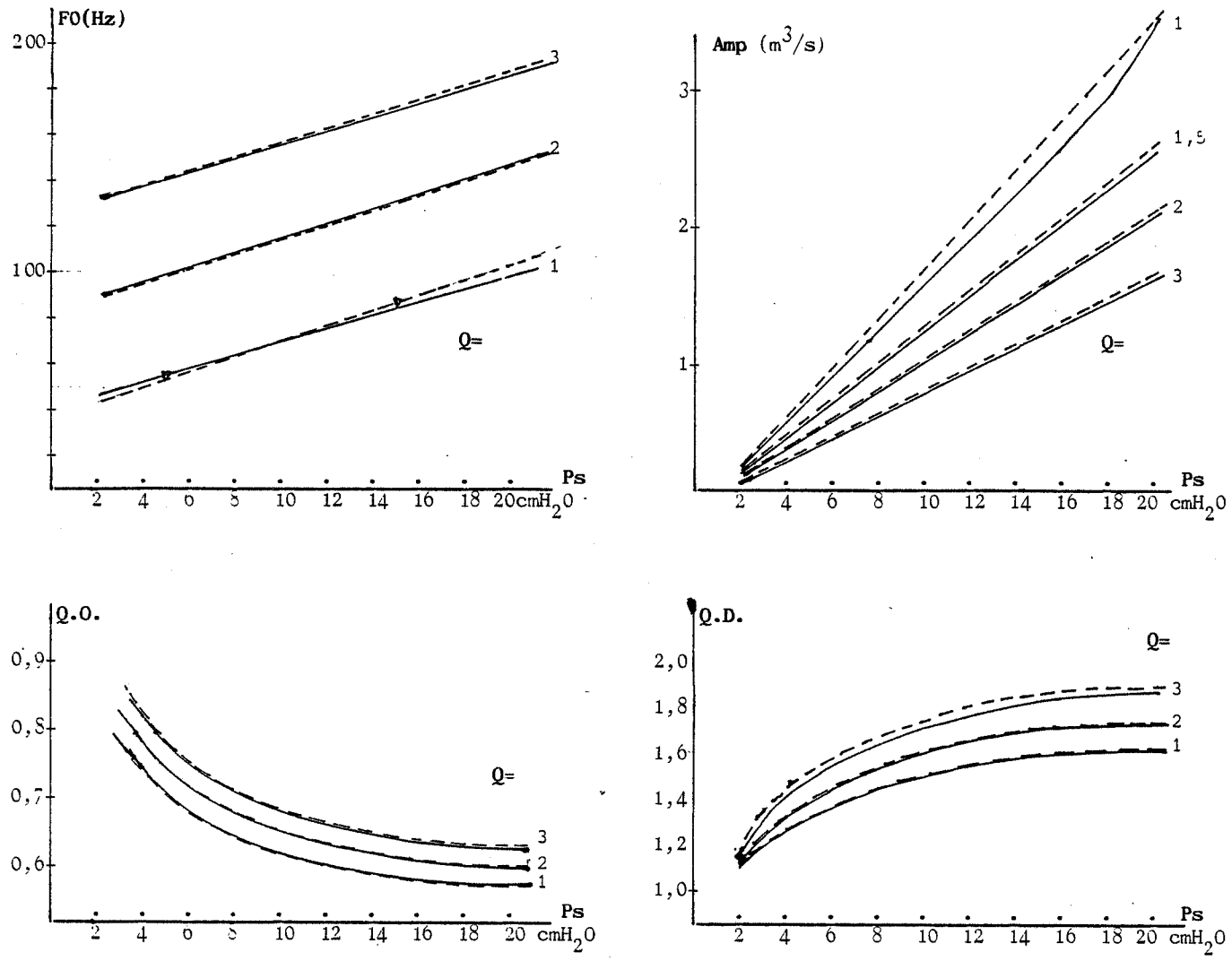


Figure 2 : Evolution des caractéristiques temporelles de l'onde de débit en fonction de Ps et Q:
En trait plein: Modèle à deux masses,
En pointillé : Modèle de débit.

Nous avons :

$$Ug(t) = \begin{cases} = 0 & \text{quand } t \quad T1 \\ = Uo.\Theta(t-t_1) & \text{quand } t \quad T2 \\ = Uo.\left[1 - \dfrac{\Theta(t-t_2)}{\Theta(T3)}\right] & \\ & \text{quand } t \quad T3 \end{cases}$$

L'interêt de cette formule est que nous ne devons calculer la fonction $\Theta(x)$ qu'une fois pour s'en servir 3 fois: dans l'intervalle t_1-t_2 : $\Theta(t - t_1)$ et dans l'intervalle t_2-t_3 : $\Theta(t - t_2)$ et $\Theta(T3)$.

De ce fait, nous n'utilisons plus le coefficient k dans la formule de l'onde proposé par G.FANT, |6|. Avec cette technique, il est certain que nous obtenons les bonnes approximations malgré les propriétés non-linéaires des caractéristiques obtenues par la simulation du modèle à deux masses. Sur la figure 2 sont aussi données les caractéristiques FO, Amp, Q.O. et Q.D. de l'onde de débit ainsi calculée.

Cette source de l'onde de débit glottique a été implantée sur un micro-ordinateur monopuce INTEL 8031 dont le schéma est donne sur la figure 5.

Pour illustrer les performances du processeur, indiquons que le temps d'éxcution d'une instrution est de 1 µs à 2 µs sauf pour la multiplication où il est

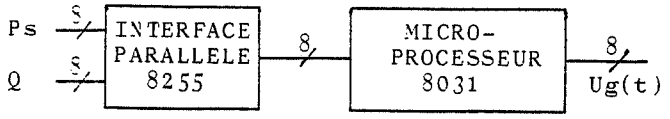


Figure 5: Schéma électronique de la soucre de l'onde de débit glottique.

de 4 µs. Cette ensemble a fourni le signal de débit en temps réel à une fréquence d'échantionnage de 10 KHz. De nombreuses opérations de divisions étaient nécessaires; pour atteindre le temps réel, avons nous dû réaliser la division par la multiplication de l'inverse du diviseur par le dividende , le diviseur étant tabulé (données sur 8 bits). Le déroulement du calcul s'effectue de la manière suivante: les deux paramètres Ps et Q sont lus à l'instant t_0 et puis nous profitons de l'intervalle t_0-t_1 où le signal Ug(t) est nul pour calculer d'abord les valeurs de F0, Q.O., Q.D., Uo et puis T0, T1, T2, T3; puis à l'avance, les premiers échantillons non-nuls dans l'intervalle t_1-t_3. Ainsi, le dernier échantillon non-nul doit-il être obtenu au plus tard à l'instant t_3.

4. CONCLUSION

La source de débit basée sur les caractéristiques du modèle à deux masses peut être connectée avec le synthétiseur à formants numériques fonctionnant en temps réel. Des tests de perception ont été conduits pour évaluer l'amélioration de qualité d'une telle source |7|. Dans tous les cas les résultats ont été positifs: les sons produits avec cette source élaborée sont jugées plus "naturels" que ceux produits avec une source impulsionnelle classique. Cela confirme l'interêt de pousuivre des études pour affiner la forme de l'onde d'excitation dans les synthétiseurs à formants.

REFERENCES

|1|.ISHIZAKA K., FLANAGAN J.L,
"Synthesis of voiced sounds from a two mass model of the vocal cords",
B.S.T.J., vol 51, pp 1233_1268,
July August 1972.

|2|.GUERIN.B,
"Contribution aux recherches sur la production de la parole. Etude du fonctionnement de la source vocale. Simulation d'un modèle".
Thèse de doctorat ès-sciences physiques à I.N.P de Grenoble et USMG, Grenoble, 1978.

|3|.GUERIN.B, BOE L.J.
"A two mass model of the vocal cords: determination of control parameters and their respective consequences".
I.E.E.E Int. Conf. on Acoust., Speech and Signal Processing, Hartford, pp 583-586, 1977.

|4|.FANT.G,
"Vocal source analysis. A progress report",
STL-QPSR 3/4, pp 31-53

|5|.AL-ANSARI A.
"Etude du fonctionement et simulation en temps réel d'un modèle de la source vocale. Etude des interactions source-cavité subglottique et source-conduit vocal".
Thèse de docteur ingénieur à I.N.P de Grenoble, Decembre 1981.

|6|.QUACH TUAN N.
"Source vocale pour synthétiseur à formants".
DEA à I.N.P de Grenoble, Septembre 1983.

|7|.AWAD S.
"Synthétiseur à formants en temps réel. Etude d'une source d'excitation élaborée. Codage optimum des paramètres de commande".
Thèse de docteur ingénieur à I.N.P de Grenoble, Decembre 1983.

Etude comparative des détecteurs du fondamental selon le principe de transformation spectrale double - premiers résultats

Helge Indefrey, Wolfgang Hess, et Günter Seeser

Lehrstuhl für Datenverarbeitung, Technische Universität München

Postfach 202420, D-8000 München 2, R.F.A.

Abstract. *The performance of double-transform pitch determination algorithms with frequency-domain nonlinear distortion is investigated. This principle comprises well-known algorithms, such as the autocorrelation algorithm or the cepstrum algorithm. Besides these two, the amplitude-spectrum and the fourth-root spectrum algorithms were implemented and tested under various environmental conditions (clean signal, telephone-quality signal, and signal degraded by additive Gaussian noise). First results suggest that the fourth-root spectrum and the amplitude spectrum algorithms are less sensitive to noise than the cepstrum algorithm and less sensitive to strong first formants than the autocorrelation algorithm.*

1. Introduction

Le fondamental (c'est-à-dire, la fréquence fondamentale F_0 même que la période fondamentale T_0) prend une position très importante dans le signal parole. Plusieurs problèmes de traitement de la parole dépendent décisivement de la qualité de la mesure du fondamental; ça vaut particulièrement pour les systèmes vocodeur.

Littéralement des centaines d'algorithmes pour la détermination du fondamental ayant été développés [voir, par exemple, Hess (1983)], les études comparatives de leur fonctionnement n'ont pas été très nombreuses. Pour cela, les raisons principales sont 1) l'effort nécessaire afin de générer une base de données contenant une courbe de référence (voir Sect.3), et 2) la difficulté d'établir un standard pour l'analyse d'erreurs.

Dans cette expérience on étudie le principe de la détermination du fondamental par transformation spectrale double et distorsion non linéaire dans le domaine fréquentiel. Ce principe contient plusieurs analyseurs[1] bien connus (voir Sect.2) et est assez facile à réaliser. La courbe de référence est obtenu par un analyseur spécial (Hess et Indefrey, 1984) qui utilise le signal de sortie d'un laryngographe.

2. Les analyseurs examinés dans cette étude

Par définition un analyseur se compose de trois étages (McKinney, 1965; Hess, 1983): 1) le prétraitement, 2) l'extracteur, et 3) le postprocesseur. L'extracteur exécute le mesurage propre: il convertit le signal d'entrée en une séquence de valeurs estimatives du fondamental. L'étage de prétraitemant fait une réduction d'information afin de faciliter le travail de l'extracteur. Le postprocesseur, enfin, peut corriger des erreurs, faire un lissage de la courbe de mélodie, ou autre chose exigée par l'application réspective.

On peut grouper les analyseurs dans deux grandes catégories (Hess, 1983): 1) analyseurs en domaine temporel, et 2) détermination du fondamental par analyse à terme court. En domaine temporel, les analyseurs examinent le signal période par période; ils sont capables de délimiter des périodes individuels. D'autre part, les analyseurs à terme court, divisant le signal en une séquence des intervalles (trames) à terme court successifs, quittent le domaine temporel par une transformation à terme court (pas nécessairement une transformation de Fourier). On désire que cette transformation réunisse toute l'information sur le fondamental, distribuée sur le trame dans le signal original, en un seul pic signifiant dans le domaine où l'extracteur fonctionne, c'est-à-dire, le domaine fréquentiel ou le domaine de décalage. Parmi les analyseurs à terme court, le principe de la transformation spectrale double avec distorsion spectrale non linéaire est assez fréquemment utilisé. Un analyseur fonctionnant selon ce principe se compose des étages suivants:

1) Prétraitement linéaire ou non linéaire en domaine temporel (optionnel);
2) transformation de Fourier discrète;
3) distorsion non linéaire spectrale;
4) transformation de Fourier inverse; et

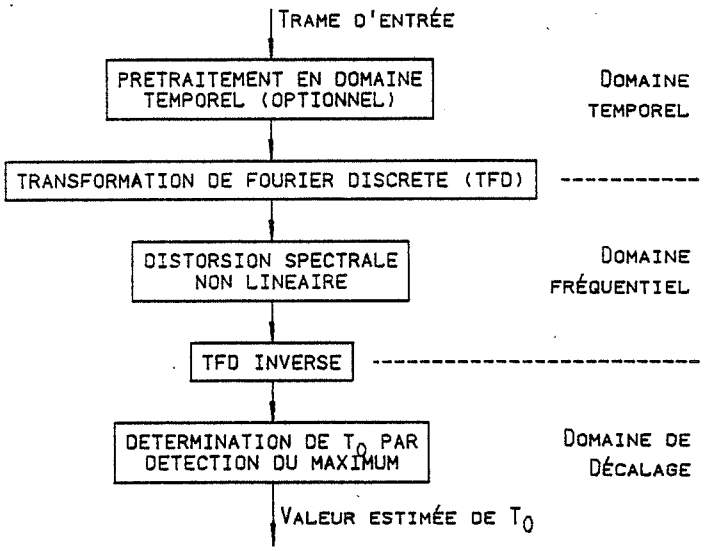


Fig.1. Schéma synoptique d'un analyseur selon le principe de transformation spectrale double avec distorsion non linéaire en domaine fréquentiel

[1] Pour raison de brévité, tout algorithme de détermination du fondamental est brièvement indiqué comme **analyseur** dans cette communication. L'analyseur utilisant la quatrième racine de spectre est indiqué comme **analyseur dit QR**, et l'analyseur utilisant le spectre d'amplitude est brièvement indiqué comme **analyseur dit AM.**

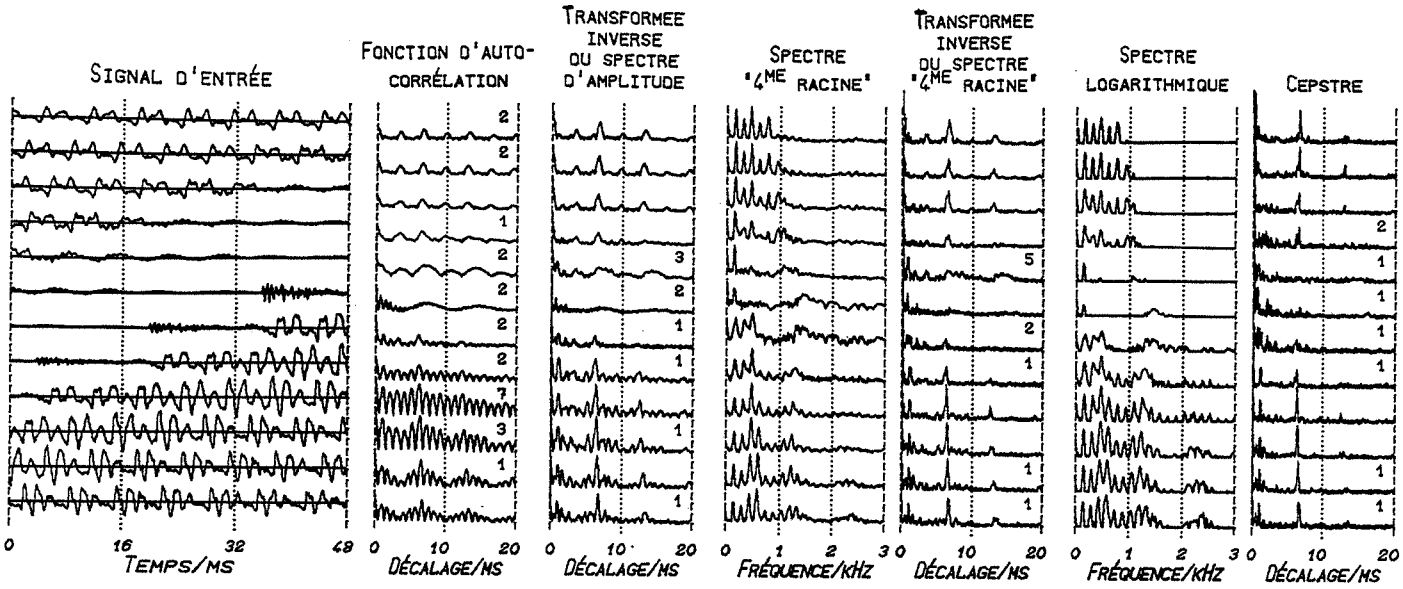


Fig.2. Détermination du fondamental par transformation spectrale double et distorsion non linéaire dans le domaine fréquentiel: exemple du fonctionnement. Le temps procède de haut en bas; distances entre deux trames successifs: 16 ms. Quatre fonctions non linéaires sont montrées. Les nombres à droite d'un part des transformées inverses indiquent combien de maxima parasitaires (excédant 70% de l'amplitude du maximum signifiant) existent dans la gamme de mesurage (1-20 ms). Signal: transition /u-d-o/, locuteur HGI (masculin), bonne qualité, pas de prétraitement en domaine temporel

5) determination de T_0 d'un maximum signifiant dans le domaine de décalage à partir de la fonction obtenue dans étage 4.

Les étages 1 à 4 faisant le prétraitement, l'étage 5 constitue l'extrateur.

Ce schéma contient plusieurs analyseurs bien connus avec bon fonctionnement, comme l'analyseur d'autocorrélation (Sondhi, 1968; Markel, 1972; Rabiner, 1977), où on prend le carré comme distorsion spectrale non linéaire, ou l'analyseur cepstral (Noll, 1967), où on prend le logarithme du spectre. De plus, on a proposé dans la littérature l'usage du spectre d'amplitude (Weiss et al., 1966) ou de la quatrième racine du spectre de puissance (Sreenivas, 1981).

Quant au prétraitement (optionnel) en domaine temporel, on préfère l'écrêtement central (Sondhi, 1968; Rabiner, 1977) ou le filtrage inverse linéaire (voir, par exemple, Markel, 1972; Fujisaki et Tanabe, 1972).

Dans notre système on a réalisé quatre fonctions spectrales non linéaires: le spectre de puissance (correspondant à l'autocorrélation), le spectre d'amplitude, la quatrième racine du spectre de puissance, et le logarithme (correspondant au cepstre). La figure 2 présente un exemple. Trois méthodes de prétraitement en domaine temporel ont été appliquées: 1) pas de prétraitement, 2) écrêtement central (Sondhi, 1968), et 3) filtrage inverse linéaire.

En fait le filtrage inverse est réalisé dans le domaine fréquentiel. Au lieu de filtrer le signal en domaine temporel on peut multiplier le spectre du signal par celui du filtre inverse. Contraire à la réalisation en domaine temporel qui exige de coéfficients explicits du filtre, une procédure quelconque peut être appliquée en domaine fréquentiel, supposé qu'elle donne un spectre utilisable. Dans notre système on a réalisé la méthode proposée par Fujisaki et Tanabe (1972). Une fenêtre très courte (durée optimale: 1.5 ms) est positionnée autour d'un pic signifiant du signal dans l'intervalle d'analyse. Cette fenêtre préserve la plupart de l'information sur les formants, mais elle ne contient plus d'information sur le fondamental. Son spectre donne donc une estimation approximative bien lissée de la fonction de transfert du conduit vocal et devient utilisable comme spectre (réciproque) du filtre inverse.

En combinant les 3 méthodes du prétraitement avec les 4 méthodes de distorsion spectrale on obtient 12 analyseurs différents. T_0 est toujours déterminé comme location du plus grand pic de la fonction en domaine de décalage dans la gamme de mesurage (1-20 ms). Parce qu'on n'a appliqué aucun posttraitement, les données présentées en Sect.5 sont les valeurs estimées comme on les obtient à la sortie de l'extracteur.

3. Courbe de référence et base de données

Lorsqu'on évalue le fonctionnement d'un instrument ou algorithme de mesurage, il faut qu'on dispose d'un autre instrument qui fonctionne mieux ou au moins aussi bien que celui sous évaluation. Comme aucun analyseur ne fonctionne sans faire d'erreurs, il n'y a que deux possibilités de venir à bout de ce problème: 1) l'usage d'un analyseur interactif avec correction manuelle d'erreurs (Rabiner et al., 1976), et 2) l'usage d'un instrument que dérive l'information sur le fondamental directement à partir des vibrations laryngiennes. Ici on trouve par exemple l'accéléromètre (Viswanathan et Russell, 1984) ou le laryngographe. Une autre possibilité est d'utiliser des signaux synthétiques le fondamental desquels est exactement connu (Dal Degan, 1982).

Dans notre étude la courbe de référence a été obtenue via le signal de sortie d'un laryngographe (Fourcin et Abberton, 1971). Pour ce but un analyseur simple mais exacte a été developpé (Hess et Indefrey, 1984). Cet analyseur utilise le point d'inflexion pendant la pente ascendante rapide du laryngogramme comme valeur estimée de l'instant de la clôture glottale (fig.3).

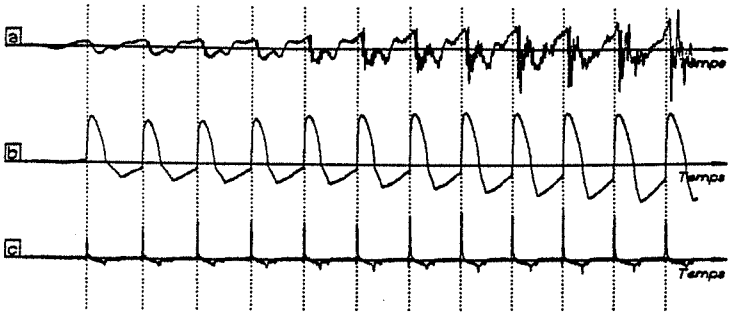


Fig.3a-c. Signal parole (a), laryngogramme (b), et laryngogramme différencié (c). Les marqueurs délimitant les périodes fondamentales individuelles ont été dérivés à partir des maxima de (c). Signal: debut de la voyelle /a/, locuteur WGH (masculin)

La base de données pour cette étude se compose du texte allemand suivant: "Jawohl, hören Sie! Ich bin Rudolf Ranick hier vom FTZ. Prüfen Sie bei Kurt Meier in der Burgstraße den Leitungsanschluß und auch das Geräusch aus den Kapseln!", lu par 4 locuteurs (1 féminin, 3 masculins). Les signaux ont été enregistrés dans une chambre sourde avec une fréquence d'échantillonnage de 16 kHz. Afin de tester la robustesse des analyseurs, 4 versions du signal d'entrée ont été appliquées: 1) signal à bonne qualité (comme enregistrée); 2) signal limité à la bande téléphonique (0,3-3,4 kHz); 3) signal dégradé par bruit gaussien additif, rapport global signal/bruit 0 dB; 4) signal dégradé par bruit et limité à la bande téléphonique. Pour la limitation de bande les signaux ont été filtrés par un filtre numérique passe-bande à phase linéaire d'un ordre de 240.

4. Analyse d'erreurs de la détermination du fondamental

Selon l'étude classique par Rabiner et al. (1976), les analyseurs commettent quatre types d'erreurs: 1) erreurs grosses de la détermination du fondamental; 2) inexactitudes de la détermination du fondamental; et 3) détermination d'un segment voisé comme non voisé, et 4) détermination d'un segment non voisé comme voisé. On a gardé cette catégorisation dans des études suivantes (El Mallawany, 1977; Viswanathan et Russell, 1984).

Les erreurs grosses de determination du fondamental sont "drastic failures of a particular method or algorithm to determine pitch" (Rabiner et al., 1976). Il est de règle qu'on regarde une erreur comme grosse si la déviation entre la valeur estimée et la valeur correcte excède le taux de change que la voix peut produire sans devenir irregulière. Dans la littérature ce seuil théorique est normalement remplacé par un seuil plus pragmatique, comme, par exemple, une différence en T_0 de plus que 1 ms (Rabiner et al., 1976) ou une déviation relative de plus que 10 % (Viswanathan et Russell, 1984). Dans notre étude le seuil proposé par Viswanathan et Russell (1984) a été appliqué.

Les inexactitudes de mesure du fondamental causent du "bruit" sur la courbe de mélodie. Ils sont des petites déviations entre la valeur estimée et la valeur correcte. On les décrit par leur valeur moyenne et l'écart type (Rabiner et al., 1976; Viswanathan et Russell, 1984).

Les erreurs "voise/non voisé" et "non voisé/voisé" ne sont pas vraiment des erreurs de la détermination du fondamental; au contraire, il s'agit d'erreurs de la détection du voisement. Comme cette étude a le but de developper des critères pour l'optimalisation des analyseurs par variation systématique des composantes individuelles, on ne regarde que les erreurs de la détermination du fondamental.

Si l'on a le problème d'évaluer un analyseur subjectivement et objectivement, il faut qu'on sache quels erreurs sont les plus ennuyantes pour l'ouie humaine. Dans l'étude par Rabiner et al. (1976) il y avait une certaine différence entre le mesurage objectif et le jugement subjectif. Viswanathan et Russell (1984) ont montré qu'on peut obtenir une bonne corrélation entre le taux objectif d'erreurs et le jugement subjectif pour les erreurs grosses de la détermination du fondamental aussi que pour les erreurs de la détection du voisement lorsqu'on les pondère avant de les compter. Cependant, pour les résultats préliminaires de cette etude il nous a suffit de seulement compter les erreurs grosses et de présenter leur pourcentage par rapport àu nombre total des trames voisés.

5. Résultats et conclusions

Le tableau 1 présente le pourcentage des erreurs grosses pour les analyseurs, signaux, et locuteurs différents. Les locuteurs masculins se comportant à peu près pareillement, la locutrice présente quelques différences.

Le degré du nivellement spectral est un critère important pour le fonctionnement de ce type d'analyseur (Sondhi, 1968). L'analyseur d'autocorrélation ne fait aucun nivellement spectral; pour cette raison il requiert un prétraitement en domaine temporel, autrement il devient sensible à des formants forts (Rabiner, 1977;

Tableau 1. Taux d'erreurs pour les algorithmes, locuteurs, et conditions d'environnement différents. Les nombres présentent les taux d'erreurs en unités de 0.1 % (à cause d'espace). Distorsions spectrales: (CE) logarithme (cepstre), (QR) quatrième racine du spectre de puissance, (AM) spectre d'amplitude, (AC) spectre de puissance (autocorrélation). Prétraitement en domaine temporel: (LI) linéaire (pas de prétraitement), (EC) écrêtement central, (FI) filtrage inverse

	Loc. masc.			Loc. fem.			Tous locuteurs		
	LI	EC	FI	LI	EC	FI	LI	EC	FI
	--- Signal original ---								
CE	14	24	8	15	30	11	14	25	9
QR	7	20	12	14	26	11	9	22	12
AM	7	19	22	26	27	19	12	21	22
AC	15	20	101	36	25	60	20	21	91
	--- Signal à bande limitée (0,3-3,4 kHz) ---								
CE	46	36	26	28	42	19	41	38	24
QR	22	27	27	26	35	23	23	29	26
AM	33	27	46	30	29	22	32	28	40
AC	47	26	122	41	28	61	45	27	107
	--- Signal avec bruit additif ---								
CE	112	148	42	198	276	125	133	180	63
QR	60	125	47	165	245	122	86	155	66
AM	43	110	101	125	208	129	64	135	108
AC	35	92	255	97	181	197	50	115	241
	--- Signal à bande limitee avec bruit ---								
CE	435	236	303	153	234	122	364	236	258
QR	216	204	187	166	201	153	203	203	179
AM	160	178	207	146	174	173	156	177	198
AC	147	167	273	133	152	219	144	163	259

voir fig.2). En prenant le logarithme du spectre, l'analyseur cepstral fait un nivellement spectral assez fort. L'algorithme dit QR fonctionne semblablement outre que les valeurs du spectre distordu tendent vers zero et pas vers $-\infty$ quand les amplitudes spectrales deviennent très basses. Quant à la sensitivité au bruit, il est à supposer que l'analyseur d'autocorrélation fonctionne le mieux parce qu'il n'augmente pas les composantes du spectre avec des amplitudes basses.

Pour le signal à bonne qualité tous les analyseurs fonctionnent bien sauf l'analyseur d'autocorrélation avec filtrage inverse qui a toujours fait plus d'erreurs que les autres. L'écrêtement central a un peu amélioré l'analyseur d'autocorrélation pour la locutrice tandis qu'il dégrade cet analyseur un peu pour les locuteurs masculins. L'analyseur dit QR et l'analyseur cepstral avec filtrage inverse fonctionnent le mieux. Pour le signal à bande limitée (l'influence de la limitation de bande étant plus grande pour les locuteurs masculins que pour la locutrice), le nombre d'erreurs grosses s'est augmenté par un facteur de 3 en moyenne. L'écrêtement central améliore le fonctionnement de l'analyseur d'autocorrélation tandis qu'il n'affecte beaucoup les autres. De nouveau l'analyseur dit QR et l'analyseur cepstral avec filtrage inverse fonctionnent le mieux.

Tandis que la plupart des analyseurs ne fonctionnent plus pour le signal à bande limitée avec bruit additif, le signal à large bande avec bruit est traité le mieux par l'analyseur d'autocorrélation sans prétraitement en domaine temporel (!) quoique le taux d'erreur ait été cinq fois plus grand que celui de l'analyseur le meilleur pour le signal à bonne qualité. Le cepstre est un peu sensible au bruit excepté si l'on applique le filtre inverse. Les analyseurs dits QR et AM fonctionnent assez bien. C'est toujours que l'écrêtement central dégrade le fonctionnement.

Comme déjà dit, ces résultats ne sont que préliminaires. Un grand problem prâtique dans cette étude est la dépense de calcul. Parce qu'on n'a pas de processeur rapide de signaux, et parce que les conditions opératoires de l'expérience exigent des transformations rapides de Fourier à 1024 points (dont chacune réquiert 0,8 s sur notre PDP-11/45 même quoiqu'on dispose d'une routine très élaborée), il faut trois heures afin d'obtenir un résultat (un locuteur, une condition d'environnement, une méthode de prétraitement, et quatre méthodes de distorsion spectrale). Néanmoins, il faut rechercher au moins quatre locuteurs additionnels afin que les résultats deviennent assez signifiants.

Une question encore pendante est celle de la robustesse au bruit. Des expériences précédentes (Noll, 1970; Weiss et al., 1966) recommandent que l'analyseur cepstral est plus sensible au bruit que, par exemple, l'analyseur d'autocorrélation. Bien que cet énoncé ait été confirmé par nos expériences pour les analyseurs sans prétraitement en domaine temporel, le filtrage inverse fortement améliore le fonctionnement de l'analyseur cepstral pour les signaux avec bruit. Des expériences plus récentes nous ont montré que le fonctionnement de tous les analyseurs (sauf ceux qui appliquent l'écrêtement central) n'est guère dégradé tant que le rapport signal-bruit est mieux que 3 dB.

La deuxième question pendante est celle du filtrage inverse. Dans nos expériences nous avons appliqué la méthode proposée par Fujisaki et Tanabe (1972) qui ne fonctionne pas bien pour l'autocorrélation. Néanmoins, plusieures études précédentes (Markel, 1972; Rabiner et al., 1976) ont montré que le filtrage inverse via prédiction lineaire fonctionne bien pour l'analyseur d'autocorrélation; cette méthode sera donc inclue dans des expériences à venir.

En résumé, la détermination du fondamental par transformation spectrale double et distorsion spectrale non linéaire contient encore des possibilités prometteuses. Deux analyseurs, ceux d'autocorrélation et du cepstre, sont bien connus et reconnus. Les deux autres analyseurs qui ont été examinés dans notre étude, ce sont les analyseurs dits QR et AM, sont au moins équivalents. Ces analyseurs sont moins sensibles au bruit que l'analyseur cepstral et moins sensibles à des formants forts que l'analyseur d'autocorrélation; ils donc représentent un bon compromis lorsqu'on ne connait pas bien les conditions d'environnement du signal.

Références

Dal Degan N. (1982): "Vocoder quality: an automatic procedure to measure the performance of pitch extractors." In **Proceedings, Globecom'82**

El Mallawany I.I. (1977): "Evaluation comparative de mélographes numériques." Recherches/Acoustique **4**, 15-34 (CNET, F-22301 Lannion)

Fourcin A.J., Abberton E. (1971): "First applications of a new laryngograph." Medical and Biological Illustration **21**, 172-182

Fujisaki H., Tanabe Y. (1972): "A time-domain technique for pitch extraction of speech." Ann. Rept. Eng. Res. Inst. Tokyo **31**, 259-266

Hess W.J. (1983): **Pitch determination of speech signals - algorithms and devices** (Springer, Berlin)

Hess W.J., Indefrey H. (1984): "Accurate pitch determination of speech signals by means of a laryngograph." **Proc. ICASSP-84** (Paper 18B.1)

Markel J.D. (1972): "The SIFT algorithm for fundamental frequency estimation." IEEE Trans. AU-**20**, 367-377

McGonegal C.A., Rabiner L.R., Rosenberg A.E. (1977): "A subjective evaluation of pitch detection methods using LPC synthesized speech." IEEE Trans. ASSP-**25**, 221-229

McKinney N.P. (1965): "Laryngeal frequency analysis for linguistic research" (Commun. Sci. Lab., Univ. of Michigan, Ann Arbor, MI; Res. Rept. #14)

Noll A.M. (1967): "Cepstrum pitch determination." J. Acoust. Soc. Am. **41**, 293-309

Noll A.M. (1970): "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate." **Symposium on Computer Processing in Communication;** ed. by the Microwave Institute (Univ. of Brooklyn Press, New York NY), Vol.19, 779-797

Rabiner L.R. (1977): "On the use of autocorrelation analysis for pitch detection." IEEE Trans. ASSP-**25**, 24-33

Rabiner L.R., Cheng M.J., Rosenberg A.E., McGonegal C.A. (1976): "A comparative study of several pitch detection algorithms." IEEE Trans. ASSP-**24**, 399-413

Sondhi M.M. (1968): "New methods of pitch extraction." IEEE Trans. AU-**16**, 262-266

Sreenivas T.V. (1981): Pitch estimation of aperiodic and noisy speech signals (Diss., Dept. of Electrical Eng., Indian Inst. of Technology, Bombay)

Viswanathan V.R., Russell W.H. (1984): Subjective and objective evaluation of pitch extractors for LPC and harmonic-deviations vocoders (Bolt Beranek and Newman, Cambridge, MA; BBN Report # 5726)

Weiss M.R., Vogel R.P., Harris C.M. (1966): "Implementation of a pitch-extractor of the double spectrum analysis type." J. Acoust. Soc. Am. **40**, 657-662

PRELANGAGE

D. Bouhénic

31, Place de Jaude
63 000 CLERMONT-FERRAND

APPRENTISSAGE DU LANGAGE

GENESE	- Cri à la naissance
de 0 à 3 mois	- Cri réflexe
communication	- Cri communication mère/enfant
	Gestes
Expression	- Gazouilli - grognement jasi
Vers 6 mois expression geste	- Gazouilli expressif - Onomatopées
Vers 12 mois Prélangage	- les premiers mots
Vers 15/18 mois	- mots enfantins
Langage	- Les mots assemblés - les phrases simples significatives

ACQUISITION DU LANGAGE DE L'ENFANT NORMAL

"L'acquisition du langage est une fonction humaine qui s'exerce comme une technique du corps" (1)

Il faut considérer que l'acquisition du langage, qui se fait dans les trois premières anrées, fait partie de la croissance et de la maturation de l'enfant.

Les émissions audibles commencent à la naissance, par les cris ; ils sont un moyen de communication, c'est par leur nuance que l'enfant interpelle soit sa mère _ (établissement du code Mère/Enfant) soit les autres personnes qui s'occupent de lui. Les cris sont produits par une décharge de forte tension survenant aussi dans des situations de besoin (faim, sommeil, etc.).

L'enfant acquiert, très vite, des possibilités de discrimination de la communication acoustique, comme il acquiert celles des gestes qui précèdent et aide celle de l'acoustique : l'oreille suit la vue.

Les cris sont alors enregistrés par le bébé qui prend conscience de son émission vocale. C'est une première communication sociale, il y a aussi le sourire, le sourire qui apparait chez l'enfant vers la fin du premier mois, souvent provoqué par les parents, c'est une manifestation spontanée, c'est un indice d'euphorie ; vers l'âge de deux mois les exercices buccaux et laryngés provoquent des sons différenciés, il utilise la mélodie. Le jasi intervient vers trois mois, les cris, les gazouillis, les jasements, sont pendant les premiers mois des comportements fonctionnels très différents.

L'enfant se construit sa motricité articulatoire à partir d'un contrôle auditif ; c'est à cette période que se fait le mécanisme.

Au contraire des cris, les jasements ne sont produits que lorsque la tension est modérée, c'est donc une activité gratuite, apparaissant lorsque les besoins sont satisfaits.

(1) M. COHEN

Nous distinguons :

a) période de roucoulement

qui commence à environ deux mois

b) période de gazouillis

qui commence à 3 mois et qui dure dans l'ensemble jusqu'à 9 mois.

Remarque : l'enfant sourd gazouille, il n'existe pas avant 5 à 6 mois de différences observables entre les productions sonores des enfants sourds et celles des entendants.

La période de gazouillis est une période de transition vers le langage, c'est une forme de prélangage au même titre que les onomatopées qui jouent un rôle important dans le développement linguistique de l'enfant, et dans les relations Mère/Enfant. Il s'agit d'un mode d'expression où l'investissement affectif et ludique est particulièrement important.

L'acquisition du langage se fait par étapes successives, l'enfant est tributaire de l'entourage, il enregistre les sons externes, les sons émis par l'adulte et ses propres sons, il les reconnaît, il les apprécie ; c'est à partir de cette période 3 ou 4 mois, qu'il se met à babiller, ce langage est principalement vocal, mais il faut faire la part des gestes.

Pour la vocale, comme pour le mime, une grande part appartient à l'invention de l'enfant conditionné par la communication Mère/Enfant.

Le principal est l'acquisition d'un ensemble de sons non significatifs par eux-mêmes, mais servant à constituer des mots qui ont une signification.

C'est donc vers le 4ème mois que l'enfant commence à utiliser sa voix (Jasi).

Les différents organes qui interviennent dans le langage sont :

- La respiration,
- le souffle : nous ne pouvons pas parler si nous n'émettons pas un souffle que nous modulons.
- la bouche, la langue,
- le voile du palais (avec ou sans prise de conscience)
- le larynx (qui comporte les cordes vocales)

Au sixième mois c'est vraiment le point important dans la communication, c'est le rythme des mélodies, c'est là qu'on dit qu'un enfant peut apprendre n'importe quelle langue il émet n'importe quelle position articulatoire.

Les premières émissions significatives apparaissent vers le neuvième mois, sous une forme simple ; c'est aussi une émission de communication. Au cours de la première année nous distinguons deux sortes de comportement :

- le cri
- le jasi

ils se rapprochent, tout en s'insérant dans un contexte de plus en plus relationnel :

- d'une part l'enfant contrôle progressivement sa tension et ainsi différencie son activité et l'adapte aux situations. Les cris sont remplacés par différentes sortes de comportement auto-stimulation, activité sur les choses, agressions appels.

- d'autre-part, les verbalisations deviennent chargées de significations affectives (désir, refus).

Les permiers mots apparaissent lorsque ces deux "lignées" fonctionnelles se rejoignent (non, maman, boire, donne etc.).

Les caractéristiques tensionnelles de l'enfant et son adaptation à l'entourage, l'amènent à développer, de façon préférentielle tel ou tel mode de comportement. Le langage se développe comme moyen privilégié de relation à autrui, sous tendu par des désirs et des oppositions.

Vers quinze mois les enfants ont beaucoup d'émissions à leur disposition mais qui ne sont pas vraiment des mots ; c'est l'installation de l'apprentissage. (Cette observation a été étudiée et publiée dans le Bulletin de la Société Linguistique de 1968). Nous assistons à la période des mots réduits, ces mots, au point de vue phonétique, sont aussi polyvalents si on les considère du point de vue de l'expression, ils sont reconnaissables par l'entourage comme des mots de la langue adulte, avec le temps ces émissions se multiplient, se compliquent et apparaissent les premiers mots enfantins (dada, pipi) c'est dans cette période qu'on peut parler de premier mot pris au langage adulte avec leur sens "donné", souvent accompagné du geste, "non" qui exprime la négation.

Il semble que certains phonèmes sont plus faciles à prononcer que d'autres, les labiales et les gingivales paraissent plus faciles à dire.

A 18 mois, étape capitale, c'est l'étape des assemblages, l'enfant ne dit pas des phrases complètes, mais il assemble des mots, l'enfant complique son langage en accumulant des éléments significatifs, avant l'intervention sur le palais qui se fait à cet âge, l'enfant ne peut, du fait des

conditions ancloniques, produire la plupart des phonèmes de la langue maternelle.

De plus ces lallations déformées ne sont pas stimulantes par la mère qui rejette ce que dit son enfant.

Les progrès sont très rapides, mais l'apprentissage reste à faire.

C O N C L U S I O N

Il est difficile de définir le thème prélangage, il est avant tout une question. Une étude aussi complexe suppose le concours de nombreuses disciplines.

Il s'agit, d'une part, de langage supposant des modèles linguistiques et psycholinguistiques et d'autre-part, de genèse supposant une étude de la mise en place et de la synthèse de nombreuses fonctions.

Chaque âge correspond à un système qui reçoit son rôle de l'ensemble de l'organisation en évolution.

Dès sa naissance l'enfant donne de la voix, une voix qui ne s'arrêtera plus durant toute une existence.

Les signes cliniques permettant de suspecter une surdité de 0 à 18 mois sont :

a) dans les premiers mois

- absence de réaction aux bruits
- sommeil trop calme
- réactions très positives aux vibrations et au toucher

b) de 3 à 12 mois

- sons émis non mélodiques
- pas d'articulation
- installation d'une communication gestuelle de désignation

c) de 12 à 18 mois

- absence de paroles articulées
- enfant inattentif à ce qui n'est pas dans son champ visuel
- émissions vocales incontrôlées.

Il est souhaitable de considérer les enfants dits à hauts risques, tel que le décrit la recommandation 02.1 dans la classification étiologique et clinique des déficients auditifs.*

* Réf. J.C. Lafon

J'ajouterai à cette fiche descriptive les troubles du comportement psycho affectifs et les troubles caractériels.

Ces troubles doivent être testés comme un signe de surdité.

****
* LEXIQUE, SYNTAXE, SEMANTIQUE *
****

ACCES RAPIDE DANS UN DICTIONNAIRE DE MOTS *

Philip Lockwood

Laboratoires de Marcoussis
C.R.C.G.E.
Route de Nozay
91460 Marcoussis
France

ABSTRACT

A speaker independent recognition system based on a global approach uses several templates per word of the vocabulary. In a recognition process, a full search procedure is used. This process is time consuming and limitative of the size of the vocabularies these systems can handle for real time applications.

This work deals with the use of a fast algorithm for the nearest neighbor computations. A tree structure is first created on the training set. In a recognition phase, tree search procedures are used (with and without backtrack).

1. Introduction

De nombreux systèmes de sont basés sur l'utilisation de la programmation dynamique (DTW). Le succès de cette aproche vient des bons résultats obtenus pour des vocabulaires de taille modérée. L'inconvenient majeur qui limite la taille des vocabulaires dans les applications temps réel est le coût de calcul. Ceci est encore plus sensible pour les systèmes multilocuteurs où l'on garde généralement plusieurs références par mot.

Des éfforts de recherche sont actuellement menés pour résoudre le problème du temps réel: à travers le développement d'architectures adaptées [6], [21] et à travers le developpement de nouvelles approches de précision équivalente mais de moindre complexité [16].

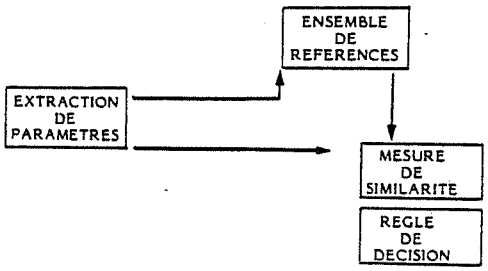
Le travail que nous decrivons va dans le sens de la reduction de la complexite d'un système fondé sur l'utilisation de la programmation dynamique.

Deux facteurs intervenant fortement dans la complexité de ces systèmes sont la place mémoire et le coût d'une comparaison. Ces deux aspects sont très liés mais chacun peut être optimisé. Le but de ce travail est de réduire le nombre de comparaisons à éffectuer en utilisant une methode de recherche rapide du plus proche voisin.

2. Le système de reconnaissance de mots isolés multilocuteur

SYRIL est un système de reconnaissance de mots isolés multilocuteurs, developpé dans notre laboratoire [1]. Ce système, initialement prévu pour un vocabulaire de 35 mots, a récemment été étendu à 130 mots [5]. Nous allons brièvement en rappeler les principes en décrivant succintement ses différentes parties constituantes

Figure : description du système de reconnaissance de mots



- Le prétraitement:

  La fréquence d'échantillonnage est de 8 KHz. Un ensemble de 9 paramètres cepstraux est calculé toutes les 16 millisecondes en prenant une fenêtre de 32 millisecondes. Ces paramètres sont déterminés après un passage dans un banc de filtres dont les fréquences centrales sont réparties selon une échelle Mel [4]. Un dixième paramètre proportionnel à l'énergie est aussi determiné [1].

- La création de l'ensemble de références:

  Un corpus comprenant une élocution des 130 mots par 100 locuteurs (50 Femmes, 50 Hommes) a été créé.

  Les 40 premiers locuteurs ont été utilisés pour créer un dictionnaire de références. Pour chacun des mots du vocabulaire, nous avons selectionné un sous ensemble de représentants par classification automatique (la méthode que nous utilisons est UWA, decrite dans [15]). Nous avons extrait 5 representants sur les 40 initiaux. Le dictionnaire final comprend 130*5 mots au total.

- La mesure de similarité:

  Un algorithme de programmation dynamique est utilisé pour le calcul d'une similarité entre deux mots (Sakoe type 1 symétrique [18]). L'algorithme tient compte des distorsions temporelles non linéaires dans la prononciation d'un mot.

  Cette mesure de similarité n'est pas une distance; cette remarque est importante car les méthodes de recherche rapide que nous envisageons travaillent dans un espace métrique.

  La mesure est positive, symétrique mais l'inégalite triangulaire n'est pas toujours verifiée. Nous avons fait la verification pour les mots du dictionnaire et pour C^3_{650} inégalités testées, 1647960 ne sont pas verifiées, soit 3.6 %.

- La règle de décision:

  La règle de décision du plus proche voisin "full search" est utilisée; le mot inconnu est comparé à l'ensemble des mots du dictionnaire, soit dans notre cas 130*5 mots.

3. Complexité du système

La détermination du score est l'opération la plus coûteuse en temps de calcul; pour le vocabulaire de 130 mots [5], un mot fait en moyenne 49 prélèvements (49 vecteurs de 10 paramètres). Le calcul de la similarité nécéssite en moyenne, avec une fenêtre de 12, 12*49 calculs de distances élémentaires pour un mot (une distance élémentaire est calculée entre deux vecteurs de paramètres, la distance euclidienne usuelle est utilisée). Le mot inconnu doit être comparé à chacun des mots de référence (650 comparaisons, 380 000 distances élémentaires).

Afin de réduire le nombre de distances élémentaires à calculer, nous avons utilisé un algorithme de segmentation de la trace [8]. Cette méthode a l'avantage de ne pas déteriorer les performances du système tout en réduisant considérablement sa complexité. L'idée est de compresser les parties "stables", qui contiennent de l'information redondante, et de conserver au maximum l'information dans les parties transitoires. Pour plus de détails sur cette méthode nous renvoyons aux références [10], [8]. Dans le calcul de la trace, nous n'avons utilisé que les 9 coefficients cepstraux. Lors de la détermination de nouveaux vecteurs par interpolation linéaire, nous avons traité le dernier paramètre séparement car c'est le logarithme de l'energie normalisée par rapport à l'energie moyenne du mot [1].

Par cette technique, tous les mots du dictionnaire ont été ramenés à une même longueur de 20 prélèvements. La complexité dans le calcul des distances élémentaires est maintenant fortement réduite, avec une fenêtre de 6, il y a 6*20 distances élémentaires pour un mot à calculer (80 000 pour l'ensemble des mots du dictionnaire).

Des techniques de quantification vectorielle [2], [11], permettraient une réduction encore plus importante en taille mémoire et donc en complexité mais ce point n'a pas été abordé dans le cadre de ce travail.

4. Quelques algorithmes de recherche rapide de plus proche voisin

Ce sujet a fait l'objet de nombreuses recherches et beaucoup d'auteurs ont proposé des méthodes pour la recherche rapide du plus proche voisin dans un espace métrique. Il se dégage deux familles de méthodes qui semblent mieux adaptées à notre problème:

- Les méthodes de condensation: l'idée est de réduire la taille du dictionnaire en ne conservant que les mots nécéssaires à la discrimination; une synthèse et des nouveaux résultats peuvent être trouvés dans [20].

- Les méthodes hiérarchiques: une structure hiérarchique est créée sur le dictionnaire initial et lors de la phase de reconnaissance, l'arborescence est utilisée pour faire un accès rapide.

L'avantage de ces méthodes hiérarchiques est dans l'accès logarithmique. Pour un arbre l-aire équilibré, le nombre d'accès sera de l'ordre de $O(Log_l\ N)$, N étant le nombre d'objets dans le dictionnaire, le nombre de comparaisons sera de l'ordre de $l(Log_l\ N)$. L'inconvénient est qu'il faudra stocker des informations supplementaires, et éventuellement créer de nouveaux représentants.

5. Structure de données et recherche rapide

5.1. Construction d'une structure de données

Deux methodes sont généralement utilisées lorsque l'on ne dispose pas de clés numériques ou autre pour construire l'arbre. Nous allons brièvement décrire deux procédures pour la création d'un arbre l-aire.

Etant donné un ensemble d'objets à classer à un noeud, on peut utiliser un algorithme de classification automatique pour repartir les objets en "l" groupes; les représentants des groupes serviront à créer "l" nouveaux noeuds. La même procédure sera alors réappliquée récursivement sur les objets d'un groupe jusqu'à ce que chaque objet initial soit affecté à un noeud terminal [7], [19].

Une autre procédure est couramment utilisée en codage vectoriel de la parole [2]. Une optimisation globale par niveaux est faite, en utilisant un algorithme hiérarchique du type Nuées Dynamiques. Le partitionnement initial se fait d'abord en "l" classes, des prototypes pour chaque classe sont determinés et stockés aux noeuds, chaque prototype est ensuite perturbé et génère "l" candidats. Ces l^l candidats initialisent l'algorithme des Nuées Dynamiques, qui est réappliqué globalement sur l'ensemble des objets. La même procédure est appliquée récursivement tant que tous les objets initiaux n'appartiennent pas à un noeud terminal.

Le choix entre l'une ou l'autre des méthodes dépend essentiellement de l'application. Dabouz [3] a montré que l'arbre obtenu par la seconde méthode était mieux equilibré et moins profond pour une application en transmission de la parole.

La création d'une structure de données sur des mots ne peut être envisagée de manière habituelle. Différents problèmes se posent principalement dû aux méthodes et moyens dont nous disposons pour comparer ces unités; par exemple, il semble difficile de faire la moyenne de deux mots différents car le "mot" résultant ne serait pas un mot moyen au sens géometrique du terme mais simplement un autre objet (nous préférons le terme objet car dans ce cas l'objet n'est plus un mot au sens phonétique). Nous avons de même fait quelques expériences en moyennant des representants d'un même mot; il n'est pas automatique que le mot résultant soit le minimax de l'ensemble. Ceci veut dire qu'il ne nous sera pas possible de faire systématiquement un moyennage pour avoir un prototype aux noeuds.

Nous avons utilisé une méthode du premier type car notre objectif est de montrer la faisabilité d'une telle approche. Cette méthode nous a semblé plus souple d'utilisation étant donne les contraintes inhérentes à notre problème. Un tel algorithme sera plus facilement adaptable localement, car à chaque noeud, nous pouvons redéfinir le nombre de classes souhaitées et ainsi pouvoir mieux exploiter la structure géometrique des objets à classer à un instant donné.

Nous allons maintenant décrire la méthode et les problèmes rencontrés.

- Notations:

S : Ensemble d'objets (Mots du dictionnaire)

N : nombre d'objets dans S

N_i: Nombre d'objets dans le groupe G_i

S_i: Ensemble d'objets associés au noeud i

M_i: minmax du groupe G_i

$$M_i = \min \{\max d(x_j, x_k)\}$$
pour tout x_j and x_k appartenant a G_i

v_i: "pseudo" distorsion totale du groupe G_i

$$v_i = \sum d(M_i, x)$$
pour tout x appartenant à G_i

Nous utilisons le terme "pseudo" car l'espace n'est pas métrique et le centroide ne peut être défini. "Pseudo" sera sous entendu dans la suite.

D(x,y) : Score entre les mots x et y (DTW).

R_i : Rayon de la classe i

$$R_i = \text{Max}\{ D(M_i, y)\} \text{ , pour y appartenant à } G_i.$$

Nos contraintes dans la construction de cet arbre étant principalement des problèmes de temps de calcul, nous avons cherché à automatiser la procédure. Pour cela il a fallu s'affranchir des problèmes d'initialisation des méthodes de type Nuées Dynamiques. Notre choix s'est donc porté sur la construction d'un arbre binaire en utilisant un algorithme optimal.

- Algorithme de classification optimal:

Cet algorithme a été proposé par Shapiro et Halarick [17]. La méthode crée une partition de N objets en deux classes en minimisant explicitement la distorsion globale totale:

$$\min\{ \sum_{i=1}^{i=2} v_i \}$$
pour tout objet à classer.

Nous utilisons le minmax comme représentant de classe dans le but de minimiser l'accroissement en taille mémoire.

- Algorithme de creation d'un arbre binaire

Initialisation: $R_0 = \infty$, $S_0 = S$

-Pas 1: Appliquer l'algorithme optimal sur les objets S_i et creer 2 groupes G_i; les objets appartenant à S_i sont tels que: $D(x, M_i) \leq R_i$.

-pas 2: Creer 2 noeuds N_i et leur associer le minmax M_i et le rayon R_i. Si M_i est le seul représentant de sa classe, il y a création d'un noeud terminal.

-Pas 3: Appliquer les pas 1 et 2 récursivement tant qu'il reste un noeud à traiter.

- Problèmes rencontrés durant la création de l'arbre:

Le problème principal est dû au recouvrement implicite introduit par l'utilisation du minmax. Dans certains cas, un recouvrement trop important a entrainé le cyclage de la methode (le volume engendré par une classe inclut entièrement l'autre classe).

- Solutions apportées:

La première idée a été de rechercher un représentant "meilleur" en cas de cyclage ou lorsque le recouvrement est trop important (nous avons fixé le seuil à 20 %, si l'accroissement dépasse ce seuil, un meilleur minmax est cherché). Le critère optimisé est la minimisation du rayon de la classe. Les representants testés sont obtenus par moyennages successifs entre des mêmes representants d'un mot pris parmi ceux de la classe (on recherche en quelque sorte l'objet le plus proche du "centroide"). L'algorithme de moyennage est decrit dans [1].

Cette phase d'optimisation a été introduite dans la procédure de création de l'arbre, mais dans certains cas, elle s'est averée insuffisante. Les noeuds concernés ont été, dans un premier temps abandonnés.

Une procedure interactive a été developpée pour permettre de traiter l'ensemble des noeuds abandonnés lors de la procedure automatique. La méthode est du type ISODATA; l'operateur a la possibilité de choisir le nombre de classes voulues, d'initialiser la classification de differentes fac,ons, de fusionner des classes, de scinder une classe en deux (soit en utilisant l'algorithme optimal précédent, soit en affectant chaque objet de la classe au plus proche objet définissant le diamètre), de faire des itérations de type Nuées Dynamiques et d'interagir avec la procédure de recherche d'un meilleur représentant décrite précédemment.

Diverses informations guident les choix de l'operateur: matrice des distances entre représentants des classes, rayon des classes, matrice des accroissements interclasses, etc.... Si le resultat n'est pas satisfaisant à un instant donné, il y a possibilité de revenir dans l'état antérieur, de même, il est possible à tout moment de réinitialiser la classification pour le noeud considéré.

En fin de traitement intéractif, la procédure automatique est réitérée.

5.2. Méthodes de recherche

Deux méthodes de recherche ont été implantées; une methode de recherche directe dont les performances devraient être proches de $O(Log_i\ N)$ accès, et une méthode avec retour arrière. Dans les deux cas, nous ferons l'hypothèse que DTW est une métrique. Nous appliquerons ensuite ce second algorithme avec la procédure tenant compte du fait que DTW n'est pas une métrique.

- Recherche directe [14]:

Cette procédure permet une recherche approchée du plus proche voisin. L'algorithme est le suivant: calculer le score entre le mot inconnu x et chacun des fils du noeud i; descendre la branche i si $D(x, M_i)$ est minimum. Répéter cette procédure tant qu'un noeud terminal n'est pas trouvé.

- Recherche avec retour arrière [7]:

  Cette méthode permet de retrouver exactement le plus proche voisin dans un espace métrique. L'algorithme est fondé sur l'utilisation de deux règles et chaque noeud i peut être testé comme suit:

  Regle 1: Aucun x_j (x_j appartenant a S_i) ne peut être plus proche voisin de x si: $B + d(x_j, R_i) < d(x, M_i)$. B est la distance au plus proche voisin courant de x, parmi les distances deja calculées.

  Regle 2: x_j (x_j appartenant a S_i) ne peut être plus proche voisin de x si: $B + d(x_j, M_i) < d(x, M_i)$.

  La procédure de recherche est la même que précédemment, mais tous les noeuds activés par la règle 1 devront être explorés. La regle 2 sera réappliquée à chaque fois lors des retours en arrière.

6. Conclusion

Nous avons proposé une procédure d'accès rapide dans un dictionnaire de mots pour réduire la complexité de calcul dans la recherche du plus proche voisin. La methode est actuellement en phase de tests et des premiers résultats seront donnés.

Cette méthode pourra être étendue à d'autres unités telles que des syllabes, demi syllabes, diphones, car ces unités peuvent aussi nécéssiter l'utilisation de la programmation dynamique.

Une telle approche pourrait être envisagée aussi dans le traitement des grands vocabulaires; soit globalement sur des mots, soit comme procédure d'accès léxical dans un dictionnaire de chaînes de caractères, chaque caractère étant un label d'une classe phonétique grossière [13], [12].

Les premiers résultats devraient nous permettre d'avoir une idée sur l'accroissement réel en taille mémoire, l'influence sur les performances (l'hypothese faite selon laquelle DTW est une métrique est elle suffisante, sinon comment en tenir compte?), et le taux de réduction en complexité.

[1] N. BRIANT, B. FLOCON, "SYRIL: Système temps réel de reconnaissance de mots indépendant du locuteur," 4ème congrès AFCET reconnaissance des formes, Paris 84.

[2] A. BUZO, A.H. GRAY jr, R.M. GRAY, J.D. MARKEL, "Speech coding based upon vector quantization", IEEE Trans. on Acoustics, Speech and Signal Proc., Vol ASSP-28, n^0 5, Oct 1980.

[3] M. DABOUZ, "Transmission de la Parole à Faible Débit par Vocodeur à Classification", Thèse Dr. Ing., ENST, Jan. 1983.

[4] S.B. DAVIS, P. MERMELSTEIN, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans ASSP, 28, N^0 4, 357-366, aug. 1980.

[5] B. FLOCON, P.LOCKWOOD, J. SAP, "Système de reconnaissance de mots isolés multilocuteurs pour un vocabulaire de 130 mots; intégration dans un poste de travail", 13èmes journées d'étude sur la parole, GALF, Bruxelles, 28-30 mai 1984.

[6] P. FRISON, P. QUINTON, "A VLSI Parallel Machime for Speech Recognition," ICASSP-84, (March 84).

[7] K. FUKUNAGA, P. NARENDRA, "A Branch and Bound Algorithm for Computing K-Nearest Neighbors", IEEE Trans on Computers, July 1975.

[8] J.L. GAUVAIN, J. MARIANI, J.S. LIENARD, "On the use of Time Compression for Word-Based Recognition," Proc. ICASSP-83, (April 83).

[9] V.N. GUPTA, M. LENNING, P. MERMELSTEIN, "Decision Rules for Speaker Independent Isolated Word Recognition", IEEE-ICASSP conf., 1984.

[10] M.H. KUHN, H. TOMASCHEWSKI, H. NEY, "Fast Nonlinear Time Alignment for Isolated Word Recognition," Proc. ICASSP-81, (April 81).

[11] P. LOCKWOOD, "Comparaison de Plusieurs Algorithmes de Classification Automatique en Vue de la Transmission de la Parole à Faible Debit," 11e Congrès International d'Acoustique, Paris, Revue d'Acoustique, Hors série, Vol 4, pp 137-140, 1983.

[12] P. LOCKWOOD, "A Speaker Independent Word Hypothesizer", Digital Signal Processing - 84, V. Cappellini and A.G. Constantinides (eds), Final Edition, North Holland 1984.

[13] J.F. MARI, J.P. HATON, "Some Experiments in Automatic Recognition of a Thousand Word Vocabulary", IEEE ICASSP, 1984.

[14] L. MICLET, M. DABOUZ, "Approximative Fast Nearest Neighbor Recognition", Pattern Recognition letters 1, july 1983.

[15] RABINER L.R., J.G. WILPON, "Considerations in Applying Clustering Techniques to Speaker-Independent Word Recognition", JASA, vol 66, 663-673 1979.

[16] L.R. RABINER, S.E. LEVINSON, M.M. SONDHI, "On the use of Hidden Markov Models for Speaker-Independent Recognition of Isolated Words from a Medium-Size Vocabulary," BSTJ, Vol 63, NO 4, April 1984.

[17] L.G. SHAPIRO, R.M. HARALICK, "Organization of relational models for scene analysis", IEEE Trans. on Pattern Analysis and Machine Intellegence, Vol PAMI-4, n^0 6, Nov 1982.

[18] H. SAKOE, S. CHIBA, "Dynamic programming algorithm optimisation for spoken word recognition", IEEE Trans. on Acoust., Speech and Signal Processing, Vol ASSP-26, 1978.

[19] G. SALTON, A. WONG, "Generation and Search of Clustered Files", ACM Trans. on Database Systems, Vol 3, No. 4, Dec. 1978.

[20] G.T. TOUSSAINT, B.K. BHATTACHARYA, R.S. POULSEN, "The Application of Voronoi Diagrams to Nonparametric Decision Rules", Proc. of Computer Science and Statistics: 16th symposium on the interface, Atlanta, Georgia, March 1984.

[21] N. WESTE, D.J. BURR, B.D. ACKLAND, "Dynamic Time Warp Pattern Matching Using an Integrated Multiprocessing Array," IEEE Trans. on Computers, Vol c-32, NO 8, Aug. 1983.

* Etude menée dans le cadre d'un contrat de la Communauté Européenne (programme ESPRIT).

DICTÉE VOCALE PAR MOTS ISOLÉS

D. Bellilty - A. Lund - J. Mariani - G. Adda - F. Néel - C. Fluhr

LIMSI-CNRS B.P. 30 91406 Orsay Cedex

ABSTRACT

Once a phoneme lattice has been obtained from an acoustico-phonetic decoder (ESOPE or SHERPA), a lookup in a large dictionary of 270,000 forms allows us to find the correct word.

The algorithm we have developped, which takes elision, addition, and confusion into account, shows that performance is rapidly degraded where the correct phoneme is not in the lattice of the first three or four candidates.

The dictionary access and organisation seem to be of great importance.

I- INTRODUCTION

Dans le cadre d'un projet à long terme sur la dictée vocale, nous étudions la possibilité de réaliser un système de reconnaissance de phrases dont chaque mot serait prononcé isolément, le vocabulaire étant constitué par un dictionnaire de grande taille.

Ce système se caractérise par la liaison entre un module de reconnaissance vocale fournissant un treillis de phonèmes, et un système d'analyse lexicale et syntaxique de chaînes phonétiques.

Cette liaison, dont il sera fait état plus particulièrement dans cet article, doit donc, à partir des candidats-phonèmes inclus dans un treillis fourni par le système de reconnaissance, trouver les mots phonétiques les plus probables, c'est à dire ceux qui ont le plus de chance d'avoir été effectivement prononcés .

II- RÉALISATION ET DESCRIPTION

1) Le treillis phonétique

Les treillis phonétiques sont fournis par un système de décodage acoustico-phonétique (ESOPE ou SHERPA).[1][2][3]

Dans tous les cas de figure, le treillis contient pour chaque phonème détecté, la liste des phonèmes qui peuvent lui correspondre, du plus probable au moins probable, ainsi qu'une note de reconnaissance destinée à chiffrer cette vraisemblance.

A partir de ce treillis, il est alors nécessaire de créer des suites de phonèmes, puis de tester leur existence en tant que mot à l'aide d'un dictionnaire.

Le nombre de suites phonétiques qu'il est possible de générer à partir d'un treillis étant souvent gigantesque, il a été nécessaire de ne conserver à chaque étape, que les suites les plus probables, cette probabilité étant obtenue à partir des notes de reconnaissance de chaque candidat pour chaque phonème.

En effet si l'on imagine par exemple un treillis constitué par six phonèmes et cinq candidats par phonème, on obtiendra 5**6 chaînes phonétiques possibles, c'est à dire 15 625 suites.

Comme, de plus, ces chaînes sont supposées erronées (les systèmes de reconnaissance n'ayant pas encore atteint un taux de reconnaissance de 100%), leur constitution n'est pas suffisante; il est nécessaire ensuite, pour retrouver le mot phonétique le plus plausible, d'explorer le dictionnaire.

2) La distance

Après avoir déterminé les chaînes phonétiques les plus probables, il faut ensuite les comparer avec les mots existant dans la langue.

Pour comparer la suite phonétique à un mot phonétique du dictionnaire, nous utilisons une distance qui sert à quantifier leur proximité.

De façon idéale, la distance entre une chaîne phonétique CHN et un mot extrait du dictionnaire phonétique MOT, représente la probabilité pour que MOT ait été prononcé, sachant que CHN est constitué de la suite des phonèmes effectivement reconnus.

Cette distance prend en compte trois types d' erreurs que le système de reconnaissance est susceptible de produire, les erreurs de remplacement, de rajout ou d'élision de phonèmes.

De manière pratique, nous avons utilisé pour

réaliser cette distance, des matrices (une matrice par type d'erreur) qui, à chaque couple de phonèmes attribue une note de pénalité, celle-ci étant d'autant plus grande que la probabilité d'obtenir ce type d'erreur pour le couple de phonèmes est faible.

Enfin, pour le mot entier, on somme les notes de pénalité de tous les phonèmes, relatives aux erreurs considérées.

Cependant, en raison des erreurs d'élisions et de rajout, on peut souvent faire correspondre le mot supposé prononcé avec le mot reconnu de multiples façons.

Par exemple:
soit le mot prononcé :ãgrãʒərɛ (engrangerais)

et le mot reconnu :ãrãʒarrɛ
Nous avons deux solutions:

1ère solution: ãgrãʒərɛ / ãrãʒarrɛ (5 remplacements)

2ème solution: ãgrãʒərɛ / ãrãʒarrɛ (1 élision,1 rajout)

Il est visible que la deuxième solution est meilleure que la première, car il est peu vraisemblable que la correspondance remarquable des phonèmes obtenus soit le fruit du hasard.
Nous devons donc calculer la distance entre deux mots d'après le deuxième cas de figure , de préférence au premier.

Aussi, la distance à considérer est le minimum des notes de pénalité que l'on obtient en considérant tous les cas de figure.

L'exploration systématique de la totalité des possibilités étant prohibitive en temps de calcul, nous avons utilisé une méthode de " branch and bound", c'est à dire une construction par étapes de l'arbre des possibilités, où, à chaque étape, seule la branche qui se termine par le noeud de plus faible pénalité est prolongée.
Cette méthode nous permet d'obtenir la meilleure solution avec le minimum d'opérations.

3) Le dictionnaire

Le dictionnaire a été constitué à l'aide d'un lexique orthographique contenant 270 000 formes, ce lexique dérivant d'un dictionnaire de base de 35 000 mots.[4]
Le dictionnaire contient la liste des mots phonétiques présents dans la langue, obtenu par phonétisation du dictionnaire orthographique.

A chaque mot phonétique est attaché une information, contenant le ou les mots orthographiques correspondant à la clé phonétique, et pour chacun d'eux des informations linguistiques, comme leur catégorie grammaticale, le nombre et le genre pour les substantifs ou les adjectifs, le temps ou le mode pour les verbes, etc.
Le dictionnaire, qui contient 120 000 clés phonétiques, est construit suivant une structure arborescente, les mots de la langue étant classés par l'ordre alphabétique de leur forme phonétique, chaque phonème étant représenté par un caractère.

4) Le parcours du dictionnaire

Le principe de la recherche des mots dans le dictionnaire est le suivant : on utilise la chaîne phonétique reconnue, comme clé pour se positionner dans le dictionnaire.
Puis on extrait les mots existant autour de la position pointée, parmi lesquels on espère trouver le mot énoncé.

Ceci suppose donc que la chaîne phonétique qui sert à se positionner n'est pas trop éloignée, du point de vue de l'ordre alphabétique du dictionnaire, du mot prononcé, et donc que les premiers phonèmes ont été correctement reconnus.
Pour trouver le mot prononcé, dans le cas où les premiers phonèmes sont mal reconnus, il a été nécessaire d'introduire un deuxième jeu de clés qui correspondent aux mots phonétiques inversés. Le problème de la mauvaise reconnaissance des premiers phonèmes est alors éliminé si les derniers phonèmes sont bien reconnus.

III- LES ESSAIS

1) Les treillis artificiels

Nous avons effectué une première série de tests , portant sur des chaînes rendues artificiellement erronées par l'ajout de manière aléatoire d'une ou de plusieurs erreurs, dans un mot phonétique correct, cette série de tests visant à valider la méthode d'analyse.
Les résultats obtenus en ont montré les limites.
En effet, le système, à l'aide d'une recherche sommaire autour de la clé phonétique pointée (5 mots de part et d'autre), aboutissait sur le bon mot dans 65% des cas , pour un remplacement simple ou pour une élision simple, et dans 85% des cas pour un ajout simple .
De plus, il faut souligner que les résultats sont meilleurs pour les mots longs que pour les mots courts, et lorsque le phonème erroné est situé au début ou à la fin du mot.
Mais lorsque l'on passe à des erreurs doubles à l'intérieur d'un même mot, les pourcentages diminuent énormément, puiqu'on ne reconnaît le mot phonétique exact que dans 30% des cas en moyenne.

Cependant le système de générations de fautes est loin de représenter la réalité d'un treillis phonétique, et nous ne pouvons tirer de ces essais que des conclusions limitées.

Toutefois, on peut noter que :

-d'une part, la méthode d'accès au dictionnaire semble déterminante, la plupart des erreurs prove-

nant du fait que le mot à retrouver était physiquement loin, dans le dictionnaire, de la chaîne erronée qui constitue la clé d'entrée.

Ceci est dû au fait que, lorsque le nombre de mots recherchés autour de la clé est trop faible, le programme extrait souvent autour du mot pointé les formes dérivées d'un même mot source, et ne parvient pas à la forme désirée; le cas est particulièrement flagrant avec une forme conjuguée d'un verbe, lorsque l'erreur se trouve en fin du mot.

-d'autre part, en conservant la même méthode d'analyse et le même parcours dans le dictionnaire , les résultats ne seront utilisables par les niveaux supérieurs (syntaxique et sémantique), que si le système de reconnaissance phonétique utilisé possède des taux de réussite supérieur à 80 %, une erreur étant ici l'absence du bon phonème parmi les trois ou quatre premiers candidats (en tenant compte des rajouts et élisions).

2) Les treillis réels

Nous avons de plus effectué des essais sur quelques treillis fournis par les systèmes ESOPE et SHERPA, où nous avons constaté, comme cela était prévisible, que les résultats obtenus dépendent beaucoup de la qualité des treillis fournis.

En effet, dans un seul cas la phrase correcte pouvait être reconstituée à l'aide des mots présents dans le treillis de mots fourni par le système.

Dans les autres cas, les mots prononcés n'étaient que rarement retrouvés par le programme, les suites extraites des treillis phonétiques étant souvent plus proche de mots faux mais existant dans la langue, que du mot prononcé.

IV- CONCLUSIONS ET PERSPECTIVES

Ce travail avait pour but de mieux cerner, à l'aide d'une approche naïve (c'est à dire la plus simple possible), les problèmes liés à la reconnaissance de très grands vocabulaires.

Nous nous sommes donc placé dans un cas extrême , en utilisant un dictionnaire de 270 000 formes orthographiques, représenté à l'aide de phonèmes; de plus nous avons choisi un ordre arbitraire (l'ordre alphabétique) pour classer les mots phonétiques.

Les résultats nous ont montré que cette approche n'était pas à rejeter, du moins en utilisant un système de reconnaissance phonétique ayant un taux de réussite supérieur à 80%.

Mais ils nous ont également montré quelles étaient les améliorations possibles (et souvent nécessaires) à apporter au système.

Tout d'abord, la distance utilisée pour calculer l'éloignement entre deux mots phonétiques doit être adaptée au système de reconnaissance utilisé, afin de mieux prendre en compte ses performances.

Ensuite, la taille du dictionnaire est telle que, à partir d'une chaîne erronée, on aboutit toujours à un mot de la langue qui n'en est pas trop éloigné, ce mot n'étant pas obligatoirement le mot prononcé. Il serait donc peut-être nécessaire si l'on désire garder le même système d'analyse, de réduire la taille du dictionnaire.

Enfin, la recherche dans le dictionnaire est aveugle, et ne tient pas comptes des confusions phonétiques courantes. En effet, il semble naturel , si l'on veut pouvoir retrouver facilement les mots dont la distance (au sens utilisé au II-2) est faible, de les placer dans la même zone du dictionnaire.

Cette dernière remarque nous renvoit aux travaux effectués sur la reconnaissance de grands vocabulaires, qui utilise le codage des mots à l'aide de classes phonétiques grossières, qui peuvent prendre en compte les erreurs dû à la prononciation et à la reconnaissance, ce qui permet de définir des sous-vocabulaires, dans lesquels on peut ensuite faire une recherche plus fine.[5][6][7]

BIBLIOGRAPHIE

[1] J. MARIANI : "ESOPE: un système de compréhension de la parole continue", Thèse d'état - Paris VI, juillet 1982.

[2] A. ANDREEWSKY , M. DESI, C. FLUHR , J-S. LIÉNARD , J. MARIANI , F. NÉEL , F. POIRIER : "Reconnaissance pour la dictée automatique de la parole continue", ATALA 1983, Vol.24 n°2.

[3] A. ANDREEWSKY , F. POIRIER , M. DESI , C. FLUHR: "SHERPA: un système de reconnaissance de la parole continue - Résultats et développement", 13ème JEP-GALF, Bruxelles, 28-30 mai 1984.

[4] A. ANDREEWSKY , J-P. BINQUET , F. DEBILI , C. FLUHR , Y. HLAL , J-S. LIÉNARD , J. MARIANI , B. POUDEROUX: "Les dictionnaires en formes complètes et leur utilisation dans la transformation lexicale et syntaxique correcte de chaînes phonétiques", 10ème JEP-GALF, Grenoble, 30 mai-1er juin 1979.

[5] D.W. SHIPMAN , V.W. ZUE: "Properties of large lexicons. Implications for advanced isolated word recognition systems", proc. IEEE ICASSP 1982, Paris.

[6] V.W. ZUE , D. HUTTENLOCHER: " Computer recognition of isolated words from large vocabularies", trends and applications IEEE computer society, may 1983

[7] J-F. MARI , J-P. HATON: "Some experiments in automatic recognition of a thousand word vocabulary", proc. IEEE ICASSP 1984, San Diego.

UNE METHODE D'APPRENTISSAGE AUTOMATIQUE DE GRAMMAIRES POUR LA COMPREHENSION AUTOMATIQUE DE LA PAROLE

Philippe Laroque, Joseph Mariani

LIMSI/CNRS, B.P.30, 91406 ORSAY Cedex

ABSTRACT

Most continuous speech recognition systems cannot deal with an utterance which only slightly deviates from their allowable syntax. In order to try to compensate for this situation, we have designed an algorithm permitting the automatic construction of a syntax from only a few examples. We think that only automatic training techniques can ameliorate the syntactic level of recognition systems in any great measure. Moreover, self-adaptation adds a flexible component to the system. Based on partial results that we have obtained on the first stage of the system, we also describe the way it operates, using it on the easier task of parsing letter strings.

II SITUATION DU PROBLEME

Les systèmes actuels de compréhension de la parole continue peuvent se ranger "grosso modo" dans deux grandes catégories, opposées en ce qui concerne leur méthode d' analyse: d' une part, ceux qui font appel à une analyse montante (ou bottom-up), c' est le guidage par les données; d' autre part, ceux qui adoptent une stratégie descendante (top-down), guidée essentiellement par la syntaxe , i.e. par la grammaire décrivant le langage à accepter.

L' avantage de la première catégorie de systèmes réside bien entendu dans une souplesse d'utilisation bien supérieure: le nombre de phrases potentiellement recevables n' est pas limité, ce qui autorise des structures syntaxiques variées. Néanmoins, cet avantage est remis en cause par le fait que, le système manquant d' informations pour guider son analyse, il est obligé d' envisager un nombre considérable d' hypothèses, d' où une perte de temps énorme, et surtout un risque constant d' explosion combinatoire. Ce risque est absent des systèmes appartenant à la seconde catégorie: ceux-ci, en effet, se contentent d' examiner les hypothèses recevables dans le cadre de leur grammaire, donc à chaque fois un nombre (plus) limité, ce qui fait d' eux des systèmes plus performants en temps que les précédents. Maleureusement, ces systèmes échouent dès le plus petit écart, qu' il soit d' ordre lexical (mot inconnu du système), ou syntaxique (tournure non prévue par la grammaire).

L'idéal, un système qui combinerait souplesse d' utilisation et rapidité en temps de calcul, étant inenvisageable en raison de l' opposition très nette entre les deux conceptions, on ne peut que tenter d' atténuer les défauts des systèmes existants. Bien qu' on puisse se permettre d' adopter une stratégie montante dans le cas de petits vocabulaires, il semble que la présence d'une grammaire soit indispensable dès qu' on veut traiter un langage "réel", i.e. qui ne soit pas uniquement un langage de commande. Le problème est donc de tenter d'assouplir cette dernière, de façon à lui faire accepter certains écarts par rapport à ce qui est initialement prévu.

Nous avons d' abord essayé une méthode simple et facile à implanter: l' adoption d' une syntaxe ne contenant qu'un nombre restreint de catégories hiérarchisées qui permettent de focaliser rapidement l'attention du système sur le(s) mot(s) important(s) de la phrase :quand un mot d'une certaine catégorie a été détecté, le système ne recherche plus que les mots des catégories supérieures (pour plus de détails, consulter [7]). Ce système fonctionnait en deux passes, la seconde s' appuyant sur des "points d'ancrages" (mot reconnus avec une bonne certitude) détectés lors de la première. Cependant, il est rapidement devenu clair que cette solution n'est valable que dans le cas d' un langage contraint (ici, les

phrases portaient sur un vocabulaire du type "standard téléphonique"), avec seulement un ou deux mots "pleins" (i.e. porteurs de sens) par phrase. En fait, il semble que le problème soit impossible à résoudre si la grammaire, si tolérante soit- elle, reste figée, en d' autres termes si le système est incapable de se modifier lui- même: il faut donc faire appel à des techniques d' apprentissage ([3],[8]), qui seules peuvent permettre au système de faire face à un cas non prévu initialement.

III LE FONCTIONNEMENT DU SYSTEME

Le programme a été écrit en LISP sur PDP 11/23. Il prend en entrée un ensemble quelconque de phrases. Comme il s' agit de générer une grammaire régulière, nous avons choisi comme mode de représentation un automate à états finis: le système, dans un état Ti quelconque, passe dans l' état Tj s' il voit le mot MOTij. La première étape consiste donc à créer autant d' états qu' il est nécessaire pour transformer les phrases d' entrée en un paquet de règles de la forme

Ti ==> MOTij Tj

L' état de départ (ou axiome) est noté "S", et un état dans lequel une phrase peut se terminer est noté "Fj" au lieu de "Tj".

Après la création de ce paquet de règles, une seconde étape de travail est chargée de le restreindre en fonction des règles déjà existantes, ou même des redondances potentielles déjà contenues dans les phrases que l' on vient d' introduire. Cette étape est plus particulièrement chargée de

- rendre la grammaire déterministe (dans le but d'éviter le backtrack -temps de calcul exponentiel- lors d'une analyse syntaxique, et de réduire la taille de l'automate),
- confondre certains états offrant une grande ressemblance (pour avoir une idée précise de la notion d' états équivalents, consulter [2]), et
- regrouper certains mots pouvant s' employer aux mêmes endroits, dans des classes de mots (repérées par le système par l' ajout du signe "$" en tête du premier mot de la classe).

Pour rendre la grammaire déterministe, il suffit de s' assurer qu' aucun état Ti n' offre plus d' un état Tj possible à la suite d' un mot donné. Si le cas se produit, on identifie alors les deux chemins et on répercute l' effet de cette identification sur l' ensemble des règles de la grammaire.

Exemple:

Ti ==> MOT Tj Ti ==> MOT Tk Tj ==> ... Tk ==> ...	devient	Ti ==> MOT Tj Tj ==> ... Tj ==> ...

Pour pouvoir confondre deux états, il faut examiner, d'une part l' ensemble des suites possibles de chacun de ces états, et d' autre part ce qui peut les précéder. S' il y a identité, ces deux états sont "unifiés", et on en répercute l' effet sur l' ensemble de la grammaire. Cette procédure sert à obtenir une généralisation de la grammaire, de façon à ce qu' elle puisse accepter d' autres phrases que celles déjà entrées.

Le regroupement de mots en classes s' effectue ainsi:

- si un mot et une classe se trouvent entre deux mêmes états, le mot est intégré à la classe.
- s' il s' agit de deux mots, une nouvelle classe portant le nom du premier est automatiquement créée.
- enfin, s' il s' agit de deux classes, l'une d' entre elles seule subsiste, après avoir intégré les éléments de la seconde.

Cette procédure permet de réduire sensiblement le nombre de règles sans perte d' information. Bien entendu, les modifications engendrées dans la grammaire par ces procédures peuvent rendre possibles d' autres modifications. C' est pourquoi le système continue d' examiner les règles jusqu' à ce que plus aucune réduction ne soit possible.

La troisième et dernière phase,assez réduite en temps de calcul, permet d'obtenir assez rapidement une grammaire générale: lorsqu' un mot seul (par opposition à une classe) se trouve entre deux états, une procédure recherche son appartenance éventuelle à l' une des classes créées précédemment, et, dans le cas d' un succès, le remplace par une copie de cette classe.

Exemple:

Supposons que la classe "$chien" a été créée précédemment, avec les mots "chat", "chien" et "loup":
($chien chat chien loup)
Alors, la règle

Ti ==> loup Tj

devient

Ti ==> 1chien Tj

où "1chien" est une copie de "$chien".

Cette troisième phase ne réduit donc pas, contrairement à la seconde, le nombre de règles contenues dans la grammaire.

IIIJ LE GENERATEUR ALEATOIRE DE PHRASES

Afin de pouvoir juger de façon objective de l' adéquation de la grammaire générée au langage souhaité, nous avons construit un générateur aléatoire de phrases. Ce générateur est guidé par des probabilités affectées aux divers arcs partant d' un noeud quelconque. Ces probabilités sont calculées en tenant compte du poids de chaque sous-arbre, ce qui donne aux phrases longues la même chance de sortir que les phrases courtes, ce qui n' est pas le cas dans un générateur où les chemins partant d' un noeud sont équiprobables. Ces poids sont donc eux-mêmes calculés récursivement à partir des états finals. Le chemin à l' intérieur de l'arbre est choisi en fonction de ces poids par génération de nombres aléatoires eux-mêmes déterminés par la pression d' une touche du clavier à un instant choisi par l' opérateur, ce qui évite tout risque de périodicité, obligatoire dans tous les cas où la génération des nombres est entièrement automatisée.

IVJ LE ROLE DES EXEMPLES NEGATIFS

La généralisation de n' importe quel fait vérifiable ne va pas sans introduire des assertions fausses dans un système, quel qu' il soit. C' est bien entendu le cas du nôtre. En outre, tout apprentissage "naturel" est fondé sur la réunion d' expériences "positives" ("á répéter") et "négatives" ("á éviter"). Il nous a donc paru utile, puis indispensable, d'introduire dans l'apprentissage un moyen de faire machine arrière en cas de généralisation trop hâtive. La technique utilisée comporte en gros deux éléments:

- d'une part, une deuxième grammaire (générée en même temps que la première) beaucoup moins générale, plus proche donc des seules phrases effectivement rentrées.
- d' autre part, un petit analyseur syntaxique qui travaille en parallèle sur les deux grammaires.

Lorsque le générateur est utilisé, l' analyseur entre en action. Si la phrase générée n' appartient pas à la grammaire 2, le système demande confirmation. S' il l' obtient, la phrase est incluse dans la grammaire 2, pour éviter que la même indéçision se reproduise. Sinon, une procédure de recherche d' erreur est activée qui retire à la grammaire 1 la possibilité d' accepter une telle phrase à l' avenir.

Exemple:

Supposons que la classe $un a été créée précédemment avec les chiffres de 1 à 9, et que l'on rentre l'exemple "dix sept":

```
S ==> dix T1
T1 ==> sept F1
```

devient, par une généralisation "hâtive":

```
S ==> dix T1
T1 ==> $1$un F1
```

où 1un est une copie de $un. Des phrases incorrectes (telles que "dix trois" par exemple) peuvent être admises. Lors de la génération de "dix trois", la grammaire 2 va échouer, le système va donc demander confirmation. La phrase devant être refusée, le mot "trois" va être retiré de 1un, et placé dans une liste au début de cette classe, de façon à refuser cette phrase à l' avenir:
(1un (trois) un deux quatre cinq six sept huit neuf)

VJ TESTS-RESULTATS

Au départ, le système était prévu pour fonctionner avec un état final et un état initial. Cependant, le fait de vouloir rendre la grammaire déterministe est incompatible avec l' exigence d'un état final unique [2], à moins de générer une grammaire bien trop générale, acceptant pratiquement n' importe quoi. Nous avons donc dû modifier en profondeur notre programme, modification qui est à peine terminée, ce qui explique que les tests soient encore peu nombreux. Le système fonctionne cependant de façon satisfaisante sur les chaînes de caractères, et pour ce qui est des phrases, nous l' avons testé sur le petit langage d' application au standard téléphonique utilisé par ESOPE [1]. Le système nous a fourni une grammaire un peu plus générale que celle d' ESOPE, quoique assez proche. Le problème principal posé est celui du temps de calcul, en O(n2), où n est le nombre de règles de la grammaire: en effet, chaque règle doit être comparée à toutes les autres lors de la seconde phase, le temps de traitement augmente donc assez vite. Un autre problème a été de constater que le temps de "convergence" vers la grammaire souhaitée dépend de l' ordre dans lequel on a introduit les exemples.

Cependant, la grammaire obtenue est dans l' ensemble satisfaisante dans tous les tests effectués (il est clair qu' une grammaire acceptant un peu plus de phrases que souhaité est préférable à une autre en acceptant un peu moins!).

VI] CONCLUSION-PERSPECTIVES

Bien que, on l'a vu, le nombre de tests effectués soit encore assez limité, les résultats obtenus sont prometteurs. Afin d'augmenter les performances du système, nous pensons étudier une possibilité d'introduire de façon automatique les exemples dans un ordre optimal, et ceci grâce à l'utilisation d' heuristiques (on a vu en effet que l' ordre d'arrivée des exemples influe sur le temps de calcul).

VII] BIBLIOGRAPHIE

[1] J. MARIANI, "ESOPE: un Système de compréhension de la parole continue.", Thèse d'Etat, Paris, juillet 82.

[2] L. MICLET, "Inférence de grammaires régulières", Thèse 3'cycle, Paris, 1979.

[3] J.R.ANDERSON, "A Theory of Language Acquisition Based on General Learning Principles", Rapport Carnegie-Mellon University, CS-81-129.

[4] J.R. ANDERSON, "Induction of Augmented Transition Networks", Cognitive Science, 77,n0 1, pp 125-157.

[5] D.COULON, D.KAYSER, "Construction of Natural Language Sentence Acceptors by a Supervised Learning Technique", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol PAMI-1, pp 94-99, 1979.

[6] M.RADY, "L'Ambiguité du Langage Naturel est-elle la source du non-déterminisme des procédures de traitement? ", Thèse d'Etat, Paris, juin 83.

[7] P.LAROQUE, "Etude d'une Stratégie Flexible en Compréhension Automatique de la Parole", rapport de DEA, Orsay, 1984.

[8] RYSZARD, R.MICHALSKY, J.G.CARBONELL, T.M.MITCHELL, "Machine Learning: an AI Approach", Palo Alto, Tioga Publ Co, 1983.

UTILISATION D'INFORMATIONS LINGUISTIQUES DANS UN SYSTEME DE DIALOGUE ORAL HOMME-MACHINE

N. CARBONELL, F. CHARPILLET, J.P. HATON, B. MANGEOL
P. MOUSEL, J.M. PIERREL, A. ROUSSANALY

C.R.I.N. - Université de Nancy I
B.P. 239, 54506 VANDOEUVRE LES NANCY

ABSTRACT

Our group has been working for the past twelve years in the design of continuous speech recognition systems based on the cooperation of different knowledge sources, from acoustics to pragmatics. We have particularly designed the MYRTILLE I and II systems for the recognition of continuous sentences in different kinds of languages. In the first part of this paper we give a brief overview of these projects together with the lessons that can be drawn for them. An important point concerns the lack of an actual dialog between the user and these systems (or of other similar systems developed throughout the world). We therefore decided to launch two new projects based on a pragmatic dialog module in two very different application areas, i.e. a dictation machine and a pseudo-natural language inquiry system. These projects are described in the two following parts of the paper. In the dictation machine the linguistic model is a simple syntactic model based on the succession of three syntactic categories and the recognition errors are recovered through a sophisticated man-machine dialog combining voice and graphic communication. The pseudo-natural language inquiry application is a very ambitious, long term project. This project started last year and we present in this paper the present state of an reflection on the use of different kinds of linguistic information and on the integration of the dialog component in the system.

INTRODUCTION

Dans le cadre de la communication orale homme-machine l'un des principaux axes de recherche de notre équipe, depuis maintenant plus de douze ans, concerne l'utilisation d'informations linguistiques en compréhension automatique de la parole continue. L'objectif de cet article est de faire le point sur l'ensemble des travaux que nous avons ainsi menés depuis 1974 et d'expliciter nos orientations actuelles, centrées essentiellement sur deux projets complémentaires : une machine à dicter et un système de compréhension de la langue naturelle.

La compréhension de la langue naturelle orale dans son intégralité ne peut en aucun cas être présentée comme un objectif raisonnable pour nos études ; nous avons donc été amenés à étudier les différents modes possibles de communication orales homme-machine et à distinguer ainsi :

1 la communication par mots (mots isolés, enchaînés ou mots-clés),

2 la communication à l'aide de langages artificiels,

3 la communication en langages quasi naturels, limités à un domaine d'expertise,

4 la dictée automatique.

Si le premier type de communication, le seul actuellement opérationnel d'un point de vue industriel, ne nécessite la prise en compte d'aucune information linguistique autre que phonétique, phonologique ou lexicale, les trois autres types au contraire conduisent à l'utilisation d'informations syntaxiques, sémantiques et pragmatiques. Dans la suite, nous nous intéressons essentiellement à ce dernier type d'informations linguistiques. Après un bilan du projet MYRTILLE qui correspondait aux points (2) et (3) (pour ce dernier uniquement limité à la compréhension de phrases), nous expliciterons les choix mis en oeuvre, tant dans notre projet de machine à dicter que dans celui de compréhension de dialogues quasi naturels.

BILAN DU PROJET MYRTILLE

Compréhension de langages artificiels : le système MYRTILLE I

La communication à l'aide de langages artificiels, particulièrement bien adaptés à des applications industrielles de type commande de processus, utilise comme support un langage entièrement défini par une grammaire à contexte libre. La mise en oeuvre de systèmes de ce type nécessite l'utilisation de trois types d'informations linguistiques :

- des informations syntaxiques pour rendre compte de la structure du langage,
- des informations lexicales limitées à une représentation phonétique des mots pour permettre une recherche lexicale phonétique,
- des informations pragmatiques nécessaires à la mise en oeuvre d'une procédure de dialogue.

Dans de tels systèmes, on peut distinguer deux composantes essentielles : la reconnaissance de phrases et la procédure de dialogue ; le système MYRTILLE I que nous avons développé durant les années 74-76 est de ce type [1]. Les principales caractéristiques d'un tel système sont :

a) l'importance du niveau syntaxique dans l'étape de reconnaissance de phrases. Ce type de système, fondé le plus souvent sur un processus d'hypothèse et test,

peut utiliser deux types de stratégies : l'une descendante qui consiste à émettre, à l'aide d'un analyseur syntaxique, des hypothèses sur les mots à reconnaître, hypothèses testées ensuite par une procédure de recherche lexicale, l'autre ascendante qui détermine à partir d'un treillis de mots préalablement construit la phrase la plus probable compte tenu des contraintes syntaxiques. Quelle que soit la stratégie choisie, le coeur du système met en oeuvre un analyseur syntaxique qui peut lui-même fonctionner de façon ascendante ou descendante, de gauche à droite ou du milieu vers les côtés [2], [3], [4]. Actuellement, on peut noter que les techniques d'analyse syntaxique ainsi mises en oeuvre sont parfaitement au point et permettent même, comme nous l'avions proposé dans MYRTILLE I [1], une certaine liberté par rapport à la syntaxe du langage en cas d'adjonction ou de suppression d'un mot non sémantiquement caractéristique ;

b) la mise en oeuvre d'une procédure de dialogue pour valider en dernier recours les résultats fournis par la reconnaissance, lever les ambiguïtés, corriger les erreurs et réaliser l'action correspondant à la demande de départ. De tels procédures intégrent un ensemble d'informations pragmatiques et peuvent être réalisées de façon purement empirique [1] ou modélisée par un automate de dialogue [5].

Compréhension de langages quasi-naturels : MYRTILLE II

Dans le cas de compréhension de langages quasi-naturels, les informations syntaxiques ne suffisent plus ; il est alors nécessaire de mettre en oeuvre des informations de type sémantique et pragmatique. En effet, la structure du langage à traiter est très vaste (structure du français parlé avec seulement quelques restrictions) et les principales contraintes apportées au langage correspondant à une limitation du vocabulaire possible à quelques centaines de mots spécifiques de l'application envisagée.

Le système MYRTILLE II que nous avons développé [6] a montré la faisabilité de la compréhension de phrases dans ce contexte. Dans ce système, plutôt que de nous appuyer sur une séparation pas toujours très bien définie entre syntaxe, pragmatique et sémantique, nous avons proposé une nouvelle classification possible des informations linguistiques en distinguant :

- la définition de la structure du langage qui regroupe l'ensemble des informations liées au langage (syntaxe et sémantique) dans un réseau à noeuds procéduraux (R.N.P.),
- la définition du vocabulaire de l'application qui regroupe, dans un lexique à hiérarchies multiples, les informations syntaxiques et sémantiques liées à chaque mot,
- les informations propres à la parole contenant, outre la représentation phonétique des mots, des informations telles que longueurs phonétiques, patrons phonétiques et patrons prosodiques.

Le principe de base de la reconnaissance reste le principe d'hypothèse et test, généralisé ici aux divers niveaux de traitement. Pour ce faire, nous avons opté pour une organisation pseudo-parallèle où la hiérarchie entre ces trois sources d'informations prises en compte lors de l'émission des hypothèses est déterminée dynamiquement à l'exécution sur la base de critères qui sont eux-mêmes hiérarchisés afin de déterminer une stratégie type lorsque le système n'a aucun critère de choix.

Les principaux processus mis en oeuvre lors de l'émission des hypothèses sont :

a) STRUCTURE qui restreint les hypothèses compte tenu de la définition syntaxico-sémantique des structures du langage et du contexte de la phrase déjà analysé,

b) SEMAN qui prend en compte les informations fournies par la définition syntaxique des mots et restreint les hypothèses en fonction des dépendances conceptuelles acceptées par la partie de phrase déjà traitée et de la structure reconnue,

c) PROPHON qui pondère ces hypothèses en fonction des traits prosodiques et de la structure phonémique du contexte à reconnaître.

Les tests de MYRTILLE II ont été réalisés sur une application de centre de renseignements météorologiques utilisant un vocabulaire de 400 mots ; elles ont permis de montrer la faisabilité d'une telle reconnaissance et de valider divers algorithmes utilisés dans la phase de reconnaissance. Mais ce système a aussi ses limites, essentiellement dues à la séparation trop artificielle entre l'étape de reconnaissance et compréhension de phrases et celle d'interprétation et gestion du dialogue. De plus, cela nous a permis de mieux poser le problème lié aux restrictions syntaxiques et/ou lexicales. Il faut en effet choisir entre un système acceptant peu de restrictions à ce niveau, au risque de perdre en efficacité et un système utilisant, pour une étape donnée de compréhension de phrases, des restrictions plus importantes spécifiques à chaque situation du dialogue. Ce dernier point nécessite une fusion ou du moins une interaction forte entre la compréhension de phrases et le traitement du dialogue.

Au terme du projet MYRTILLE, nous avons été amenés à orienter nos recherches suivant deux axes complémentaires : un système de machine à dicter et un système de compréhension de dialogue, qui posent un ensemble de problèmes très variés.

PROJET DE MACHINE A DICTER

Il s'agit en partant de l'expérience acquise dans les systèmes MYRTILLE de réaliser un prototype de machine à dicter automatique ou plus précisément de système de traitement de texte utilisant une entrée vocale complétée par une entrée clavier et un système de désignation de type souris pour permettre une correction rapide des erreurs de reconnaissance. En effet, le module de reconnaissance ne pouvant encore résoudre toutes les ambiguïtés, celui-ci propose plusieurs interprétations possibles du treillis phonétique. La sélection est ensuite effectuée manuellement à l'aide de la souris. Le projet se distingue essentiellement des systèmes MYRTILLE I et II par la spécificité du problème de la dictée automatique. En effet, le niveau syntaxique est simplifié par le fait que l'on dicte des textes écrits ce qui permet une meilleure formalisation du langage et la constitution aisée d'un corpus représentatif. La simplification apparaît également au niveau sémantique qui ne devient plus qu'une source d'information d'appoint ; en effet, la compréhension des phrases n'est pas nécessaire

dans l'optique adoptée qui consiste à résoudre les ambiguïtés manuellement (le module de reconnaissance propose plusieurs solutions possibles). Enfin, on peut amener facilement, sans trop de contraintes, le locuteur à une locution soignée et ponctuée par de nombreuses pauses.

Nous avons réalisé un éditeur qui autorise l'utilisation d'un vocabulaire composé de l'ordre de 1 000 racines dans un français courant, sans restriction syntaxique si ce n'est au niveau du modèle qui peut être toujours pris en défaut par une tournure particulière, non apparue dans le corpus. Le système est principalement composé de trois modules :

- un décodeur acoustico-phonétique qui fournit un treillis de phonèmes,

- un module de reconnaissance de phrases,

- un éditeur de texte qui fonctionne en mode parole et en mode clavier aussi bien pour l'édition que pour l'interpréteur de commande. Pour ce qui est le mode commande-parole, une version identifie les chaînes phonétiques par programmation dynamique tant au niveau lexical qu'au niveau syntaxique [7].

Le module de reconnaissance est bâti autour d'un analyseur lexical utilisant des contraintes phonologiques locales. L'analyse lexicale est fondée sur l'évaluation heuristique d'un taux de dissemblance entre le treillis de phonèmes fourni par le décodeur et les formes de référence contenues dans le lexique. Le taux de dissemblance est évalué en utilisant les techniques de programmation dynamique qui permettent, de façon optimale, de récupérer les erreurs de segmentation et d'identification du décodeur. Les règles phonologiques permettent d'évaluer les erreurs de segmentation du décodeur en fonction de ses caractéristiques propres et de règles phonologiques générales.

Exemple 1 : cas de l'élision d'une plosive sourde après une pause :
si phon-prec=silence
si référence=plosive-sourde
alors élision=très-probable.

Exemple 2 : règle permettant de tenir compte de la bonne détection des noyaux vocaliques :
si phon-prec=consonne
si phon-cour=consonne
alors insertion=très-peu-probable

(l'insertion d'une voyelle entre deux consonnes remettrait en cause la détection des noyaux vocaliques).

La stratégie de reconnaissance de phrase utilisée est une adaptation de l'algorithme "level building" [8] avec une recherche en faisceau. Cet algorithme présente l'avantage de ne faire des choix qu'entre des chaînes composées du même nombre de mots, ce qui évite de privilégier la reconnaissance des chaînes composées de mots courts.

Le lexique contient pour chaque mot, d'une part, l'ensemble des classes grammaticales et traits syntaxico-sémantiques qui lui sont attachés et, d'autre part, sa description phonétique. Celle-ci contient les différentes variantes possibles d'un même mot regroupées en trois catégories : élisions, insertions et substitutions, ce qui permet de prendre en compte la plupart des variations inter- et intra-locuteurs, ainsi que les erreurs prévisibles du décodeur.

Du fait de la diversité des textes susceptibles d'être saisis, le modèle syntaxique devait être le moins restrictif possible ; ceci nous a conduit à choisir un modèle syntaxique local constitué par l'ensemble des suites binaires et ternaires de classes grammaticales admissibles. La simplicité de cette approche a l'avantage de permettre un apprentissage automatique du modèle sur un corpus de texte préalablement étiqueté (association pour chaque mot de la classe grammaticale qu'il possède dans le contexte où il apparaît dans le texte). L'étiquetage est engendré automatiquement sur le corpus d'apprentissage en utilisant un modèle markovien [9] comprenant 65 classes grammaticales.

Afin de compléter le modèle syntaxique, un certain nombre de règles grammaticales ont été introduites. Ces règles s'appuient sur les traits syntaxico-sémantiques du lexique. Les règles sont de deux types :
- des règles prédictives qui permettent de déterminer les traits que doit posséder le prochain mot ; exemple : accord du verbe avec le pronom personnel qui précède,
- des règles restrictives ; exemple : vérification de la construction de verbe suivi d'un infinitif.

L'avantage de la double approche, règles et modèle triclasses, est de permettre de limiter le nombre de catégories grammaticales et donc de conduire à une convergence plus rapide du modèle. Signalons par ailleurs que les règles autorisent l'introduction de contraintes syntaxiques d'ordre supérieur à 3.

Le système a été implanté sur un système de développement Exormacs (68000 Motorola). Il est actuellement en cours de test en utilisant le décodeur acoustico-phonétique des systèmes MYRTILLE. Le module de reconnaissance, lorsque le treillis phonétique contient au moins 80% de bons phonèmes, fournit dans 90% des cas la bonne interprétation parmi celles qu'il propose. Entre 70% et 80% de bons phonèmes, il fournit une interprétation valide à un mot près. Ces résultats sont donnés à titre indicatif, le système n'ayant pas encore été testé sur un corpus suffisant.

PROJET COMPREHENSION DE DIALOGUES [10]

Notre objectif est de construire un système informatique susceptible de constituer un interface oral homme-machine efficace et confortable pour l'utilisateur, dans le cadre d'applications informatiques destinées au grand public. Nous envisageons donc un système qui soit capable de comprendre et satisfaire les demandes en provenance d'usagers s'exprimant librement (pas d'apprentissage préalable du langage), à partir d'un dialogue oral finalisé qu'il devra, en partie au moins, gérer et structurer et d'une expérience dans un domaine (tâche et univers de connaissances) relativement simple et circonscrit. Les applications visées correspondent à la consultation de bases de données relativement simples (l'application test que nous avons choisie traite des renseignements administratifs, tels qu'ils apparaissent dans les pages roses de l'annuaire téléphonique).

Très schématiquement, on peut dire que le système

doit d'abord comprendre la demande de l'utilisateur, puis satisfaire cette demande en fournissant, dans le cas de notre application, les informations administratives souhaitées. Ces fonctions comportent toutes deux la compréhension et la construction d'énoncés ainsi que la gestion d'un dialogue finalisé ; elles mettent donc en oeuvre des connaissances de nature identique auxquelles viennent s'ajouter, pour la seconde, une source d'informations spécifique : les connaissances relatives au domaine d'application choisi.

Il ne nous est pas possible d'expliciter ici l'ensemble des choix que nous avons faits pour la mise en oeuvre de ce système ; nous nous limiterons donc à la présentation des sources d'informations linguistiques utilisées (à l'exclusion des informations acoustico-phonétiques) que l'on peut regrouper en trois grands types :

(i) les informations statiques et dépendantes du type d'applications ; ce sont essentiellement :
- des informations sur la structure du langage; nous avons opté pour une représentation sous forme de R.N.P. permettant une analyse du milieu vers les côtés,
- une description des différentes structures prosodiques d'un énoncé : patrons élémentaires, structures complexes,
- une composante lexicale correspondant à un sous-ensemble du vocabulaire général de la langue (mots outils en particulier) et précisant pour chaque mot ses attributs syntaxiques et sa représentation phonétique),
- des connaissances sur les stratégies de contrôle et de gestion de dialogues finalisés sous forme de schéma et structure de dialogues ;

(ii) les informations statiques propres à chaque application avec en particulier :
- une définition du lexique spécifique à l'application traitée associant à chaque mot une composante syntaxique (classes grammaticales possibles), sémantique (sous forme de traits, par exemple : "inanimé", "animé") et phonétique (représentation phonétique et règles d'altération phonologique),
- une description de l'application et de la tâche (suffisante pour permettre de satisfaire les requêtes de l'usager, le système étant en position d'expert) comprenant :
  . une définition des concepts utilisés dans l'application précisant pour chaque concept sa déclaration, son profil et sa classe ; exemple : sexe (personne) [masculin, féminin]
  . les relations entre concepts, sous forme d'expressions de réécriture permettant diverses déductions, par exemple : nationalité (personne, français) → nationalité (personne étranger)
  père (x,y) → (fils(y,x)/fille(y,x)) & masculin(y),
  . la formulation de la base de données concernant l'application sous forme d'une liste de formules instanciées, par exemple :
  délai (obtenir (personne, C.I.), 10 jours)
  couleur (permis de conduire, rose)
  . des informations pragmatiques, liées aux requêtes dans l'univers de la tâche, par exemple : but : obtenir (x, C.I.)
  precond : mineur (x)
  res : accompagner(x,y) 2 parent(x,y)
  . un ensemble de valeurs par défaut ; de même type que les informations sur la base de données ; ces informations interviennent différemment dans la stratégie de raisonnement, par exemple :
  nationalité (locuteur, française)
  ville (Nancy) ;

(iii) des informations dynamiques, modifiables par les processus : elles concernent l'énoncé en cours de traitement ou le dialogue lui-même dans son ensemble ; les différents processeurs peuvent le consulter mais également les mettre à jour ; il s'agit d'hypothèses ou d'interprétations concernant l'énoncé en cours de l'analyse ou l'ensemble du dialogue dont ce dernier fait partie. Ce sont essentiellement l'interprétation syntaxico-sémantique et l'historique du dialogue.

La mise en oeuvre de ces sources d'information, plus particulièrement celles liées à l'application, nous conduit à developper la composante dialogue du système et à lui donner un rôle prépondérant dans les structures de contrôle. Nous pensons que l'utilisation de telles connaissances permettra de simplifier la compréhension des requêtes de l'usager en fournissant des hypothèses sémantiques susceptibles de guider le processus de compréhension ou, tout au moins, de restreindre, pour un énoncé donné, le nombre des interprétations sémantiques à envisager, de combler les lacunes, de lever les ambiguïtés sans systématiquement faire appel à l'usager. Plus généralement, nous escomptons simplifier la compréhension de la demande de l'usager en la réduisant à la sélection, au sein d'un ensemble de schémas a priori de celui (ou ceux) qui rend(ent) compte le plus adéquatement de la requête. Ces connaissances devraient également contribuer à limiter le nombre, le volume et la complexité des échanges au cours du dialogue.

CONCLUSION

Comme nous venons de le voir, le type et le rôle des informations linguistiques à mettre en oeuvre dans un système de compréhension automatique de la parole continue dépend fortement du type de système visé : langages artificiels de commande, machine à dicter ou système de dialogues quasi-naturels finalisés. Notre expérience nous a montré que si l'on possédait maintenant des techniques assez robustes pour conduire à la reconnaissance de phrases, il est nécessaire de développer et d'intégrer beaucoup plus la composante dialogue pour aborder des applications grand public. Enfin si, volontairement, nous avons passé sous silence le problème de l'utilisation d'informations phonétiques, ce n'est pas que nous les considérons comme secondaires ; bien au contraire, l'obtention d'un meilleur décodage acoustico-phonétique reste l'une des principales conditions de réussite de nos projets et nous fondons beaucoup d'espoir sur la démarche systèmes experts initialisée dans notre équipe depuis maintenant plus de deux ans [11].

BIBLIOGRAPHIE

[1] J.M. Pierrel, "Contribution à la reconnaissance automatique du discours continu", Thèse de 3ème cycle, UNiversité de Nancy I, 1975.

[2] J.M. Pierrel, "Un système de compréhension du discours continu utilisant des contraintes morphologiques, syntaxiques et sémantiques", RAIRO Informatique, vol. 12, N° 2, 1978.

[3] J.F. Mari, "Reconnaissance du discours continu par ilôts de confiance", RAIRO Informatique, Vol. 15, N° 2, 1981.

[4] J.P. Haton, R. Mohr, "A parsing algorithm for imperfect patterns and its applications", 3th IJCPR, Coronado, 1976.

[5] S. Lafont, "Etude et mise en oeuvre d'outils informatiques pour non-voyants : console braille et commandes vocales", Thèse de Docteur-Ingénieur, Université de Nancy I.

[6] J.M. Pierrel, "Etude et mise en oeuvre de contraintes linguistiques en compréhension automatique du discours continu (systèmes MYRTILLE I et MYRTILLE II)", Thèse d'Etat, Université de Nancy I, 1981.

[7] F. Charpillet, J.P. Haton, J.M. Pierrel, "Apport de la programmation dynamique en reconnaissance automatique de la parole continue", 4ème congrès AFCET-INRIA RF-IA, Paris 1984.

[8] C. Meyers, S. Levinson, "Connected Word Recognition Using a Syntax-Directed Dynamic Programming Temporal Alignment Procedure", Proc. ICASSP-81, 1981, pp. 956-959.

[9] F. Jelinek and al., "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech", I.E.E.E. Trans. of Inf. Theory, Vol. II-21, N° 3, 1975, pp. 250-256.

[10] N. Carbonell et al., "Dialogue oral homme-machine. Bilan du projet MYRTILLE et perspectives", Actes du Séminaire "Dialogue à composante orale", GRECO-GALF, Nancy, octobre 1984.

[11] N. Carbonell, D. Fohr, J.P. Haton, F. Lonchamp, J.M. Pierrel, "An Expert System for the Automatic Reading of French Spectrograms", Proceedings ICASSP, San Diego, March 1984.

UTILISATION DE CONNAISSANCES ACOUSTICO-PHONETIQUES ET LEXICALES POUR L'IDENTIFICATION DE MOTS DANS LE DISCOURS CONTINU

H. Meloni[1], J. Gispert[2], J. Guizol[3]

(1) Faculté des Sciences, 33 rue Louis Pasteur, 84000 AVIGNON
(2) GIA, Luminy, 70 route L. Lachamp, 13288 MARSEILLE Cédex 9

This paper describes an analytic word recognition system in continuous speech using various fields of knowledge (Acoustical, Phonetic, Phonologic, Prosodic, Lexical) stated in a declarative form in PROLOG. Data is constituted by speech segments (pseudo-phonetic events) distinguished by 27 parameters exhibiting the different phases of a phoneme as well as regular or random transitions between phonetic units. Results are presented in the form of a lattice of the words most clearly recognized in the processed statement. Strategies controlling the recognition process optimise dynamically the use of available information depending on context (Data and Knowledge).

1. INTRODUCTION

Nous présentons un système de traitement de connaissances pour la Reconnaissance Automatique de la Parole Continue (RAPC) qui permet l'identification analytique de mots dans le discours continu.

Les techniques employées dans notre système sont à rapprocher de celles utilisées par ailleurs dans le traitement de connaissances déclaratives acoustico-phonétiques pour la lecture de sonagrammes |1|, |4| ou l'identification d'unités phonétiques |2|,|9|. Toutefois, ces réalisations concernent chacune un domaine précis d'expertise. Notre approche permet d'intégrer dans un même formalisme l'ensemble des connaissances nécessaires au traitement automatique d'un énoncé.

Les données du système sont constituées de segments infra-phonémiques de parole produits de manière algorithmique. Les résultats sont présentés sous la forme d'un treillis de mots vraisemblables.

2. DONNEES DU SYSTEME

Compte tenu des bons résultats obtenus dans la segmentation du signal en portions stables et transitoires dont la durée coïncide avec les quelques étapes (au plus 3 ou 4) de variations significatives et aléatoires dans un phonème, il nous a semblé raisonnable de retenir ce type d'unités comme données pour le système.

Le processus employé est pratiquement identique à celui que nous avons décrit en détail par ailleurs |6|,|7|,|8| ; nous ne donnerons donc ici que les indications nécessaires à la compréhension de l'exposé.

Les paramètres utilisés sont obtenus à partir de spectres lissés calculés au moyen des 14 premiers coefficients de LPC. Conjointement à la segmentation du signal sur des critères de variations spectrales et temporelles, chaque unité est rangée dans l'une des 6 macro-classes suivantes :

- *VO → segment vocalique stable,*
- *VC → segment vocalique transitoire,*
- *OC → segment occlusif suivi d'une explosion*
- *CC → segment constrictif,*
- *CO → segment consonantique,*
- *SI → silence.*

Les unités sont caractérisées au moyen de paramètres pseudo-phonétiques évalués sur la zone de plus grande stabilité spectrale pour les segments VO, VC et CO, sur l'explosion pour les segments OC et sur la portion médiane des segments CC. Les 27 variables retenues sont proches des propriétés d'indices caractérisant les traits acoustiques des phonèmes, elles sont représentées sur des échelles dont la hauteur ajustable permet une normalisation approximative des phénomènes à coder.

3. REPRESENTATION DES CONNAISSANCES

Les connaissances du système sont à chaque instant regroupées d'une part, dans un ensemble fixe de faits, règles et méta-règles exprimés en PROLOG et, d'autre part, dans un treillis en constante évolution qui représente les données et résultats obtenus à une étape quelconque de l'analyse.

3.1. Représentation des données et résultats

Les données et résultats du processus de reconnaissance sont représentés dans le système par une structure en treillis constituée de clauses PROLOG. Les évènements pseudo-phonétiques sont les

éléments initiaux du treillis et notent les points de jonction des futures données. La structure s'enrichit, à chaque étape de la reconnaissance, d'unités nouvelles (noyaux phonémiques, noyaux pseudo-syllabiques, mots, groupes de mots, syntagmes, propositions, phrases, etc.). Les éléments des niveaux les plus bas sont toujours accessibles en un point quelconque de l'analyse ; cela permet de réexaminer à la lumière d'un contexte plus précis les portions d'un énoncé qui auraient été insuffisamment ou mal traitées.

3.2. Règles caractérisant les connaissances

Toutes les informations autres que les données et résultats sont représentées au moyen de clauses PROLOG. Le prédicat d'appel de chaque règle permet de distinguer diverses classes de connaissances qui sont susceptibles d'être évoquées dans certains contextes. Le corps de chaque règle est composé généralement de prédicats définis pour effectuer de multiples traitements sur les paramètres (fonctions logiques et numériques simples, calcul de distances, tests divers, manipulation de listes, déplacements dans le treillis,etc.).

Les connaissances sont données sous forme déclarative ; des clauses de contrôle les activent au moment opportun. Si la stratégie proposée par l'interpréteur PROLOG convient dans certains cas pour l'enchaînement des règles, nous avons été conduits à interposer entre les règles et les clauses de contrôle des méta-règles permettant de définir déclarativement les diverses manières d'utiser les clauses terminales.

4. CONNAISSANCES UTILISEES

Les connaissances utilisées peuvent difficilement se regrouper en rubriques disjointes car leurs imbrications sont nombreuses : acoustiques et phonétiques, acoustiques et prosodiques, etc.. Nous les présenterons cependant pour plus de clarté dans des paragraphes distincts.

4.1. Connaissances acoustiques

Ces connaissances n'apparaissent pas toujours directement dans le système. Elles permettent de définir, à travers le modèle paramétrique, les évènements pseudo-phonétiques qui constituent les données.

Les connaissances acoustiques sont étroitement liées, par l'intermédiaire des paramètres, à toutes les règles qui définissent les limites et la substance des unités phonémiques et prosodiques: noyaux phonémiques et syllabiques, traits pseudo-phonétiques, accentuation, frontières prosodiques, etc... Elles sont donc implicites dans la plupart des clauses qui représentent les informations phonétiques du système.

4.2. Connaissances phonétiques

Ce sont les connaissances les plus nombreuses du système ; elles permettent de caractériser les sons, par comparaison directe ou au moyen de macro-traits pseudo-phonétiques, et d'interpréter phonétiquement les données de manière inductive ou prédictive.

4.2.1. Caractérisation des phonèmes.

Afin de permettre le calcul d'une distance entre les éléments caractérisant un segment ou une séquence de segments et une unité phonémique pressentie, nous avons défini chaque phonème par l'affectation de valeurs particulières à ses paramètres stables et représentatifs. Les variables retenues et leurs valeurs diffèrent suivant le locuteur ; un processus d'apprentissage automatique permet d'ajuster au mieux ces connaissances. Ces informations caractérisent moins la représentation idéale d'un phonème (qui serait totalement indépendante du contexte) qu'une situation intermédiaire constituée à partir d'un ensemble de ses "bonnes" réalisations.

Exemple :

parametres(on,7.16.16.18.7.6.nd.nd.nd.5.5.7.nd.4.
nd.nd.0.nd.3.8.nd.nd.8.9.nd.6.5)→
parametres(mm,nd.15.12.16.7.4.5.nd.0.8.4.9.nd.1.
nd.nd.0.8.9.nd.nd.nd.nd.nd.8.8.4)→

où *mm* et *on* sont les noms des phonèmes */m/* et */ɔ̃/*.

Par ailleurs, chaque phonème est également défini de manière formelle par un ensemble de traits : *aigu, grave, ouvert, mi-ouvert, fermé, mi-fermé, compact, nasal, diésé, bémolisé* pour les voyelles ; *occlusif, constrictif, nasal, semi-consonne, liquide, sourd , sonore, aigu, grave, compact, diffus, strident* pour les consonnes. Ces attributs symboliques sont à mettre en relation avec les traits pseudo-phonétiques que nous affectons aux noyaux phonétiques (paragraphe 4.2.3.).

4.2.2. Interprétation inductive des segments

Il s'agit ici de connaissances qui permettent d'interpréter certains segments ou suites de segments bien caractérisés sans autres informations contextuelles que celles présentes dans la séquence des évènements pseudo-phonétiques.

Les premiers éléments définis par ce type de connaissances sont les noyaux phonémiques déterminés par des règles de la forme :

```
noyau-vocalique(s) →
    segment-vocalique(s)
    vocalique(s)
    precede(s1,s)
    segment(s1)
    precede(s,s2)
    segment(s2)
    moins-vocalique(s1,s)
    moins-vocalique(s2,s)

noyau-consonantique(s) →
    segment-consonantique(s)
    non(transition(s));
```

D'autres segments reçoivent une interprétation qui ne tient compte que des données,ce sont des évènements transitoires que l'on ne saurait considérer comme des parties significatives d'une voyelle ou d'une consonne.

4.2.3. Macro-traits pseudo-phonétiques

Les noyaux vocaliques et consonantiques sont affectés de macro-traits pseudo-phonétiques afin d'indiquer de façon aussi précise que possible le type de son qu'ils représentent. Ils constituent les éléments d'un filtre phonétique limitant les comparaisons lors de la phase d'identification des phonèmes . Nous disposons de 14 traits pour les noyaux vocaliques : 6 *degrés d'ouverture* non disjoints, *aigu/grave, compact/écarté, nasal/oral, diésé/bémolisé* et 13 traits pour les consonnes : *occlusif, constrictif, sourd/sonore, aigu/grave, compact/écarté, interrompu/continu, buzz, bruité/non bruité.*
Exemple : *"determination du trait grave/aigu".*

```
acuite(<u,k,<s,t>>,voy-aigue) →
        ou(superieur(maximum-energie(t),3),
           inferieur(moment-spectre(t),6),
           superieur(second-pic(t),5).
           inférieur(difference-ap1-ap3(t),6));
```

4.2.4. Ajustement d'une séquence de phonèmes aux segments

Ces connaissances définissent toutes les manières d'associer aux phonèmes d'une séquence quelconque les évènements pseudo-phonétiques qui les caractérisent le mieux. Une version initiale du modèle est produite de manière automatique |3| par l'examen de phrases choisies pour caractériser les situations phonémiques probables.

Exemple :

```
regle-phon-seg(x.p.y,s1.s2.s1,s2,s2) →
          cu(con-interrompue(p),
             cons-constrictive(p))
          voyelle(x)
          segment-vc(s1)
          segment-consonantique(s2);
```

4.3. Connaissances phonologiques

Des connaissances phonologiques ordinairement traitées dans un système de RAPC, nous n'avons retenu, dans cette étape d'identification de mots, que les informations permettant de décrire de manière remontante des pseudo-syllabes et traiter les problèmes posés par les réalisations diverses de certaines unités phonémiques (/ə/ caduc par exemple).

4.3.1. Définition des pseudo-syllabes

La suite de données composant une pseudo-syllabe est formée d'un noyau vocalique précédé de tous les segments qui le séparent du segment le plus fermé. Ce dernier est intégré dans la séquence proposée. Cette chaîne d'unités est ensuite décomposée en une ou plusieurs sous-chaînes phonémiques dont le noyau consonantique initial et le noyau vocalique final (avec les phonèmes probables en ce points du discours) constituent des accès lexicaux vraisemblables.

4.3.2. Altérations contextuelles des phonèmes

Le traitement des phénomènes de liaison et d'élision suppose la connaissance de contextes graphiques, phonémiques, prosodiques et syntaxiques. Nous avons donc retardé à une étape ultérieure l'application des règles définissant ces informations. L'identification des mots est effectuée sur leur portion inaltérable par l'environnement, les phonèmes restants, introduits par une liaison ou non élidés, seront examinés dans une autre phase de la reconnaissance.

Pour résoudre simplement le problème des réalisations du /ə/ caduc à l'intérieur des mots, nous avons d'une part toléré sa disparition dans tout contexte phonémique où sa présence n'est pas obligatoire et, d'autre part, interdit les accès lexicaux à partir des pseudo-syllabes contenant cette voyelle.

4.4. Connaissances lexicales

Ces connaissances, dans le cadre du système que nous proposons, sont réduites à la description graphique et phonémique des mots et au codage d'un certain nombre d'accès précompilés.

4.4.1. Description des mots.

N'ayant pas introduit de règles de morpho-phonologie, chaque mot est intégralement représenté par sa forme graphique et la liste des phonèmes qui le composent. Les phonèmes de liaison sont particularisés afin d'en faciliter le traitement.

Exemples :

```
<mot33,petit,pp.ee.tt.ii.tt_l.nil> → ;
<mot253,montagne,mm.on.tt.aa.gn.ee.nil> → ;
```

4.4.2. Accès au moyen des pseudo-syllabes

Lors du codage du lexique, toute séquence phonémique comportant une voyelle différente du /ə/ muet et précédée d'au moins une consonne est précompilée en un accès au mot correspondant.Ces informations constituent des filtres phonologiques de sélection des items lexicaux.

Exemple : les accès au mot *montagne* sont

```
<lx-acces-phon,mm.on,mot253> → ;
<lx.access-phon,tt.aa,mot253> → ;
```

5. CONTROLE DU PROCESSUS DE RECONNAISSANCE

Un grand nombre de stratégies, parfois très sophistiquées |5||10| , ont été proposées pour effectuer la reconnaissance de mots dans le discours continu. Nous avons essayé, dans ce système, de simuler les méthodes empiriques que nous utilisions pour identifier, à partir de paramètres visualisés, les sons de la langue dans un ensemble de phrases dont certains mots nous étaient connus.

5.1. Identification des noyaux phonémiques

Cette phase de la reconnaissance permet de repérer dans la séquence d'évènements pseudo-phonétiques des segments particuliers qui caractérisent sans trop d'ambiguités certaines phases de phonèmes. Un ensemble de traits pseudo-phonétiques est associé à chaque noyau à partir duquel on propose la liste des phonèmes les mieux valués. Ces informations sont ajoutées dans le treillis sous la forme suivante :

```
<unite95,<35,36>, voyelle(fer-mi,fer,voy-grave.nil,
                          unite96)> → ;
```

<unite96,oui83> → ;
<unite93,au84> → ;
<unite96,on,70> → ;

où le noyau considéré caractérise le phonème /$\tilde{o}$/de *montagne*.

Les règles actuellement disponibles ont permis de localiser correctement 93 % des noyaux phonémiques puis de proposer pour ces derniers le bon phonème dans 70 % des cas en première position et 85 % dans les 3 premiers meilleurs choix.

5.2. Identification des pseudo-syllabes d'accès aux mots

Cette étape consiste à rajouter dans le treillis, les pseudo-syllabes constituant des accès probables à des éléments du lexique. Pour ces nouvelles unités nous calculons une valuation en fonction de celles de la consonne initiale et de la voyelle qui la composent. Lorsque le taux de vraisemblance est compris dans la fourchette valide dans le contexte où s'effectue cette opération et qu'il existe un item lexical comportant cette pseudo-syllabe, nous ajoutons dans le treillis les informations qui permettront de sélectionner les mots à envisager. Parmi tous les accès possibles aux mots constituant l'énoncé, nous en obtenons effectivement 76 %. Considérant qu'un mot est accessible dès l'instant où l'une de ses pseudo-syllabes a été identifiée, 85 % des mots possibles sont traités.

5.3. Vérification des mots

Pour chaque mot sélectionné, nous opérons un alignement des phonèmes et des données à partir de la pseudo-syllabe d'accès. Ensuite, les règles de mise en correspondance des unités phonémiques et des évènements pseudo-phonétiques nous permettent de calculer, pour la séquence de phonèmes, un taux de corrélation avec la portion d'énoncé traitée. Lorsque cette valuation est satisfaisante, le mot est ajouté dans le treillis sous la forme suivante :

<unite147,<33,42>,<mot,mot253>> → ;

Parmi les mots accessibles par cette technique et appartenant effectivement à l'énoncé,80 % ont été reconnus.

6. CONCLUSION

Le système que nous avons réalisé,bien qu'il soit conçu pour s'intégrer dans un processus plus général de RAPC fournit,sans utilisation de connaissances linguistiques de niveau supérieur (prosodie, syntaxe, sémantique, pragmatique), dès résultats très satisfaisants. Toutefois,le nombre des accès lexicaux (pseudo-syllabes) proposés dans la phase remontante demeure relativement important et nous nous employons à le réduire en affinant les règles d'interprétation phonémique.

BIBLIOGRAPHIE

|1| N.Carbonell,D.Fohr,J.P.Haton,J.M.Pierrel, F.Lonchamp.*Système expert de décodage acoustico-phonétique et invariance;* 13èmes JEP,Bruxelles 28-30 mai 1984.

|2| R. De Mori,M. Gilloux,G.Mercier,M.A. Simon, C. Tarridec,J.Vaissiere,D. Gillet,M.Gérard. *Integration of acoustic,phonetic,prosodic and lexical knowledge in an expert system for speech understanding;* Congrès ICASSP.1984.

|3| J. Guizol,H.Meloni. *Apprentissage de règles d'interprétation d'évènements pseudo-phonétiques;* 13èmes JEP,Bruxelles 28-30 mai 1984.

|4| L.F.Lamel,V.W. Zue. *Properties of consonants sequencies whithin words and accross words boundaries ;* Congrès ICASSP.1984.

|5| W.A. Lea (Ed.).*Trends in speech recognition ;* Englewood Cliffs, Prentice-Hall,1980.

|6| H. Meloni. *Etude et Réalisation d'un Système de Reconnaissance Automatique de la Parole Continue ;* Thèse de Doctorat d'Etat , Université d'Aix-Marseille II,Faculté de Luminy, février 1982.

|7| H. Meloni,J. Guizol. *Identification d'évènements pseudo-phonétiques pour la reconnaissance automatique de la parole ;* 11èmes ICA, Toulouse, juillet 1983.

|8| H. Meloni. *Traitement des contraintes linguistiques en reconnaissance de la parole ;* TSI Vol. 2 n° 5, septembre-octobre 1983, 349-363.

|9| J. Vaissière, M. Gilloux, C. Tarridec, M.A. Simon. *PROSEIDON : détection automatique des indices prosodiques contenus dans la parole continue ;* 13èmes JEP, Bruxelles 28-30 mai 1984.

|10| W.A. Woods. *Optimal Search Strategies for Speech Understanding Control ;* Artificial Intelligence 18,1982,295-326.

UNE STRATEGIE AUTO-ADAPTATIVE DE TRAITEMENT DE LA PAROLE

D. Béroule

LIMSI (CNRS) BP 30
91406 ORSAY Cedex

ABSTRACT

This paper aims at describing the processing strategy implemented in our model of Adaptative; Dynamic and Associative Memory. After a survey of the strategies used in current speech understanding systems, or proposed by psycholinguistic studies, the principles of this model are described.

We then focus on its learning strategy, that leads to a topological representation of the external events, and that involves an identification process based on direct access to the internal representation.

First results of a computer simulation, and a project of auto-adaptive speech understanding system derived from the model, are also presented.

INTRODUCTION

Déterminer une stratégie de traitement automatique de la parole, c'est chercher à combiner de façon optimale les différentes sources de connaissances impliquées dans l'identification du message vocal. Traditionnellement, la stratégie employée est définie par la façon dont la procédure de reconnaissance scrute les composants du message vocal et les différents niveaux de la base de connaissance, ainsi que par le mode de gestion de la mémoire de travail. On parle donc plutôt de "stratégie de reconnaissance" que de "stratégie d'apprentissage", la complexité de la conduite de la reconnaissance répondant à la simplicité de la phase d'acquisition de données.

Bien que quelques tentatives aient été effectuées pour fournir aux systèmes une certaine autonomie quant aux choix de représentants pour les formes traitées, l'apprentissage est toujours circonscrit à une période restreinte, ce qui interdit toute possibilité d'adaptation en cours de traitement.

On peut cependant envisager un apprentissage élaboré, dont les instants et mode d'intervention lors du traitement sont décidés par le système en fonction de critères pré-établis. Ce type d'apprentissage est mis en oeuvre dans notre modèle Adaptatif, Dynamique et Associatif de Mémoire, dans lequel l'identification d'une forme résulte simplement de l'accès direct à sa représentation interne.

Après un bref rappel des différents types de stratégies utilisés dans les système actuels de compréhension de la parole continue, ou proposés par les psycholinguistes, nous présentons les principes du modèle. On montrera notamment comment sa stratégie de traitement évolue avec la représentation interne des formes.

STRATEGIES DE RECONNAISSANCE

Psycholinguistes et informaticiens chercheurs en traitement de la parole se penchent avec un intérêt égal sur les problèmes de stratégie de compréhension, cependant leurs objectifs diffèrent. Les Psycholinguistes réalisent des expérimentations pour éprouver certaines hypothèses concernant l'activité du système humain et faire progresser la connaissance de ce système, alors que les informaticiens cherchent essentiellement à améliorer les performances de leur algorithme de reconnaissance, dont la ressemblance éventuelle avec des processus psychologiques n'est que secondaire.

On peut adopter plusieurs attitudes vis-à-vis des résultats et modèles proposés par la psycholinguistique... La première est de les ignorer en considérant que les solutions adoptées par le système humain ne s'appliquent pas a priori à l'ordinateur, les contraintes matérielles étant trop différentes. La seconde est de s'inspirer de procédures de traitement suffisamment globales pour être exploitées par des moyens informatiques. La troisième est de mettre à profit ces résultats au moyen d'un système fondé sur un mode de représentation et d'utilisation de l'information plus proche du modèle humain. C'est cette dernière approche que nous avons choisie.

Résultats et Modèles de Psycholinguistique

Le fait que différents types d'informations interviennent dans le processus de compréhension d'énoncés est unanimement reconnu. Des sources de divergences proviennent cependant de l'incertitude concernant la manière dont ces informations cohabitent dans le lexique mental. Ainsi, le rôle accordé aux informations d'ordre supérieur est variable : celles-ci pourraient intervenir pour prédire les événements ultérieurs et interdire

certaines solutions. Pourtant des expériences semblent montrer que l'accès lexical s'appuie uniquement sur les niveaux inférieurs, et que les contraintes contextuelles agissent uniquement à l'issue d'un traitement ascendant autonome. La majorité des contributions à cette thèse provient de l'étude des effets de contexte sur l'interprétation des polysèmes qui consiste à déterminer si toutes les significations d'un mot ambigu sont atteintes lorsque le contexte est fortement contraignant |1|

D'après les résultats de cette étude, le contexte sémantique n'améliore pas la rapidité du traitement, comme ce serait le cas s'il intervenait au cours de l'accès au lexique pour éliminer les significations impropres d'un polysème.

En plus de ces résultats, des modèles fournissent des descriptions qualitatives de stratégie de compréhension...

- Le modèle de recherche autonome de Forster |2|

Ce modèle met en oeuvre une procédure de recherche à l'intérieur du lexique qui est très comparable à la scrutation séquentielle d'une zone-mémoire par l'ordinateur. L'information sur les mots est contenue dans un fichier principal dont les entrées sont accessibles par l'intermédiaire d'un certain nombre de fichiers, comme par exemple le fichier phonétique. Celui-ci possède une série de descriptions phonétiques des mots, chacune d'elles étant associée à un pointeur qui permet l'accès direct à l'entrée correspondante du fichier principal. La scrutation séquentielle du fichier phonétique s'interrompt dès qu'un candidat est trouvé ; les descriptions phonétiques des mots les plus fréquents sont donc placée en tête de fichier, afin d'optimiser la procédure de recherche.

Lorsqu'on accède de cette façon à une entrée du fichier principal, aucune information de niveau supérieur n'est mise à contribution. Syntaxe et sémantiques interviennent à l'issue de ce processus autonome pour participer à un contrôle post-accès lexical qui consiste à vérifier l'adéquation du mot-candidat avec le stimulus initial et le contexte.

- Le modèle des Logogènes de Morton |3|

Etabli à partir de l'observation des sujets aphasiques, ce modèle propose un lexique partitionné en zones de traitements spécialisés tels que : perception auditive, perception visuelle, traitement syntaxique, production du langage... La destruction d'une partie du système, la rupture de liaisons entre zones expliquent ainsi un cas pathologique particulier. Cette spécialisation du traitement est conduite jusqu'au niveau de chaque mot, associé à une procédure appelée Logogène (du grec Logos: fabrique de discours).

Chaque Logogène contient la description complète d'un mot ; il fonctionne comme un compteur qui s'incrémente chaque fois que le résultat des analyses sensorielles ou contextuelles correspond à une composante de la description. Chaque compteur admet un seuil dont le dépassement caractérise la reconnaissance du mot associé.

- Le modèle de Cohortes de Marslen-Wilson |4|

Ce modèle propose essentiellement une stratégie gauche-droite d'accès au lexique, avec interaction des niveaux inférieurs et supérieurs. La sélection initiale d'un ensemble de candidats, ou Cohorte, s'effectue à partir d'informations perceptives sur le début du mot à identifier. Tous les mots admettant le même début sont activés simultanément. Puis au fur et à mesure que se déroule la séquence de sons composant le mot, les éléments qui découvrent une incompatibilité entre leur représentation et l'entrée du système s'excluent de manière irréversible de la cohorte, dont la taille est ainsi progressivement réduite. Lorsqu'un seul candidat subsiste, il s'agit du mot reconnu.

Chaque processeur associé à un mot est également capable d'analyser sa validité syntaxique et sémantique au cours de l'accès lexical. Les membres de la cohorte incompatibles avec le contexte sont ainsi éliminés, ce qui permet d'obtenir plus rapidement un candidat unique.

Stratégies employées en Traitement automatique de de la parole

Les systèmes de compréhension de parole continue permettent l'évaluation effective de certaines stratégies. Celles-ci peuvent être classées en fonction du mode d'interaction des informations contenues dans la base de connaissance, en fonction également de la manière de scruter les composants du message à reconnaître et leurs représentations internes.

- Interaction des niveaux de traitement

Le message vocal est porteur de plusieurs informations de nature différente qui doivent collaborer à son identification. Entre les paramètres acoustiques et l'énoncé complet, on peut distinguer des unités de décision appartenant à des niveaux intermédiaires (acoustique, phonétique, lexical). Les niveaux supérieurs contiennent des informations sur la structure des énoncés (niveau syntaxique) et leur signification (niveaux sémantique et pragmatique).

Etant donné les performances encore modestes du décodage acoustique-phonétique, il semble actuellement préférable de prendre en compte les contraintes de niveau supérieur qui limitent le nombre de combinaisons possibles des entités à reconnaître |5| . Cette approche descendante présente pourtant l'inconvénient de ne pas admettre de nouvelles combinaisons, non établies par avance. S'il n'est pas capable d'intégrer dans sa représentation un événement imprévu (apprentissage par *Différenciation*), un système obéissant à une stratégie descendante est cependant capable de l'assimiler à un évènement déjà représenté (apprentissage par *Généralisation*). L'apprentissage par différenciation est seulement envisageable avec une approche ascendante, qui prend en compte l'ensemble des interprétations possibles avant d'utiliser les contraintes de niveau supérieur |6| |7|.

- Mode de scrutation des représentations

Dans les systèmes actuels, la reconnaissance d'une forme acoustique repose sur la comparaison des composants de cette forme avec des représentations stockées préalablement dans la base de connaissance.

Une fois le signal de parole transformé en une séquence non déterministe de symboles (paramètres acoustiques, phonèmes), il est en principe possible de débuter l'analyse en n'importe quel point de la séquence. La méthode la plus fréquemment utilisée est de suivre l'ordre chronologique en commençant par les premiers symboles suivant une stratégie gauche-droite. Mais le caractère indéterministe du décodage acoustique peut inciter à considérer en premier lieu les composants de la séquence les mieux identifiés, contituants des "ilôts de confiance" sur lesquels s'appuie l'analyse |8| .

A chaque instant, il est possible de conserver le meilleur candidat, quitte à revenir sur cette décision si les comparaisons ultérieures échouent |9|. En conservant plusieurs candidats à chaque instant, le retour arrière est évité

PRESENTATION DU MODELE

D'inspiration Psychophysiologique |10| , ce modèle met en oeuvre un mode de représentation topologique très éloigné de celui auquel nous habitue la pratique de l'informatique, et qui appelle donc quelques éclaircissements préalables. Nous décrirons ensuite sa structure, qui correspond à des fonctions déterminées.

Représentation codée et représentation topologique

On peut distinguer deux modes d'exploitation d'un espace-mémoire, auxquels correspondent deux modes de représentation de l'information.

Chaque classe de faits associée à un élément de mémoire M peut être représentée par un *code* spécifique contenu dans M (Représentation codée), ou par la *position* de M à l'intérieur de l'espace-mémoire (représentation topologique).

- Représentation codée

La représentation codée est la moins contraignante en ce qui concerne la procédure de stockage, puisque le code associé à un fait à mémoriser peut être indifféremment placé dans n'importe quel élément M disponible. Par contre, l'exploitation de ce type de représentation nécessite l'existence d'une procédure de recherche qui examine séquentiellement chaque élément M. Cet examen consiste en une série de comparaisons entre un code à identifier et les codes stockés préalablement. Une comparaison aboutissant à une décision positive caractérise la reconnaissance du fait associé au code identifié.

- Représentation topologique

La représentation topologique implique une forme élaborée de stockage qui doit prendre en compte l'ensemble des faits déjà mémorisés pour déterminer la position spécifique (x_i, y_i, z_i) d'un nouveau fait.

Chaque élément de mémoire $M_i(x_i, y_i, \ldots)$ possède un marqueur non spécifique placé dans un certain état lorsqu'un évènement externe est identifié. Une façon de positionner cet indicateur est d'utiliser la valeur d'autres indicateurs présents en différents points de la mémoire, par exemple au moyen d'un superviseur effectuant des opérations logiques du type :

SI $M_i(x_i, y_i, z_i)$ ET $M_j(x_j, y_j \ldots) \ldots$ ALORS $M_k(x_k, \ldots)$

Le traitement peut également être distribué à l'intérieur d'une mémoire rendue associative et dynamique en reliant physiquement les éléments et en leur donnant la faculté de propager les marqueurs le long de ces liens.

Si chaque marqueur admet plusieurs niveaux rendant compte de l'intensité des évènements traités, il faut introduire un seuil qui établit la séparation entre une hypothèse (à confirmer) et un évènement réalisé (marqueur à propager). Chaque élément M devient ainsi une unité de décision qui envoie en d'autres points du réseau un marqueur (ou niveau d'energie) si la somme de ses entrées dépasse un certain seuil.

A une classe de configuration de l'entrée doit correspondre l'activation automatique d'un élément M de coordonées spécifiques, point de convergence d'autres éléments activés. Le processus d'identification procède donc d'un accès direct à la représentation interne ; sa réalisation convenable dépend de la structure du réseau, établie par apprentissage.

Structure du réseau

On considère une mémoire à représentation topologique comme une boîte noire admettant un certain nombre d'entrées et de sorties. Le problème est alors de trouver la structure en réseau qui détermine la mise en activité d'une sortie spécifique pour chaque classe de configurations des entrées par diffusion d'énergie à travers le réseau. La solution que nous proposons est adaptée au traitement de configurations qui évoluent au cours du temps.

Le principe de base sur lequel repose le fonctionnement du modèle est de supposer que la propagation d'énergie le long d'un chemin du réseau a besoin pour se réaliser d'être stimulée à différents instants par des unités de mémoire qui convergent vers ce chemin de façon ordonnée ; ces unités "stimulantes" sont associées à des évènements externes.

Chaque élément dynamique du chemin se contente de pré-activer l'élément suivant, c'est-à-dire de l'activer en-déssous de son seuil de propagation ; la convergence simultanée d'autres niveaux d'énergie vers cet élément est nécessaire pour que son seuil soit dépassé et qu'il propage vers son descendant, qui se trouve à son tour préactivé. Ce processus se renouvelle jusqu'à l'unité située à l'extrémité du chemin, dont la mise en activité

Caractérise la réalisation d'une séquence d'événements.

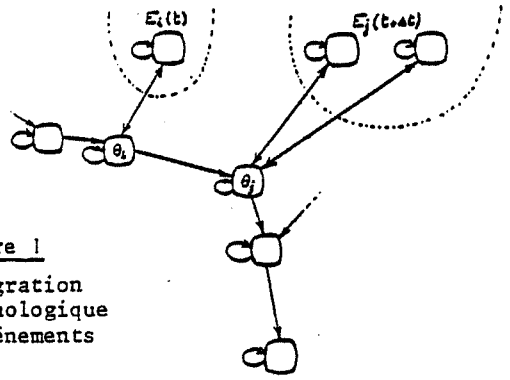


Figure 1

Intégration chronologique d'événements

On peut maintenant remarquer que le rôle de chaque unité composant le chemin est d'amorcer le fonctionnement de l'unité correspondant à l'événement suivant, phénomène assimilable a une prédiction. Or, plusieurs événements sont alors susceptibles de se réaliser, ce qui nous conduit à compléter cette chaîne d'intégration en la rendant arborescente (figure 2).

Chaque unité pré-activée constitue ainsi une hypothèse concernant les événements susceptibles de se réaliser dans un certain intervalle de temps. Pour désactiver les hypothèses non confirmées, on fait appel à un principe d'organisation bien connu en neurophysiologie sous le terme d'inhibition latérale ; dans le modèle, il se matérialise par des liens négatifs entre les branches d'arborescence. Ainsi, la réalisation effective d'un fait prédit va provoquer la diffusion d'énergie le long d'une branche et désactiver les branches voisines. Enfin, des boucles rétro-actives reliant les feuilles d'arborescence à la racine permettent de relancer le processus d'intégration chaque fois qu'une séquence de faits a été filtrée.

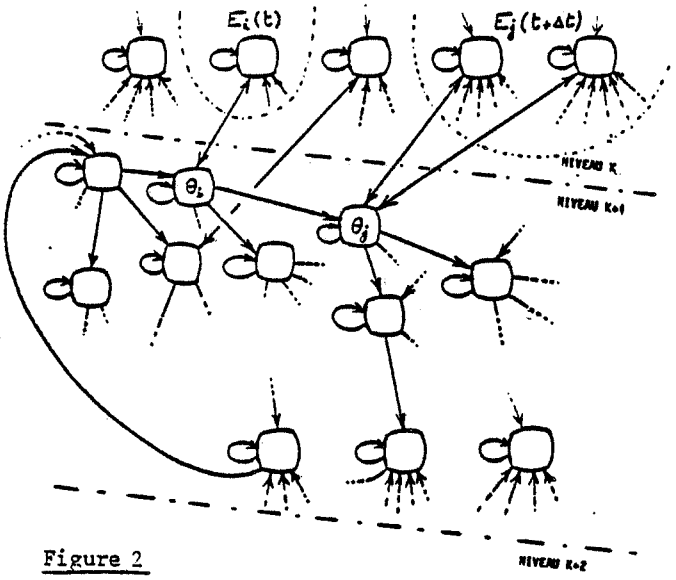
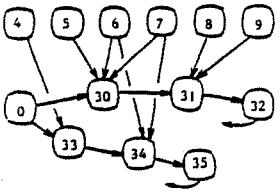


Figure 2

Un niveau d'intégration



Exemple de traitement effectué par la simulation informatique du modèle.

```
                                                        niveau d'activité
                                                                  | seuil
Evenement No 1    0005 = 9, 0006 = 5, 0007 = 1                    |
))PROPAGATION de 0005 (30:666) Vers 0030 (166/133)                |
))PROPAGATION de 0006 (16:533) Vers 0030 (219/133), 0034 (68/308) |
))PROPAGATION de 0007 (3:100) Vers 0030 (229/133), 0034 (79/308)  |
--          DESACTIVATION     Niveau=1                            |
++      RENFORCEMENT: 0030    Fact. trans.: 296, 226, 198, 114    Exci
))PROPAGATION de 0030 (229:1000) Vers 0031 (128/158)

Evenement No 2    0008 = 8, 0009 = 2
))PROPAGATION de 0008 (40:666) Vers 0031 (194/158)
))PROPAGATION de 0009 (10:333) Vers 0031 (227/158)
--          DESACTIVATION     Niveau=1    0030, 0033, 0034
++      RENFORCEMENT: 0031    Fact. trans.: 319, 226, 156    Excitabil
))PROPAGATION de 0031 (227:1000) Vers 0032 (100/83)
--          DESACTIVATION     Niveau=1    0031
                                          (*(*(*REPONSE*)*)*)  .../BCD.
++      RENFORCEMENT: 0032    Fact. trans.: 296    Excitabilite: 120
))PROPAGATION de 0032 (100:566) Vers 0000 (100/50)
--          DESACTIVATION     Niveau=1    0032
++      RENFORCEMENT: 0000    Fact. trans.: 212, 100    Excitabilite:
))PROPAGATION de 0000 (100:1000) Vers 0030 (296/341), 0033 (296/308)
```

UNE STRATEGIE D'APPRENTISSAGE

Suivant la terminologie utilisée pour caractériser une stratégie de reconnaissance, l'accès aux représentations internes des composants du message vocal s'effectue dans notre modèle par une approche ascendante gauche-droite avec activation parallèle de candidats. Nous décrivons maintenant la stratégie d'apprentissage autonome qui intervient constamment au cours du traitement.

Généralisation et Différenciation

Les événements susceptibles de se réaliser à chaque instant sont représentés dans le réseau par des unités pré-activées. Un événement imprévu ne participe pas à l'activation complète de l'une de ces unités et ne guide donc pas la propagation d'énergie le long d'un chemin spécifique. La prise en compte de cet événement revient à compléter le réseau de deux manières possibles.

- Généraliser

L'alternative la moins coûteuse en nombre d'unités nouvelles à mettre en oeuvre et le nombre de liens à créer, consiste à assimiler l'événement en question à l'un de ceux qui étaient prédits. Cette opération revient à étendre la classe des configurations associées à une réponse spécifique (figure 4).

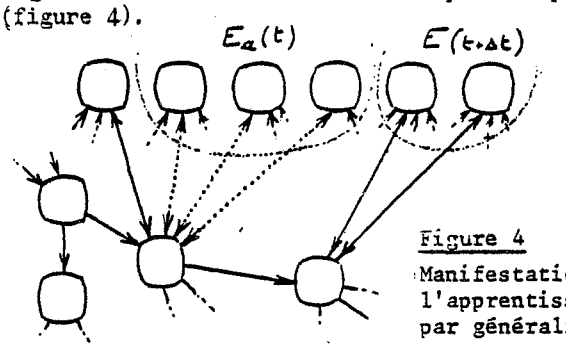


Figure 4

Manifestation de l'apprentissage par généralisation

(les liens représentés par des pointillés sont créés par apprentissage)

- Différencier

L'autre alternative est de considérer au contraire que les hypothèses générées ne correspondent pas à l'événement réalisé, qui se produit pour la première fois dans ce contexte et admet donc une signification différente. En pratique, une branche issue de la dernière unité activée et aboutissant à une nouvelle unité-réponse reçoit les unités associées à l'événement différenciateur et à ceux qui le suivent. Ce processus se termine avec la séquence d'événements par un bouclage reliant la nouvelle feuille d'arborescence (unité-réponse) à la racine (figure 5).

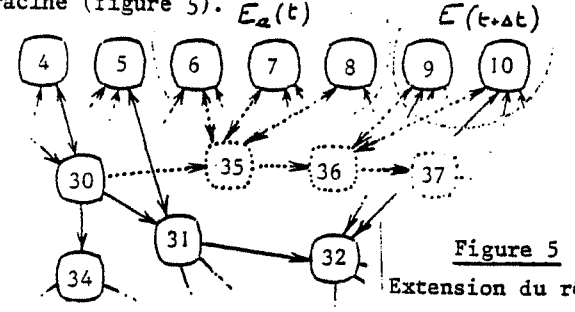


Figure 5
Extension du réseau

```
Evenement No 1   0004 = 9
))PROPAGATION de 0004 (90:666) Vers 0030 (697/575)
--        DESACTIVATION     Niveau=1     0000
++     RENFORCEMENT: 0030     Fact. trans.: 608, 453     Excitabilite: 1
))PROPAGATION de 0030 (697:1000) Vers 0031 (170/200), 0034 (296/375)

Evenement No 2   0006 = 8, 0007 = 2
                    EXTENSION DU RESEAU

Lien-contexte:        0030------)0036
++     RENFORCEMENT: 0030     Fact. trans.: 608, 453     Excitabilite: 1
Liens inter-niveaux: 0006------0036, 0007------0036
--        DESACTIVATION     Niveau=1     0030
Evenement No 3 = 0008 = 8, 0009 = 5, 0010 = 1
))PROPAGATION de 0008 (26:666) Vers 0032 (66/300)
))PROPAGATION de 0009 (16:533) Vers 0032 (119/300)
))PROPAGATION de 0010 (3:100) Vers 0032 (129/300)

Evenement No 4
Lien-contexte:        0036------)0037
++     RENFORCEMENT: 0036     Fact. trans.: 100, 100, 100     Excitabili
Liens inter-niveaux: 0008------0037, 0009------0037, 0010------0037
--        DESACTIVATION     Niveau=1     0031, 0034, 0030

Lien-Reponse:         0037------)0038
++     RENFORCEMENT: 0037     Fact. trans.: 128, 100, 100, 100     Excit
Bouclage:             0038------)0000
++     RENFORCEMENT: 0000     Fact. trans.: 100, 212     Excitabilite: 1
--        DESACTIVATION     Niveau=1     0032, 0026
++     RENFORCEMENT: 0038     Fact. trans.: 100     Excitabilite: 120
--        DESACTIVATION     Niveau=1     0037
--        DESACTIVATION     Niveau=1     0036
++     RENFORCEMENT: 0000     Fact. trans.: 100, 212, 100     Excitabili
))PROPAGATION de 0000 (450:1000) Vers 0030 (608/758)
```

Constitution des niveaux d'intégrations

En limitant le nombre d'événements successifs pouvant être associés lors d'une phase d'apprentissage par différenciation, on interdit par exemple que des structures lexicales intègrent directement de nombreux détecteurs de propriétés du signal d'entrée. Il est nécessaire de faire intervenir des niveaux d'intégrations intermédiaires, l'apprentissage des entités d'un niveau donné s'appuyant sur les structures déjà élaborées aux niveaux inférieurs.

La hiérarchie de traitement qui résulte de ce mode d'apprentissage n'est pas stricte : l'unité-réponse d'une structure d'intégration syllabique intervient par exemple au niveau suivant comme composant d'entités lexicales et aux niveaux supéen tant que mot monosyllabique ; c'est le contexte qui détermine son interprétation.

Renforcement des structures acquises

L'activation de la représentation interne d'un événement est toujours accompagnée de son renforcement, c'est-à-dire de la diminution du seuil des unités qui la composent. Ce processus facilite la propagation d'énergie dans le réseau, qui devient moins dépendante de la réalisation effective d'événements et favorise une prédiction à plus long terme. Restreinte initialement à la pré-activation des descendants immédiats d'une untité active, elle intervient ensuite au niveau des petites-filles et des unités-réponses du niveau inférieur. la diffusion d'énergie se produit alors d'un niveau d'intégration vers les niveaux inférieurs (figure 6).

Cette évolution de la stratégie de reconnaissance, qui accorde progressivement plus d'importance aux informations de niveau supérieur, est un fait remarquable du modèle.

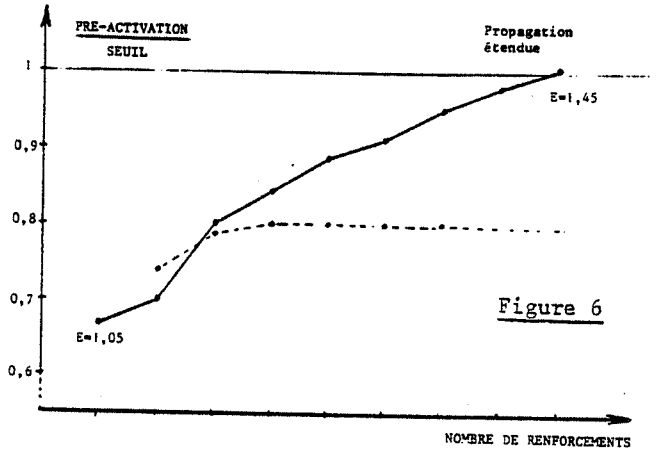


Figure 6

Evolution du degré de pré-activation d'un élément du réseau sous l'effet de renforcements successifs.
Seuil = Activation incidence maximale / Excitabilité E
En pointillé : E = 1,2
En continu : E varie de 5% à chaque renforcement

CONCLUSIONS ET PERSPECTIVES

Le choix d'une stratégie de compréhension est contraint par la qualité des procédures de prétraitement du signal de parole, à l'heure actuelle insuffisante. L'approche descendante semble préférable puisqu'elle permet de pallier l'indéterminisme de l'interprétation des informations acoustiques, à condition que les niveaux supérieurs soient parfaitement définis à l'avance. Cette approche est pourtant incompatible avec une éventuelle capacité d'adaptation des systèmes à des événements imprévus lors du traitement. De récents résultats de psycholinguistique semblent de plus privilégier l'approche ascendante en montrant que les niveaux supérieurs exercent essentiellement leur influence à l'issue de l'accès lexical.

Dans le modèle de mémoire adaptive proposé, l'accès à la représentation topologique des composants du message de parole s'effectue suivant un

mode ascendant, le contexte d'apparition d'un événement déterminant a posteriori son interprétation. Sous l'effet de renforcements successifs des représentations dynamiques mises à contribution au cours du traitement, l'effet prédictif augmente progressivement. La stratégie décrite par ce modèle est principalement une stratégie d'apprentissage, qui détermine l'évolution du réseau au cours du traitement (phases de différenciation ou généralisation, renforcements).

Pour réaliser un système auto-adaptatif de traitement de la parole fondé sur ce modèle, il est nécessaire de mettre en oeuvre un organe d'analyse qui extraie du message vocal des événements élémentaires, points d'entrée d'un premier niveau d'intégration. Contrairement à la procédure classique de segmentation-codage de l'ensemble du spectre, il s'agit pour être compatible avec le schéma d'intégration proposé de sélectionner à différents instants un sous ensemble restreint de canaux fréquentiels dans lesquels se produit une variation significative d'énergie. Le système envisagé repose sur une double analyse simultanée du signal, les événements intégrés étant définis par les réponses d'un banc de filtres à bande étroite, les instants d'intégration de certaines de ces réponses étant décidés au moyen d'un banc de filtres à bande large, à définition temporelle précise. ...Un module d'inhibition latérale pour renforcer le contraste des réponses... Un retard qui varie proportionnellement avec le canal fréquentiel pour étaler au cours du temps des événements simultanés... Voici esquissée une dernière stratégie qui concerne le filtrage des composants les plus élémentaires du message vocal, opération fondamentale dont la qualité détermine celle des représentations internes des formes traitées par un système auto-adaptatif.

BIBLIOGRAPHIE

|1| D.A. Swinney, "Lexical Processing during Sentence Comprehension, Effects of higher Order Constraints and Implications for Representation". The Cognitive Representation of speech, North-Holland, 1981.

|2| K.I. Foster, "Accessing the mental lexicon", New approches to language mechanisms, North-Holland, 1976.

|3| J. Morton, "Interaction of information in Word Recognition", Psychological Review, 1969.

|4| W.D. Marslen-Wilson,"Speech Understanding as a Psychological Process", Spoken Language Generation and Understanding, J.C. Simon (ed.), 1980.

|5| J.J. Mariani, "ESOPE : Un Système de Compréhension de la Parole continue", Thèse de Doctorat d'Etat, Orsay, 1982.

|6| P. Quinton, "Contribution à la reconnaissance de la parole : Utilisation de méthodes heuristiques pour la reconnaissance de phrases", Thèse de Doctorat d'Etat, Rennes, 1980

|7| G. Perennou, M. de Calmes, M. Bui Van, "Décodage Lexical et composante syntactico-sémantique dans ARIAL II" 3ème congrès AFCET-INRIA, RFIA, Nancy, 1981.

|8| J. Mariani, J.S. Liénard, "Reconnaissance Automatique de la Parole utilisant la notion de spectre différentiel", Seminar on Pattern Recognition, Sitel-ULG, Liège, 1977.

|9| J.M. Pierrel, "Contributions à la Compréhension Automatique du discours continu", Thèse de Spécialité, Nancy, 1975.

|10| D. Béroule, "Un modèle de mémoire dynamique, associative, distribuée et adaptative", congrès AFCET-INRIA informatique, Paris, 1985.

MISE EN CORRESPONDANCE TEMPORELLE DE DESCRIPTIONS PHONOLOGIQUE ET PROSODIQUE DE MOTS DANS LE SYSTEME DE RECONNAISSANCE DE LA PAROLE KEAL

R. VIVES

Centre National d'Etudes des Télécommunications
Route de Trégastel - 22301 - Lannion

RESUME

Nous décrivons dans cet article une méthode de mise en correspondance temporelle de descriptions phonologique, syllabique et prosodique de mots avec les sorties de l'analyseur phonémique du système KEAL. Cette méthode s'apparente aux méthodes classiques de mise en correspondance temporelle utilisant des techniques de programmation dynamique. Son originalité vient de l'élargissement des conditions de continuité habituellement utilisées et des pénalisations par défaut permettant de diminuer le nombre de chemins à tester dans les zones où l'on n'obtient pas de bonnes mises en correspondance.

1. INTRODUCTION

Le problème le plus délicat, lié au traitement automatique de la parole, est celui du contrôle des variantes de prononciations possibles des mots, auxquelles se mêlent les variabilités dues aux conditions d'enregistrement.
En reconnaissance automatique de la parole (RAP) on peut distinguer deux moyens pour résoudre ce problème :

a) par une description exhaustive des variantes

b) en décrivant seulement un petit nombre de variantes de référence et en comparant ces formes de référence aux expressions qui ont été prononcées, à l'aide d'un algorithme de mise en correspondance temporelle.

Les nombreuses équipes qui décrivent les mots directement par des paramètres acoustiques tentent de contrôler les variabilités présentes dans la parole à l'aide du second moyen.
Cette approche est habituellement qualifiée de globale par opposition à l'approche analytique qui caractérise les systèmes de RAP mettant en oeuvre des niveaux d'analyse intermédiaires entre l'analyse acoustique et le niveau lexical.

Dans l'approche analytique de la RAP, on trouve trois résolutions sensiblement différentes du problème des variabilités.

La première est appliquée par les équipes qui estiment pouvoir résoudre le problème à un niveau intermédiaire entre l'analyse acoustique et le niveau lexical :
. au niveau lexical, le mot prononcé ne peut être que la suite des unités (phonèmes, pseudo-phonèmes, syllabes) prévues lors de la confection du dictionnaire (voir pour le Français les travaux du CERFIA, de l'ENSERG, du GIA-LUMINY, de TUBACH).

Nous classons dans un second groupe les travaux des équipes pour lesquels une partie des incertitudes du (ou des) niveau(x) intermédiaire(s) est transmise au niveau lexical et qui procèdent à une description explicite des variantes. On trouve dans ce groupe toutes les équipes qui ont choisi l'approche statistique : un mot est décrit par l'ensemble des variantes qui ont été observées lors d'un apprentissage (voir pour le Français les travaux de l'IRISA).

La troisième résolution du problème apparaît dans les travaux pour lesquels la plus grande part des variabilités des mots ou des expressions de référence ne sont pas décrites explicitement, mais sont acceptées par l'algorithme de mise en correspondance temporelle qui est utilisé :
On a une description implicite des variantes mais comme dans le groupe précédent, aucune prise de décision radicale n'est effectuée dans les niveaux intermédiaires précédant le niveau lexical (voir pour le Français les travaux de la CGE, du CRIN, de NANCY 1, du LIMSI, de l'ENSIMAG).

L'approche que nous proposons s'apparente à cette dernière classe.

L'originalité de la méthode, que nous présentons succintement dans cet article, vient de l'élargissement des conditions de continuité habituellement utilisées et des pénalisations par défaut permettant de diminuer le nombre de chemins à tester dans les zones où l'on n'obtient pas de bonnes mises en correspondance. C'est une méthode de mise en correspondance temporelle par ilôts de confiance entre la transcription phonémique

standard (TPS) d'un mot ou d'une expression, et le spectre phonémique (SP) d'un mot ou d'une phrase que l'on a prononcé, et qui a été élaboré par les analyseurs phonémique [1] et prosodique [2] du système de reconnaissance automatique de la parole KEAL développé au CNET.

Nous décrivons dans le chapitre II les grandes lignes de la nouvelle méthode de calcul d'un indice de ressemblance que nous utilisons dans le système KEAL et notamment quelques points qui n'ont pas fait l'objet de publications.

L'indice que nous avons défini tient compte de l'information prosodique et des éléments les plus stables utilisés pour décrire les mots et les expressions recherchés.

Nous verrons au chapitre III, qu'il permet d'élaborer au niveau syntaxique une mesure des suites de mots, cohérente avec celle qu'il induit au niveau lexical.

Nous discutons dans le chapitre IV de la prise en compte des phénomènes phonologiques. Nous donnons en fin d'article les premiers résultats que nous avons obtenus.

II. DESCRIPTION DE LA METHODE DE CALCUL D'UN INDICE DE RESSEMBLANCE DANS LE SYSTEME KEAL -

En reconnaissance de mots isolés, nous choisissons les 1er et dernier segments détectés comme limites du mot prononcé.

En détection de mots dans la parole continue, chaque segment peut être, à priori, considéré comme un début (ou une fin) possible de mot.

Nous pouvons distinguer 3 étapes :

- la construction d'une matrice de coïncidence entre le SP et une TPS
- la recherche de tous les chemins possibles mettant en relation dans cette matrice les 2 formes comparées
- l'évaluation de ces chemins.

II-1 - Matrice de coïncidence :

La figure 1 donne un exemple de matrice de coïncidence : le mot recherché est "quarante" dans l'expression "huit cent quarante cinq".

La case M (i, j) de la matrice contient une valuation comprise entre 0 et 1 de la ressemblance phonétique maximale existant entre le ième élément de la TPS et le jième segment du SP.

La ressemblance phonétique que nous utilisons a déjà été décrite [3].

Figure 1 : Exemple de matrice de coincidence :

TPS \ SP	d	ɥ	d	ɑ̃	a	ɑ	w	ɛ̃	s	ə	ʁ	z
k	1	0	0,75	0	1	0	0,75	0	0,75	0	1	1
a	0	0,55	0	1	0	0,93	0	0,74	0	0,93	0	0
R	0,75	0	0,25	0	0,75	0	1	0	0,25	0	0	0,25
ɑ̃	0	0,74	0	1	0	0,81	0	1	0	1	0	0
t	1	0	1	0	1	0	1	0	1	0	1	0,75

On recherche la TPS du mot "quarante", /kaRɑ̃t/, dans un SP obtenu pour l'expression "huit cent quarante cinq".

II-2 - Recherche des chemins possibles

On appelle chemin une suite d'éléments de la matrice de coïncidence définis par les couples (i_1, j_1), (i_2, j_2), ..., (i_n, j_n) et vérifiant les conditions suivantes :

$$\text{II-2.1)}\quad M(i_p, j_p) > S \quad \forall p \in [1,n]$$

Cette relation définit les îlôts de confiance : on ne veut s'intéresser qu'aux couples (phonème de la TPS, segment du SP) ayant une ressemblance phonétique suffisante.

II-2.2) Fenêtres de recherche pour les début et fin de chemin

Il existe des zones de la matrice de coïncidence qui ne peuvent pas, potentiellement, être à l'origine de cadrages intéressants.

Ces zones varient en fonction du seuil de détectabilité SD qui est associé au mot que l'on recherche. Dans le cas où l'on cherche un mot de N phonèmes entre les segments numéro JDEB et JFIN du SP, les couples (i_1, j_1) et (i_n, j_n) représentant les premier et dernier éléments d'un chemin, devront vérifier les conditions :

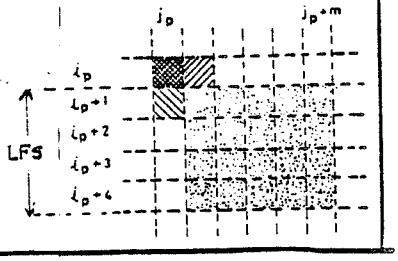
$$i_1 < N(1-SD)+1 \text{ et } j_1 < JFIN+1-N*SD,$$
$$i_n > N*SD \text{ et } j_n > JDEB-1+N*SD.$$

II-2.3) Conditions de continuité

La figure 2 décrit l'ensemble des points (i_{p+1}, j_{p+1}) pouvant être les suivants directs d'un point (i_p, j_p). La largeur LFS de la fenêtre de recherche est paramétrable en nombres entiers de segments.

Figure 2 : Conditions de continuité d'un chemin

Le successeur de (i_p, j_p) peut se trouver dans la case /// en cas de dispersion, dans la case \\\ en cas de fusion ou dans la zone en pointillés.



La nature des déformations entre (i_p, j_p) et un élément de la zone en pointillé ne fait intervenir que des combinaisons d'hypothèses de substitution, d'omission ou d'insertion qui sont définies par l'algorithme ci-dessous :

```
i = i_p + 1
j = j_p + 1
tant que (i ≠ i_{p+1} ou j ≠ j_{p+1})
  si (i = i_{p+1}) alors insertion : on
    considère que les segments j à (j_{p+1}-1)
    sont insérés.
    Stop.
  sinon si (j = j_{p+1}) alors omission : on
    considère que les phonèmes i à (i_{p+1}-1)
    de la TPS ont été omis.
    Stop.
  sinon on considère qu'il y a substitution du
    ième phonème de la TPS avec les phonèmes du
    jème segment du SP
    i = i + 1
    j = j + 1
  fin si
 fin si
 fin tant que
 Stop.
```

C'est parce qu'il nous paraît "hardi" de vouloir faire des hypothèses fines de cadrage dans les zones de la matrice de coïncidence où l'on obtient une mauvaise mise en correspondance, que nous avons choisi de définir par défaut, la nature des déformations entre deux ilôts de confiance.

L'idée d'un cheminement par ilôts de confiance n'est pas nouvelle [3]. Récemment, LEUNG et ZUE ont employé une procédure similaire pour cadrer automatiquement des transcriptions phonétiques sur la parole continue [4].

II-2.4) Vérifications prosodiques avec Fo

Nous avons décrit, dans un précédent article [2] comment nous prenions en compte les marques prosodiques extraites de Fo et définies par J. VAISSIERE.

II-2.5) Support de détection

A chaque chemin $(i_1, j_1), (i_2, j_2), \ldots \ldots, (i_n, j_n)$ ainsi défini nous associons un début fictif (i_D, j_D) et une fin fictive (i_F, j_F) qui vont nous permettre de faire des hypothèses d'insertion d'omission ou de substitution de segments en début ou en fin de détection comme nous le faisons entre deux points quelconques du chemin.

Le couple (j_D+1, j_F-1) des numéros de segments correspondant respectivement au début réel et à la fin réelle de la détection dans le SP sera appelé support de la détection.

La détermination des limites d'un mot ou d'une expression dans une phrase prononcée lorsqu'on a aucune indication a priori sur ces limites s'effectue par cadrage phonétique.

Nous nous servons de (i_1, j_1) et de (i_n, j_n) pour effectuer ce cadrage.

- si $i_1=1$, on sait que le premier phonème de la TPS a été cadré sur le j_1 ième segment du SP : il n'y a pas de cadrage phonétique à faire. Dans ce cas $(i_D, j_D) = (0, j_1-1)$.
- si $i_1 > 1$, on sait que les (i_1-1) premiers phonèmes de la TPS n'ont pas été cadrés. Dans ce cas on procède à un cadrage phonétique qui consiste à dire que l'expression recherchée débute sur le jo ième segment du SP, avec $j_o = j_1-(i_1-1)$.
  Dans ce cas $(i_D, j_D) = (0, j_o-1)$.
  Nous utilisons le même principe pour effectuer le cadrage phonétique de fin d'expression.

II-3 - Evaluation des chemins

La fonction $r(\bar{\Phi}, \Upsilon / \Upsilon)_\sigma$ que nous définissons satisfait au principe d'optimalité énoncé par BELLMAN à propos de la programmation dynamique dans la mesure où tout sous chemin d'un chemin optimal est optimal. Elle permet d'évaluer chaque chemin $\Upsilon = (i_D, j_D), (i_1, j_1), \ldots, (i_n, j_n), (i_F, j_F)$ que l'on peut associer au support de détection $\sigma = (j_D+1, j_F-1)$. Elle exprime une ressemblance entre la TPS $\bar{\Phi}$ et le SP Υ selon le chemin Υ considéré.

Sa valeur maximale est 1. On pose :

$$r(\bar{\Phi}, \Upsilon / \Upsilon)_\sigma = 1 - (P_1+P_2+P_3)$$

avec : P_1 pour comptabiliser les dissemblances des éléments qui ont été mis en correspondance.

P_2 comptabilise les pénalisations liées aux déformations provoquées par les insertions, les omissions, les substitutions, les dispersions d'éléments sur plusieurs segments et les fusions de plusieurs phonèmes sur un seul segment.

P_3 comptabilise les pénalisations de vérifications segmentales et suprasegmentales que l'on effectue en cours de cadrage ou en fin de cadrage d'une TPS sur une partie du SP et qui concernent les éléments les plus stables (traits, phonèmes, syllabes...) d'une description de mot ou d'expression de référence [6].

Les contraintes de place nous obligent à demander au lecteur de se reporter à [5] pour les détails concernant les différentes pénalisations P_1, P_2, P_3.

II-4) Mise en forme des résultats :

La fonction $r(\bar{\Phi}, \Upsilon / \Upsilon)_\sigma$, intervient directement dans la formulation de l'indice de ressemblance $R(\bar{\Phi}, \Upsilon)_\sigma$ que nous proposons. On pose :

$$R(\bar{\Phi}, \Upsilon)_\sigma = 1 - \min_{\Upsilon \in E_\sigma} (P_1+P_2+P_3)$$

avec E_σ : ensemble des chemins Υ ayant pour support σ.

III. COHERENCE DE MESURES ENTRE LES NIVEAUX LEXICAL ET SYNTAXIQUE -

En mode de détection de TPS dans le SP d'un flot continu de parole, le niveau lexical peut être suivi d'un niveau syntaxique dans lequel on évalue la vraisemblance d'une suite de détections (voir la thèse de P. QUINTON, Université de Rennes - 1980).

Nous nous proposons de définir une mesure de cette vraisemblance, cohérente avec l'indice de ressemblance que nous avons défini au niveau lexical :

- la vraisemblance d'une suite de détections devra être égale à la ressemblance qu'aurait obtenue la TPS formée de la suite concaténée des TPS détectées.
- le support de détection de la suite concaténée aura pour origine l'origine du premier support de détection de la suite et pour extrémité, l'extrémité de son dernier support de détection.
- Après le contrôle de l'état d'application des phénomènes phonologiques de liaison, on comptera comme insertion les écarts entre deux détections consécutives, on comptera comme fusion les chevauchements sur 1 segment et comme omissions les chevauchements de plus d'un segment [5].

IV - PRISE EN COMPTE DES PHENOMENES PHONOLOGIQUES -

La génération systématique de toutes les variantes de prononciations pour tous les mots et expressions du lexique est coûteuse en temps de calcul si elle est appliquée au moment où l'on effectue la comparaison avec le SP. Elle est encombrante si elle a été établie par avance et mémorisée.

En reconnaissance automatique de la parole on peut se poser la question de savoir si elle est utile.

ZUE a montré, pour l'anglais, qu'il était, à la limite, peu important pour un système de reconnaissance d'identifier correctement les phonèmes appartenant aux syllabes non accentuées [4].

Il n'y a pas, en français, d'accent distinctif sur les mots comme en anglais mais on observe une certaine similitude entre les deux langues : en français, la variabilité de prononciation se situe surtout au niveau des phonèmes dominés (voir "Speech Recognition : a Tutorial in Computer Speech processing" de J. Vaissière - Advanced course Cambridge Univers. press - edit. Fallside-Woods). En anglais cette variabilité opère essentiellement sur les syllabes non accentuées.

Il était tentant d'opter, comme ZUE, pour une description donnant une grande importance aux phonèmes les plus stables des expressions ou des mots, et de nous intéresser aux variantes de prononciations que dans la mesure où elles pouvaient se montrer indispensables à une bonne reconnaissance. Rappelons aussi qu'avec l'indice de ressemblance que nous utilisons, il n'est pas nécessaire de connaître à l'avance toutes les descriptions des variantes phonologiques des mots ou expressions que nous voulons reconnaître mais seulement d'en connaître une description proche.

Nous avons choisi de faire une distinction nette entre deux modes de description des mots ou expressions du lexique : les Transcriptions phonétiques standards (TPS) et les Transcriptions de complaisance (TC).

Les TPS rendent compte des descriptions phonétiques les plus courantes.

Les altérations qui sont prises en compte dans les TPS sont uniquement provoquées par des phénomènes articulatoires, par des accents, par des liaisons bien connus des phonéticiens.

Les transcriptions de complaisance (TC) permettent de récupérer des insuffisances régulières de l'analyseur phonétique sur des phonèmes difficilement segmentables et identifiables dans un parlé relâché.

La méthode que nous avons présentée n'a pas encore été testée sur des applications mettant en jeu la syntaxe. Les résultats que nous avons obtenus sont partiels : ils ont tous été obtenus en mode mots isolés décrits par des TPS et prononcés par des locuteurs ayant participé à aucun apprentissage. La première série de résultats montrait l'intérêt des procédures de vérification effectuées sur les éléments dominants des mots : 88 % de bonne reconnaissance pour 19 mots (dont les 10 chiffres) prononcés 3 fois par 4 locuteurs (2 hommes et 2 femmes) [6]. La seconde série de résultats concerne la reconnaissance de 58 expressions tirées des pages roses de l'annuaire - on a obtenu 85 % de reconnaissance pour 113 expressions prononcées par 2 locuteurs (hommes) sans procédures de vérification.

REFERENCES

[1] MERCIER G., "Rules and strategies for syllabic segmentation, phoneme identification and tuning in continuous speech recognition" Proceedings of Symposium Towards Robustness in speech recognition, Santa Barbara, California 24 nov. 1983 (à paraître).

[2] VIVES R., LE CORRE C., MERCIER G., VAISSIERE J., "Utilisation pour la reconnaissance de la parole continue de marqueurs prosodiques extraits de la fréquence du fondamental" 8° J.E.P. Aix-en-Provence 1977 - pp. 353-363.

[3] VIVES R., "L'analyse lexicale dans le système KEAL pour la reconnaissance de la parole continue". 7° J.E.P. Nancy 1976. pp. 115-127.

[4] LEUNG H.C. et ZUE V.W., "A Procedure for automatic alignment of phonetic transcriptions with continuous speech", Proceedings of the IEEE International Conf. on ASSP. San Diego 1984, Vol. 1, pp. 2.7.1-2.7.4.

[5] VIVES R., "Mise en correspondance temporelle des descriptions phonologique et prosodique de mots dans le système de reconnaissance de parole KEAL". NT/LAA/TSS/243 - CNET-LANNION - 1985.

[6] VIVES R., Voir p. 117 à 127 des Actes du Séminaire GALF - groupe de la communication parlée - Processus d'encodage et de décodage phonétique - Toulouse sept. 1981.

TRADUCTION PHONETIQUE/GRAPHIQUE ANALYSE SYNTAXIQUE

J-Ph DELPIROUX

CERFIA TATP UA 824 du CNRS
188 route de Narbonne - 31062 TOULOUSE CEDEX

ABSTRACT

As part of the ARIAL II project we currently study a syntactic parsing unit whith intent to check morphological parsing proposals , and to product a spelling typewriting sentence . Using "Prolog" we are in a position to settle directly a natural grammar and resolve ambiguities stemed from previous processing .

INTRODUCTION

La transcription orthographique du signal sonore de la parole préocupe de nombreux laboratoires (LMSI,CRIN,IBM, etc ...). C'est le cas du CERFIA qui dans le cadre du projet ARIAL II s'est fixé cet objectif. Nous discutons ici une approche de l'analyse syntaxique faisant appel aux techniques de programmation en logique et utilisant la base de données lexicales BDLEX-0 du GRECO .

La conversion d'une suite de mots phonétiques en une suite de mots graphiques pose de nombreux problème . En effet dans un système de reconnaissance du type ARIAL II les différents traitements phonologiques et morpho-phonologiques construisent des solutions lexicales dont la forme n'est pas déterministe (7) (8) . Un premier traitement consiste donc à reconnaitre la justesse des suites retransmises . Dans les cas où les suites sont reconnues érronées il faut générer des contraintes à partir des suites partielles bien formées . Ces contraintes seront ensuite destinées aux niveaux inférieurs (découpage prosodique , recherche morpho-phonologique) . Dans les autres cas , l'utilisation des règles de constructions syntaxiques , les traitements linguistiques doivent permettre de résoudre les ambiguités locales qui résident encore dans les suites . Ces ambiguités proviennent tout d'abord de l'absence de marque de certaines flexions au niveau phonétique (genre , nombre) et de la production d'homophonies .

Travaux éffectués au département TATP du CERFIA sous la direction de G Pérennou .

Ainsi nous sauvegarderons les résultats partiels de l'analyse au niveau de certains groupes syntagmatiques et , outre les problèmes de flexions ou d'homophonies , nous traiterons aussi le cas de l'élision .Nous aborderons encore les difficultés posées par l'indétermination d'une forme phonétique aux analyses antécédentes .

ANALYSE MORPHOLOGIQUE

Elle est ici simulée sur le lexique du projet BDLEX-0 du GRECO (base de données relationnelle du français courant) (3) . Il est composé de plus de 6000 racines complétées de leurs différentes désinences. Une recherche directe est effectuée sur un fichier de mots spécifiques ou courants (déterminants,pronoms,conjonctions,etc...). Le traitement d'une forme phonétique construit une "clause donnée" pour chaque solution retrouvée , c'est à dire pour chaque flexion de chaque homophone de la forme. Par exemple , le traitement de la forme : "f+" ("fon") nous donne

```
NOM(f+,fond ,M,S,0,0 ,N).
NOM(f+,fonds,M,P,0,0 ,N).
VCJ(f+,fonds,0,0,1,pi,V.fondre).
VCJ(f+,fonds,0,0,2,pi,V.fondre).
VCJ(f+,fond ,0,0,3,pi,V.fondre).
VCJ(f+,fonds,0,0,2,ip,V.fondre).
VCJ(f+,font ,0,0,6,pi,V.faire ).
```

Nous retrouvons les différentes formes du mot : le nom au singulier et au pluriel , le verbe fondre aux trois premières personnes du singulier du présent de l'indicatif et à la 2° de l'impératif et le verbe faire à la 3° personne du pluriel du présent de l'indicatif ;6 .Pour une autre forme : "l" nous obtenons

```
DET(1,l',i,S,0,v,d.art.def).
PRO(1,l',i,S,0,v,P.per.com).
DET(1,le,M,S,0,c,d.art.def).
PRO(1,le,M,S,0,c,P.per.com).
```

Là nous observons les deux formes syntaxiques (pronom et déterminant) et les deux variations dues au problème de l'élision . Chaque clause comprend donc la forme graphique , la forme phonétique et tous les arguments syntaxiques de la solution .

ECRITURE DES REGLES

L'analyse se base essentiellement sur une grammaire orthographique reflètant les structures naturelles du langage . A terme elle devra être réalisée par des linguistes . Elle est du type hors-contexte et est constituée de deux sortes de règles : les règles terminaleq correspondant aux catégories syntaxiques du mot isolé, et les règles de construction aux dérivations possibles d'une unité syntagmatique. Dans ces règles, un certain nombre de paramètres doivent être pris en compte sous forme de variables de transfert. Par exemple, pour traduire une flexion en genre et en nombre dans un syntagme, nous écrirons :

SN(ac_g_n)---> DET(ac_g_n)+NOM(ac_g_n).

Chacune des règles de grammaire correspond à une clause Prolog. Une unité syntagmatique correspond à un prédicat spécifique. Leurs arguments permettent la prise en compte des variables de transfert et la gestion du positionnement relatif des mots dans la phrase (phonétique et graphique). Voici quelques exemples de règles de construction :

```
proposition--->
     +syn_nom(ac_v_s)
     +syn_ver(ac_v_s,ac_comp,non_sub)
     +syn_nom(ac_g_n).
proposition--->
     +syn_nom(ac_v_s)
     +syn_pro(ac_comp)"pronom complément"
     +syn_ver(ac_v_s,ac_com,non_sub)
     +syn_nom(ac_g_n).
syn_ver(ac_v_s,ac_comp,sub?)--->
     +vcj(ac_v_s,sub?).
syn_ver(ac_v_s,sub?)--->
     +ne1
     +vcj(ac_v_s,sub?)
     +ne2.
```

Quelques règles terminales :

```
vcj(ac_v_s,sub?)--->
     +VCJ(ac_v_s,temps) "analyse morph."
     +sub(temps,sub?).
nom(ac_g_n)--->
     +NOM(g,n)
     +G_N(g,n,ac_g_n).
```

(Les parametres de positionnement ont été volontairement enlevés.)

STRATEGIE

Tout le problème de la stratégie se situe au niveau du choix de l'hypothèse à développer en priorité. La hiérarchisation des règles de grammaire associées aux hypothèses est donc nécessaire. Elle sera effectuée en fonction des formes syntaxiques les plus courantes. Nous pouvons considérer que l'analyse parcourt un graphe de résolution ET/OU où chaque noeud représente tour à tour une unité syntagmatique ou une règle correspondante. La résolution de la suite de règles par l'interpréteur se déroule strictement sur ce graphe. C'est là l'intérêt d'utiliser ce langage de programmation. la stratégie de recherche (directement liée au moteur de résolution) est de type "depth first" (A/O, sans heuristique, en profondeur d'abord) (6) . En cas d'échec la nouvelle hypothèse, le nouveau noeud développable choisi sera issu du dernier choix possible, du noeud le plus proche. On remarque alors qu'une différence minime entre deux règles de construction peut entrainer l'abandon de toute une partie juste de la résolution. L'introduction d'une nouvelle forme de règle par l'intermédiaire d'un prédicat spécial élimine ce problème. Elles sont du type "P--->S + V + (C)" ou "(A)" correspond à la propriété facultative du syntagme A. Deux formes de cette propriété ont été introduites : l'une favorise la recherche du syntagme, l'autre donne la priorité au saut. Le graphe de résolution est alors fortement réduit et l'efficacité sensiblement améliorée.

Dans le but de préserver les résultats partiels d'une analyse, nous sauvegardons toutes les informations retrouvées aprés la résolution de certaines règles (syntagmes nominaux ou verbaux). Ainsi, en cas d'échec nous avons la possibilité de les retransmettre aux niveaux inférieurs (décodage lexical). Celles-ci vont permettre de nouveaux découpages en particulier dans les groupes de mots dont la structure n'a pas été déterminée.

De même l'analyse morphologique peut ne pas retrouver la correspondance d'un segment phonétique. Nous avons donc mis en place la possibilité d'extrapoler un syntagme élémentaire pour éviter l'échec, continuer la résolution et en cas de réussite, de retrouver ce syntagme grâce à des informations syntaxiques supplémentaires transmises aux niveaux inférieurs.

IMPLANTATION

L'implantation a été réalisée sur DPS/8 au CICT à l'aide du système MULTICS. La base de données lexicales BDLEX-0 a été créée en 1983 sous MRDS et le logiciel morphologique écrit en PL/1. L'analyseur lui-même est écrit en Prolog (CNET V2.00). (1)

Seule une grammaire comprenant une faible partie des formes déclaratives du français courant a été introduite. Elle donne cependant d'assez bons résultats.

Voici quelques exemples résolus par ce système :

Entrée phonétique :"l& pOm ke nu AvJ+ mxZé rul"

Sortie graphique :"les pommes que nous avons mangées roulent"

Dans l'introduction de la proposition relative par "que" (pronom relatif complément) une variable d'accord a été instanciée (les pommes --->que) elle se répercutera alors sur le syntagme verbal et entrainera l'accord du participe passé.

E.P. : "nu AvJ+ mxZé l& pOm"

S.G. : "nous avions mangé les pommes"

Ici, lors du traitement du syntagme verbal aucun accord en complément n'a été retrouvé (variable de transfert libre). Il n'y a donc pas d'accord effectué.

E.P. : "il f+ mxZé l peti gArs+"

S.G. : "ils font mangé le petit garçon"

Là, certains paramètres sémantiques ont été pris en compte : les verbes tels que "faire", "vouloir", "pouvoir" admettent être suivis par un infinitif.
E.P. : "il regArd l AvJ+"
S.G. : "il regarde l'avion"
"ils regardent l'avion"
Outre le cas d'homophonie au niveau de la phrase, il y a aussi ici une gestion de l'élision
E.P. : "l& pOm ke ? AvJ+ mœZé rul"
S.G. : "les pommes que nous avions mangées roulent"
E.P. : "il mœZ l& ?"
S.G. : "il mange les (nom pluriel)"
"ils mangent les (nom pluriel)"
Ici, "?" représente une indétermination d'une forme phonétique lors de l'analyse morphologique. Nous sommes en mesure de pallier ce problème en introduisant des règles terminales dont les arguments sont délibérément libres. C'est ainsi que l'on retrouve une partie des arguments syntaxiques du mot. S'ils appartiennent à un ensemble grammatical fermé (pronoms déterminants) alors par inférence sur cet ensemble introduit sous forme de clauses morphologiques, nous les retrouvons directement.

CONCLUSION

Ce système n'est évidemment pas déterministe. La recherche de la première solution est relativement rapide mais la recherche des solutions suivantes est, elle, beaucoup plus lente (parcours de tous les points de choix). Si cette méthode se révèle efficace sur un corpus dont la grammaire réduite ne décrit que des phrases noyaux du langage, elle devient très lourde lorsque ces formes se compliquent. Il est en effet inopportun d'ajouter une règle ad-hoc pour chaque forme de phrase particulière. D'autres formes de recherches existent : par exemple, l'analyse par les relations de dépendance.(2) (9) Ici, ce n'est plus l'application de règles strictes de grammaire, mais une recherche de "relations pertinentes" entre les mots qui devra permettre la résolution des différentes ambiguités. Même si cette forme de traitement ne circonscrit pas autant les structures syntaxiques de la phrase, elle est plus efficace lors de cette résolution sur un corpus plus élargi. Encore faut-il une description morphologique des mots plus conséquente (arguments sémantiques, pragmatiques).
En résumé, nous pouvons noter l'interêt d'étudier la complémentarité de ces diverses formes d'analyses.

BIBLIOGRAPHIE

(1) A.Colmerauer,H.Kanoui,M.Van Caneghem "Prolog,bases théoriques et développements actuels". T.S.I. vol 2 n° 4 1983

(2) M.Dyer;T.CWolf "MacDypar, a demon parser" 2° école d'été sur les applications de l'I.A. Aix en Provence (fascicule 2) 1984

(3) GRECO de la communication parlée Opération : "Une base de données lexicales du français parlé". Rapport Scientifique 1983

(4) G.E.Heidorn, K.Jensen,L.A.Miller, R.J.Byrd,M.S.Chodorow "The EPISTLE text critiquing system" I.B.M. journal vol 21 n° 3 1982

(5) H.Meloni "Traitement des contraintes linguistiques en reconnaissance de la parole" T.S.I. vol 2 n° 5 1983

(6) N.J.Nilsson "Problem solving,methods in A.I." Mac-Graw-Hill 1971

(7) G.Pérennou "ARIAL II, System for speech recognition" IAPR (4 pages) 1980 (Automatic speech analysis and recognition) Reidel Pub. Cy pp 269-275

(8) G.Pérennou "The ARIAL II speech recognition system" J.P.Haton (ed) 1982

(9) B.Vauquois "La traduction automatique à Grenoble" Document de linguistique quantitative, Dunot 1975

UN SYSTEME DE TRANSCRIPTION AUTOMATIQUE DE LA STENOTYPIE.

A.M. Derouault, B. Mérialdo

Centre Scientifique IBM-France,

36 avenue R. Poincarré, 75116 Paris FRANCE.

ABSTRACT

We present a system for automatic transcription of stenotypy into French. Stenotypy is a pseudo-phonetic coding of speech. Our system achieves the phoneme to text decoding part of the speech recognition process. It includes a 140000 entries dictionnary, and a probabilistic language model for French. It has been implemented on a micro-computer IBM PC-XT.
A more sophisticated language model is proposed, which combines the Markovian approach with a probabilistic grammar, and allows to improve transcription quality.

INTRODUCTION

Notre but est d'étudier le problème du passage phonétique-orthographique en vue d'une application future à la dictée automatique.
Comme première étape, nous abordons le problème de la transcription de la sténotypie en Français. La mise au point d'un prototype de transcription permet de développer des outils de décodage linguistique pour un gros dictionnaire, et un discours naturel sans contraintes.
Après avoir rappelé les principes du codage sténotypique, et décrit le dictionnaire de base, nous présentons le modèle de langage markovien choisi, et la stratégie de décodage. L'analyse des erreurs de transcription indique alors les directions de recherche à prendre en vue d'une amélioration. Nous proposons un nouveau modèle qui prend en compte des informations syntaxiques globales au niveau de la phrase.

POSITION DU PROBLEME.

L'ENTREE PSEUDO-PHONETIQUE.

La méthode de sténotypie Grandjean [1] permet de coder manuellement la parole sous une forme pseudo-phonétique, à l'aide d'un clavier spécial de 21 touches. Une frappe consiste à enfoncer simultanément plusieurs touches, pour coder phonétiquement une syllabe, avec des confusions systématiques entre certains phonèmes (par exemple entre voisé et non voisé).
Il n'y a pas de coupure entre les mots, ni en général de virgules. Seules les fins de phrases sont marquées, par la frappe "*" représentant le point.
L'intérêt de la sténotypie est qu'elle permet la saisie en temps réel de la parole (conférences dans des domaines divers).

SOURCES D'AMBIGUITES.

La suite de syllabes obtenue est très ambiguë. Ceci est dû à trois facteurs:

- les homonymies, par exemple: "donné", "donnés", "donner", ...
- les confusions systématiques provenant de la méthode: "tout", "doux" ou "barre", "balle"
- les différentes coupures possibles en mots, par exemple: "au-delà" et "eau de la".

A cause de l'existence en français de nombreux mots courts et de nombreuses homonymies, le nombre total d'orthographes possibles correspondant à une sténotypie donnée est souvent très élevé (10 mots possibles en moyenne à chaque syllabe).

LE DICTIONNAIRE.

Le dictionnaire dont nous disposons comprend 140000 formes fléchies, avec pour chacune les informations utilisées dans la transcription, à savoir: sa sténotypie, ses classes grammaticales, et leurs fréquences ([4]).
Il a été organisé pour faciliter la recherche de tous les mots correspondants à une sténotypie donnée: il est trié par ordre décroissant de sténo, et chaque enregistrement contient tous les mots dont la sténo est incluse dans la sténo en tête de l'enregistrement.

Remarque: le dictionnaire a été récemment augmenté, et contient en fait 250000 formes. Les résultats donnés dans la suite de cet exposé ont été obtenus avec 140000 formes. Cependant, les premières expériences avec 250000 indiquent que

le taux d'erreurs du modèle markovien reste sensiblement le même.

MODELE DE LANGAGE.

Dans une langue naturelle, on peut considérer que l'émission d'un mot dépend de façon probabiliste des mots précédents. Cette dépendance est bien prise en compte par la notion de source de Markov: Nous prenons pour modèle du français un automate fini non-déterministe où chaque transition est affectée d'une probabilité. Ce modèle permet de calculer la probabilité d'une suite de mots. L'orthographe retenue par le système est celle qui a la plus forte probabilité (au sens de notre modèle).
Cette approche présente l'avantage de pouvoir apprendre automatiquement les paramètres du modèle.

DEFINITION DE LA SOURCE DE MARKOV.

Nous considérons comme ensemble d'états l'ensemble des couples de classes grammaticales (c1,c2) et comme alphabet le dictionnaire français. Une transition passe d'un état (c1,c2) à un état (c2,c3) en produisant un mot m3 de classe c3. La probabilité de cette transition est prise égale à la probabilité d'apparition de la classe c3 après les classes (c1,c2), multipliée par la probabilité que le mot m3 soit prononcé, sachant que la classe est p3.
Le premier facteur est obtenu par une interpolation entre les fréquences triclasses et biclasses. Par exemple, la fréquence triclasse est donnée par:

$$f(c3/c1,c2)=nb(c1,c2,c3)/nb(c1,c2) \qquad (1)$$

où nb(c1,c2) et nb(c1,c2,c3) sont les nombres d'occurrences des bi-classes et tri-classes dans le corpus d'apprentissage. Les coefficients de l'interpolation sont optimisés automatiquement [3].
Le deuxième facteur, probabilité conditionnelle d'un mot sachant sa classe, est indiqué dans le dictionnaire pour chaque mot.
92 classes grammaticales ont été choisies. Les fréquences triclasses et biclasses ont été collectées sur un corpus d'un million de mots. Le modèle résultant possède environ 4000 états et 50000 transitions.

Etant donnée une phrase m(1..n), sa probabilité au sens du modèle considéré sera donc

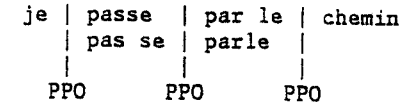
$$p(m(1..n))= \prod p(m(i)/c(i)).p(c(i)/c(i-2),c(i-1))$$

F. Jelinek ([2]) a utilisé en reconnaissance de la parole sur un vocabulaire de 5.000 mots, un modèle de Markov où les états sont les couples de mots (m1,m2), et où les probabilités de transition sont basées sur les fréquences bi- et tri-mots. Un tel modèle nécessite un corpus d'apprentissage important pour avoir des fréquences significatives. Ne voulant pas nous restreindre à un vocabulaire limité, nous avons donc développé un modèle basé sur des classes grammaticales.

DECODAGE.

POINTS DE PASSAGE OBLIGE

A partir de la bande sténotypée, une recherche dans le dictionnaire sténo-français permet d'obtenir tous les mots dont la sténotypie figure sur la bande, ainsi que leurs classes respectives. On obtient donc un graphe dont les noeuds sont les séparations entre syllabes sténotypées et les arcs sont les mots, chaque arc indiquant les syllabes utilisées par le mot associé.
Dans ce graphe, certains noeuds servent pour tous les chemins possibles. On les appelle "points de passage obligé" (PPO). En particulier, ce sont toujours des frontières de mot.



L'intérêt des points de passage obligé est que, à chaque étape de l'énumération, les chemins correspondent à la même portion de la sténotypie, et donc leurs probabilités sont comparables. On évite ainsi les problèmes de normalisation en fonction de la longueur. Expérimentalement, on trouve des PPO tous les 1 à 3 mots (exceptionnellement plus).

DECODAGE

Le décodage s'effectue phrase par phrase. On opère de gauche à droite, de PPO en PPO. Supposons que nous avons deux points de passage obligés, PPO1 et PPO2 tels que:

- la transcription a été fixée jusqu'à PPO1.
- entre PPO1 et PPO2 nous avons p chemins candidats d'au moins deux mots chacun.

On détermine alors le prochain PPO, soit PPO3, tel que tout chemin candidat entre PPO2 et PPO3 contienne au moins 2 mots. On effectue un préfiltre sur tous les chemins entre PPO2 et PPO3 consistant à:

1. garder seulement les chemins dont le nombre de mots est minimum, minimum plus 1, ou minimum plus 2.

2. calculer pour chaque chemin une probabilité d'après un modèle simplement basé sur les biclasses, et garder les dix meilleurs parmi ceux dont la probabilité est supérieure à 10**-3 fois la plus grande.

A la sortie de ce préfiltre, il reste q chemins sélectionnés entre PPO2 et PPO3. Expérimentalement, ce préfiltre n'a jamais éliminé à tort le bon mot.
Pour choisir définitivement la transcription entre PPO1 et PPO2 on calcule les probabilités des pq concaténations d'un chemin entre PPO1 et PPO2 et d'un chemin entre PPO2 ET PPO3. Ces probabilités font appel cette fois au modèle de Markov interpolé avec biclasses et triclasses.

Le chemin entre PPO1 et PPO2 apparaissant dans la meilleure concaténation est alors fixé comme transcription. On recommence la même opération en repartant de PPO2: recherche de PPO4, préfiltre entre PPO3 et PPO4, choix entre PPO2 et PPO3, etc... jusqu'à la fin du texte.

IMPLEMENTATION ET RESULTATS.

Le système de transcription a été développé sur un IBM 3031, puis implémenté sur un ordinateur personnel IBM PC-XT. Il a été testé sur un texte de 1800 mots provenant de journaux télévisés. En sortie, on trouve 8,7% de mots mal transcrits. L'analyse des erreurs montre la répartition suivante:

- 1,7% de noms propres inconnus,
- 1% de frontières de mots incorrectes,
- 2,4% de fautes grammaticales,
- 3,6% d'autres fautes.

Les accords grammaticaux sont impossibles à prédire par un tel modèle lorsque les mots en rapport sont distants dans la phrase, par exemple sujet et verbe séparés par une relative, un complément etc...
Le dernier type d'erreurs relève de la sémantique, ou même de la pragmatique, très difficile à traiter dans un cas de discours oral ouvert. Par exemple, *"qu'en pense ton arôme"* au lieu de *"qu'en pense t'on à Rome"* ...

GRAMMAIRE PROBABILISTE.

Afin de diminuer les 2,4% de fautes classées "grammaticales", dans l'expérience précédente, nous avons développé une grammaire susceptible de suppléer aux insuffisances du modèle local.
Cependant, une grammaire qui accepte le discours oral sans contrainte, avec un vocabulaire illimité, parait une tâche non encore résolue. En fait, pour transcrire la sténotypie, ou plus généralement pour la reconnaissance de parole, on peut cerner les caractéristiques souhaitées pour une éventuelle grammaire:

- elle doit accepter une proportion suffisante de phrases, sans pour autant accepter n'importe quoi. Il faut trouver un compromis sur le nombre de règles que l'on va retenir.
- Elle doit essentiellement discriminer les structures de surface "bonnes" ou "mauvaises".
- Nous n'essaierons pas non plus de produire absolument une et une seule analyse. La définition de probabilités sur les règles permettra de classer les multiples analyses, ou de fabriquer une pseudo-analyse en raccordant les analyses partielles avec une probabilité faible lorsqu'il n'y a pas de noeud couvrant la totalité de la phrase.

DESCRIPTION DE LA GRAMMAIRE.

Les règles sont non contextuelles, écrites dans une forme proche de la forme normale de Backus, avec la possibilité de déclarer des attributs attachés à chaque constituant. Les règles peuvent contenir des conditions sur les attributs, ce qui permet de simuler des règles contextuelles.
Les vrais terminaux de notre grammaire sont les 92 classes grammaticales du dictionnaire.
Chaque mot m du dictionnaire est en fait représenté par des règles de la forme p<-m, où p est une classe grammaticale de m.
Environ 200 règles non terminales constituent la grammaire proprement dite. Notons que la possibilité de tester des conditions logiques sur les attributs permet souvent de condenser plusieurs règles en une seule.
Une évaluation partielle de cette grammaire sur un ensemble de phrases ne comportant pas de fautes de frappe, d'accents, ou de mot inconnu, a montré que le taux de phrases analysées est alors de 65%.

APPRENTISSAGE DES PROBABILITES.

Pour toute suite de mots m(1,2,...,n), on désire définir la probabilité qu'elle soit produite par la grammaire. Dans le cas où il existe une analyse complète de la suite de mots m(1,2...n), soit r(1...n) la suite de règles de dérivations qui la produit depuis le noeud initial. La probabilité de l'arbre syntaxique associé est le produit des probabilités de chaque r(i).

- Si r(i) est une règle terminale, de la forme m<-p, sa probabilité est simplement la fréquence relative du mot m dans la classe p (stockée dans le dictionnaire pour chaque mot).
- Si r(i) est une règle non terminale, de la forme A<-BC, nous prenons pour estimé de sa probabilité sa fréquence relative d'utilisation dans un ensemble assez grand de phrases d'apprentissage.

Remarque:

Jelinek a adapté l'algorithme de Baum pour ajuster les probabilités des règles d'une grammaire non contextuelle probabiliste ([5]). Nous avons testé cet algorithme pour ajuster les probabilités des règles, mais les résultats obtenus avec ces probabilités au décodage sont équivalents à ceux obtenus avec les fréquences.

Dans le cas où la suite de mots ne peut être couverte par le noeud initial, l'analyse fournit des arbres partiels disjoints. Chacun des arbres a une probabilité calculée comme précédemment par le produit des probabilités de dérivation. La probabilité de la suite est alors prise égale au produit de celles des arbres et d'un nombre très petit pour chaque "trou" entre deux arbres.

DECODAGE.

Le décodage se fait phrase par phrase, le principe étant toujours de choisir la transcription de probabilité maximale. Mais cette fois la probabilité est prise égale au produit de la probabilité de la phrase au sens du modèle de Markov, et de la probabilité au sens de la grammaire (à normalisation près).

Seules les phrases composées de groupes de mots sélectionnés par le préfiltre biclasse sont envisagées.

RESULTATS.

Nous avons comparé ce nouveau modèle combiné et le modèle de Markov sur le même texte télévisé, pour une sténotypie sans fautes, les mots inconnus ayant été mis dans un dictionnaire utilisateur. Dans ces conditions, le modèle de Markov fait 7% d'erreurs, et le modèle combiné 5,5%. Signalons que la grammaire seule (choisissant la transcription dont l'arbre syntaxique est le plus probable), produit de moins bons résultats que le modèle markovien.

Voici des exemples qui montrent bien la différence entre les deux modèles. Les mots en gras sont ceux choisis par le modèle combiné lorsqu'ils diffèrent de la transcription avec le modèle markovien seul.

- *"Le patronat à proposer* (a proposé) *la réduction..."*
- *"Mais il faut peut-être se demander si les polonais de Pologne n'avait* (n'avaient) *pas des raisons particulières d'acclamer le cardinal...".*

CONCLUSION.

Le décodage linguistique pour la reconnaissance de parole concerne les trois niveaux linguistiques: lexical, syntaxique, sémantique. La modélisation markovienne peut intervenir aux deux étages syntaxique et sémantique, mais localement. La source que nous avons élaborée pour la sténotypie opère au niveau syntaxique local, tandis que la grammaire opère globalement. Les résultats obtenus avec le modèle combiné montrent la faisabilité, et tout le bénéfice à tirer d'une approche souple et multiple, qui sans opposer ni trancher entre les deux méthodes, profite de leurs avantages complémentaires.

BIBLIOGRAPHIE

[1] Méthode de sténotypie; Ed. Sténotype Grandjean, 15 rue Soufflot, 75240 Paris Cedex 05.

[2] F. Jelinek, R.L. Mercer, L.R. Bahl: Continuous Speech Recognition: statistical methods; C.S.R. group, IBM T.J. Watson Research Center, Yorktown Heigths NY 10598.

[3] A.M. Derouault, B. Mérialdo: "Language modeling at the syntactic level" Proceedings of 7th International Conference on Pattern Recognition, August 84, Montreal.

[4] B. Mérialdo, A.M. Derouault : "Recognition complexity with large vocabulary" Proceedings of International Conference on Acoustics, Speech, and Signal Processing, March 84, San Diego.

[5] T. Fusijaki, B. Greene: "A probabilistic approach for dealing with ambiguous syntactic structure" IBM T.J. Watson Research Center. Research report.

****
* RECONNAISSANCE *
****

RECONNAISSANCE MULTILOCUTEUR DE MOTS ISOLES

A. Mokeddem - H. Hugli - F. Pellandini

Institut de microtechnique de l'université de Neuchâtel
71 rue de la Maladière - 2000 Neuchâtel 7
Suisse

RESUME

Dans une première partie, nous effectuons une analyse factorielle sur les prononciations multilocuteur. La visualisation de la répartition des prononciations d'un ou plusieurs mots prononcés par divers locuteurs dans un espace à deux dimensions montre l'existence de classes de locuteurs.

La deuxième partie concerne la création des références multilocuteur: nous déterminons pour chaque mot un nombre restreint de classes à partir des prononciations de ce mot par un grand nombre de locuteurs; les représentants de chaque classe constitueront les références utilisées pour la reconnaissance multilocuteur. A cet effet, nous proposons deux algorithmes AEC et ASC. La comparaison avec d'autres algorithmes (ISODATA et UWA) sur la base d'un test de reconnaissance montre l'avantage des algorithmes proposés.

Finalement, nous présentons une méthode qui permet de choisir un nombre variable de classes par mot.

1. INTRODUCTION

La reconnaissance multilocuteur de la parole vise à reconnaître la parole telle qu'elle est prononcée par chaque individu d'une large population. Elle doit surmonter les difficultés liées à la grande variation inter-locuteur de la prononciation d'un même mot. Une approche qui a déjà fait ses preuves /2/,/5/ consiste à utiliser, pour reconnaître un même mot, un ensemble de plusieurs références, c'est-à-dire une description multiple du même mot. Lors de la phase d'apprentissage, il s'agit alors de trouver, pour l'ensemble des formes différentes (les diverses prononciations) d'un même mot telles qu'elles sont rencontrées pour une large population, un nombre réduit de références qui les décrivent au mieux. La classification automatique constitue un moyen pour résoudre ce problème.

Avant de chercher des classes au sein d'un ensemble de prononciations, on fait implicitement l'hypothèse que ces classes existent. La question qui se pose est de savoir s'il existe réellement des classes de prononciation au sein d'une population. L'application de l'analyse factorielle permet de répondre à cette question. Avec cette approche, on peut représenter les prononciations d'un ou plusieurs mots par plusieurs locuteurs dans un espace de dimension réduite. Nous pouvons ainsi analyser visuellement la distribution des prononciations.

Ensuite, nous développons et appliquons à la reconnaissance multilocuteur des mots isolés deux algorithmes de classification basés sur une fonction-critère. Il s'agit de l'algorithme d'échange basé sur une fonction-critère (AEC) et de l'algorithme séquentiel basé sur une fonction-critère (ASC). Une comparaison par rapport à d'autres algorithmes (Basic Isodata /4/ et UWA /3/) récemment utilisés en reconnaissance multilocuteur de mots isolés est effectuée.

Il est bien connu que, dans un vocabulaire donné, il existe des mots faciles à reconnaître et d'autres plus difficiles à reconnaître. Nous présentons et testons une méthode permettant l'utilisation d'un nombre variable de références par mot.

2. SYSTEME DE RECONNAISSANCE

Le spectre d'énergie à court-terme du signal vocal est mesuré à l'aide d'un banc de filtres à quatorze canaux, couvrant de façon non uniforme la gamme des fréquences comprise entre 75 Hz et 4.8 kHz.

Après les divers traitements /8/, nous obtenons les caractéristiques de la locution à traiter sous la forme d'une matrice binaire avec 280 bits d'information.

La distance entre une référence et un mot inconnu est déterminée par l'algorithme de programmation dynamique (DTW) /1/. Ici, l'algorithme DTW a été modifié de telle façon qu'il compense les erreurs de détection de début et de fin de mot (claquement des lèvres au début de la prononciation, plosive en fin de mot perdue, etc.).

3. ANALYSE DE LA VARIATION INTER-LOCUTEUR

Soient I éléments obtenus par la prononciation d'un (ou plusieurs) mot(s) par I locuteurs. La distance D (i,j) entre les éléments i et j est déterminée pour tout i et j.

Sous certaines conditions /7/,/9/ on peut représenter les I éléments dans un espace euclidien E^q de dimension q ($q \leqslant I-1$) de telle façon que les distances entre les éléments soient exactement respectées.

Une représentation dans l'espace E^q n'est évidemment ni praticable ni propice à une analyse visuelle. On cherche alors une représentation de tous les éléments (les prononciations) dans un sous-espace F^m dont la dimension m est beaucoup plus petite que q (pour une analyse visuelle m doit être plus petit ou égal à 3). La distance D' (i,j) dans le sous-espace F^m doit être la plus proche possible de D (i,j).

Les différentes étapes pour arriver à une représentation dans F^m sont les suivantes /7/,/9/: à partir de la matrice des distances D(i,j) on détermine la matrice de variance-covariance, puis ses m plus grandes valeurs propres a_m. Les vecteurs propres $\underline{v}_m$ qui leur correspondent définissent les m premiers axes factoriels. L'échantillon i a pour composantes les i-èmes composantes des vecteurs $\sqrt{a_m}\,\underline{v}_m$.

Soit t_m /7/,/9/ le pourcentage de la quantité d'information projetée sur le m-ième axe factoriel.

La figure 1 montre la projection des prononciations du mot "deux" par 25 hommes et 25 femmes sur le plan F^2 (le plan formé par le premier axe factoriel et le deuxième axe factoriel). Les pourcentages d'information conservée valent: t_1 = 29%, t_2 = 11%.

La figure mentionnée ci-dessus montre clairement l'existence d'un groupe masculin et d'un groupe féminin. Ainsi, nous pouvons déjà dire qu'il faut au moins deux groupes pour représenter un mot particulier.

La figure 2 montre la projection des prononciations du mot "cinq" par 50 hommes sur le plan F^2. Les pourcentages d'information conservée valent : t_1 = 17%, t_2 = 10%.

Nous observons la présence de classes: ceci est dû naturellement à l'existence de "types" de prononciations au sein de la population constituée par les cinquantes hommes.

La figure 3 montre la projection du couple de mots "six" et "sept" sur le plan F^2. Chacun des deux mots est prononcé par 25 hommes. Les pourcentages d'information conservée valent: t_1 = 25%, t_2 = 11%.

Cette figure illustre bien la difficulté de la reconnaissance multilocuteur. Elle montre qu'une reconnaissance des mots "six" et "sept" ne pourra se faire que si l'on arrive à regrouper les différentes locutions dans plusieurs classes caractéristiques.

En résumé, la visualisation selon les deux premiers axes factoriels des prononciations de mots isolés par divers locuteurs a mis en évidence l'existence de classes de prononciations pour un même mot. Les classes les plus évidentes sont les "classes masculines" et les "classes féminines"; elles sont dues à des différences anatomiques. D'autre part, nous avons montré qu'il existe aussi des classes de prononciation au sein d'une population masculine. Enfin, nous avons montré (visuellement) que l'utilisation de plusieurs classes par mot pourrait diminuer la confusion entre les mots. Les tests de

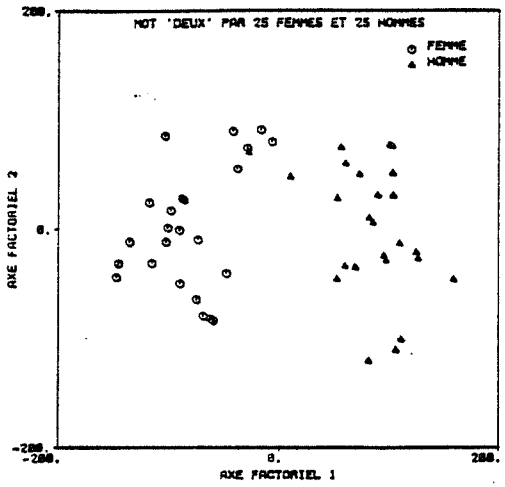


Fig. 1 Projection des prononciations du mot "deux" par 25 locuteurs et 25 locutrices sur le plan déterminé par les deux premiers axes factoriels.

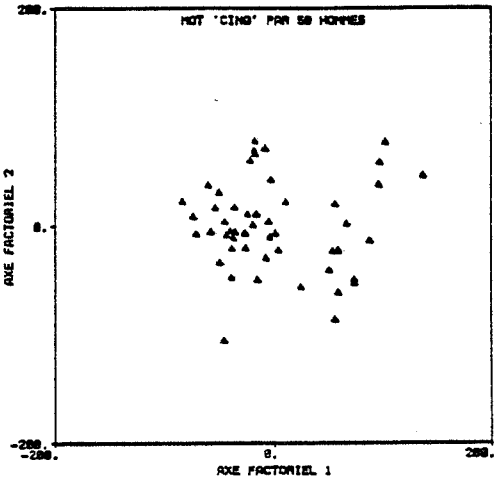


Fig. 2 Projection des prononciations du mot "cinq" par 50 locuteurs masculins sur le plan déterminé par les deux premiers axes factoriels.

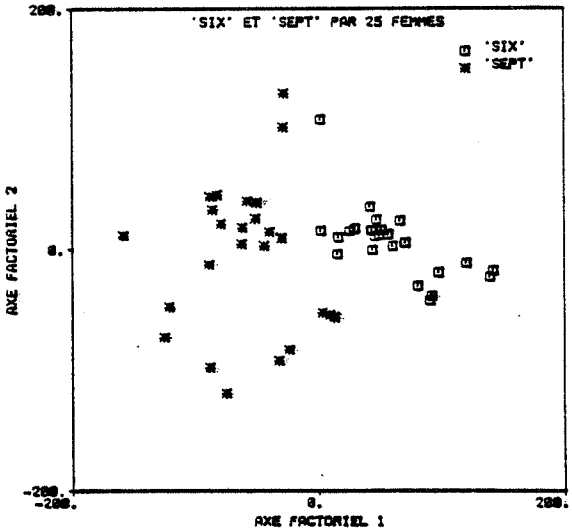


Fig. 3 Projection des prononciations des deux mots "six" et "sept" par 25 locuteurs sur le plan déterminé par les deux premiers axes factoriels.

reconnaissance présentés plus loin montrent, effectivement, que l'utilisation de plusieurs classes par mot permettent de diminuer de façon significative le taux d'erreur.

4. PROCEDURE DE CLASSIFICATION

Soit C l'ensemble des éléments Xi, i=1,...,I (différentes prononciations d'un même mot).

4.1 L'algorithme AEC /6/,/8/

L'algorithme d'échange basé sur une fonction-critère (AEC) produit un nombre m de classes, fixé à priori. En partant d'une partition initiale de m classes, on cherche à minimiser une fonction F en échangeant les éléments d'une classe à l'autre.

4.2 L'algorithme ASC /6/,/8/

L'algorithme séquentiel basé sur une fonction-critère (ASC) crée les classes l'une après l'autre. A chaque itération, on cherche la meilleure classe parmi un ensemble de classes-candidates A(Xi) déterminées à l'aide d'un seuil de distance. La meilleure classe est celle qui minimise une fonction-critère Hq. Les éléments de la classe ainsi déterminée sont ensuite retirés, et l'on recommence la même procédure jusqu'à ce qu'il n'y ait plus d'éléments à classer.

4.3 Fonction-critère

A un élément Xi de la classe Ck, nous associons la métrique suivante :

$$L_q(Xi,Ck) = \left(\frac{1}{n_k - 1} \sum_{Xj \in Ck} d(Xi,Xj)^q \right)^{1/q}$$

A partir de $L_q(Xi,Ck)$, nous définissons les fonctions d'homogénéité suivantes :

$$M_q(Ck) = \min_{Xi \in Ck} L_q(Xi,Ck)$$

Différentes fonctions-critères ont été définies et testées pour les deux algorithmes AEC et ASC. On montre /6/,/8/ que les fonctions-critères suivantes sont bien adaptées :

Fonction-critère pour AEC :

$$F_q = \sum_k M_q(Ck) \cdot (n_k - 1)$$

où n_k est le nombre d'éléments de Ck.

Fonction-critère pour ASC :

$$H_q = M_q(A(Xi))$$

4.4 Représentant d'une classe

Le représentant R(Ck) d'une classe Ck est l'élément qui minimise la métrique L_q.

$$R(Ck) = Xi^* \in Ck \text{ tel que } L_q(Xi^*,Ck) = M_q(Ck)$$

4.5 Nombre variable de références par mot

Lorsque le nombre total de références pour un vocabulaire donné est fixé, il est plus raisonnable à priori d'assigner un plus grand nombre de références pour les mots difficiles à reconnaître, et un plus petit nombre pour les mots faciles à reconnaître.

Le nombre de références c(m) par mot est fixé selon la règle suivante /8/ :

$$c(m) = (Er(m)/Er) \cdot (c-1) + 1$$

où Er(m) est le taux d'erreur de reconnaissance obtenue pour le m-ième mot, Er le taux d'erreur global et c le nombre moyen de références par mot.

5. TESTS DE RECONNAISSANCE MULTILOCUTEUR

Le vocabulaire de test est composé des mots suivants : "zéro, un deux, trois, quatre, cinq, six, sept, huit, neuf, en-avant, en-arrière, terminer".

L'ensemble des éléments utilisés pour la création de la référence multilocuteur pour un mot donné est constitué des prononciations de ce mot par 25 hommes et 25 femmes. Un autre ensemble, constitué par 3 prononciations de chacun des mots du vocabulaire par 5 hommes et 5 femmes, est utilisé pour les tests de reconnaissance. Notons bien que le locuteur-référence est toujours différent du locuteur-test.

La figure 4 donne les taux d'erreur en fonction du nombre de références par mot. Elle compare les performances des différentes méthodes de création de références, à savoir :

1) choix aléatoire des références
2) références sélectionnées par AEC + F_1
3) références sélectionnées par Isodata /4/
4) références sélectionnées par ASC + H_1
5) références sélectionnées par UWA /2/,/3/

Dans le cas des deux algorithmes séquentiels ASC et UWA /2/,/3/, le seuil de distance est choisi, pour chaque mot, de telle façon que l'on ait m classes (m=1,...,6).

Nous pouvons tirer les conclusions suivantes :

- par rapport à un choix aléatoire, la supériorité des performances obtenues avec des références sélectionnées par les algorithmes de classification est évidente;
- nous constatons la décroissance rapide du taux d'erreur en fonction du nombre de références par mot (ou nombre de classes par mot).

En ce qui concerne la comparaison des algorithmes de classification automatique, nous retenons que :

- Parmi les algorithmes d'échange, l'algorithme proposé AEC est plus performant que Basic Isodata /4/.
- Parmi les algorithmes séquentiels, l'algorithme proposé ASC est plus performant que UWA /3/.

La figure 5 donne les résultats de reconnaissance (algorithme ASC) dans les cas suivants :

1) même nombre de références par mot
2) même nombre de références par mot avec élimination au préalable des isolés /8/ (il s'agit des 'mauvaises prononciations' (EI sur fig. 5)
3) nombre variable de références par mot (NVRM sur fig. 5)
4) nombre variable de références par mot avec rejet au préalable des isolés (EI + NVRM sur fig. 5).

Les performances obtenues lorsque le nombre de références par mot est variable sont supérieures au cas où l'on a le même nombre moyen de références par mot.

6. CONCLUSION

Dans ce travail, nous avons montré, au moyen de l'analyse factorielle, que les prononciations d'un même mot par un grand nombre de locuteurs se regroupent dans un nombre limité de classes. Nous avons ensuite présenté et appliqué deux algorithmes de classification basés sur une fonction-critère (AEC et ASC). Les tests de reconnaissance ont montré l'avantage de ces deux derniers par rapport à d'autres algorithmes déjà appliqués en reconnaissance multilocuteur de la parole.

Finalement, nous avons montré l'avantage de l'utilisation d'un nombre variable de références par mot.

Remerciements

Ce travail a été supporté par la 'Commission pour l'Encouragement des Recherches Scientifiques' (CERS no 1158, Berne, Suisse) et les entreprises suivantes : ASULAB SA, CSEM SA, METTLER SA, HASLER SA, AUTOPHON SA and CIR SA.

Références

/1/ H. Sakoe, S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans on ASSP, Vol. ASSP-26 No 1, pp. 43-49, Feb. 1978.

/2/ L.R. Rabiner, "On Creating References for Speaker Independent Recognition of Isolated Words", IEEE Trans on ASSP, Vol. ASSP, No 3, pp. 34-42, Feb. 1978.

/3/ L.R. Rabiner, J.G. Wilpson, "Considerations in Applying Techniques to Speaker-independent Word Recognition", J. Acoust. Soc. Am., Vol. 66, No 3, September 1979.

/4/ Niles, Lee, Harvey F. Silverman., N. Rex Dixon, "A Comparison of Three Feature Vector Clustering Procedures in a Speech Recognition Paradigm". Proc. ICASSP 83, pp. 765-768, 1983.

/5/ S.E. Levinson, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpson, "Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Issolated Word Recognition", IEEE Trans on ASSP, Vol. ASSP-27, No 2, April 1979.

/6/ A. Mokeddem, H. Hugli, F. Pellandini, "Evaluation of criterion based clustering procedures for generating multiple template references in speaker independent speech recognition", 7th ICPR August 1984, Montreal.

/7/ A. Mokeddem, "Analyse factorielle appliquée aux échantillons multilocuteurs de la parole". Rapport interne IMT Neuchâtel.

/8/ A. Mokeddem, H. Hugli, F, Pellandini, "Criterion based clustering techniques applied to SISR", Conf. on "Digital Processing of Signals in Communications", Loughbrough, England, April 1985.

/9/ Diday E., Lemaire J., Pouget J., Testu F., "Eléments d'analyse de données", Dunod, Paris, 1982.

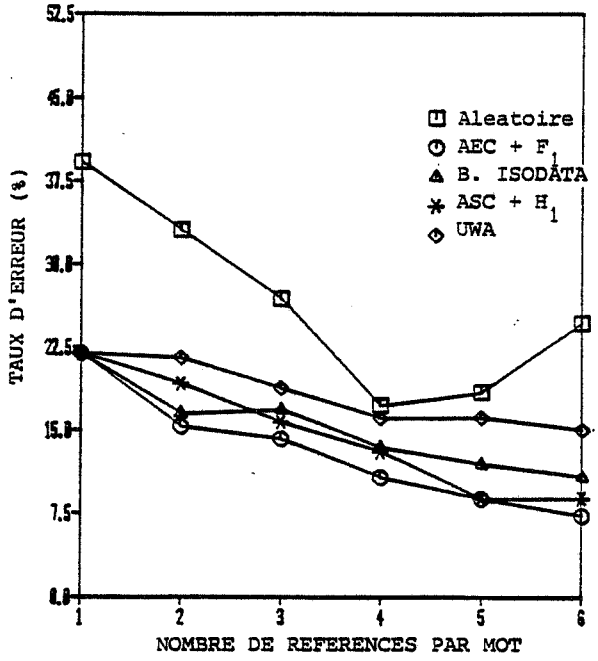


Fig. 4 Comparaison des performances des algorithmes de classification appliqués à la reconnaissance multilocuteur.

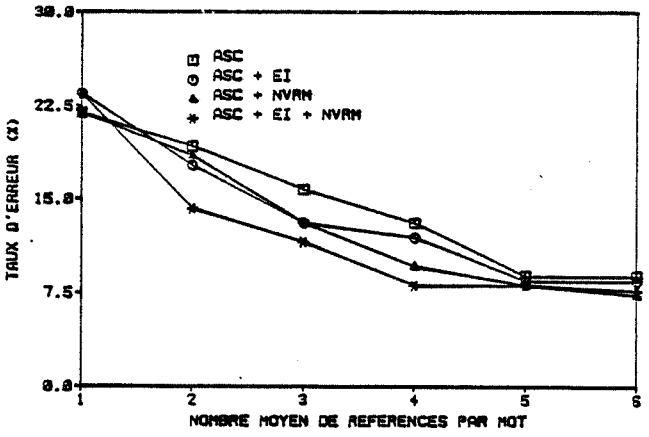


Fig. 5 Effet de l'utilisation d'un nombre variable de classes par mot.

RECONNAISSANCE DE MOTS ISOLES PAR DIPHONEMES ENCHAINES

M. DECKER, J-L. GAUVAIN, J. MARIANI

LIMSI-CNRS BP 30 91406 ORSAY

ABSTRACT

In order to recognize any sequence of speech using a limited number of references, a decision unit that is smaller than a word must be used.
We have chosen to test the diphone as a possible decision unit for speaker dependant recognition of CVCV strings.
Myers and Rabiner's level-building connected word algorithm has been adapted for use with diphones. Tests were carried out on 5 series of 11 words. Without lexical constraints we obtain a phonetic recognition rate of 80% and a 90% rate for word recognition with lexical constraints.
Results show that, for certain phonemes which are very context-dependant, a decision unit larger than the diphone is necessary.

INTRODUCTION

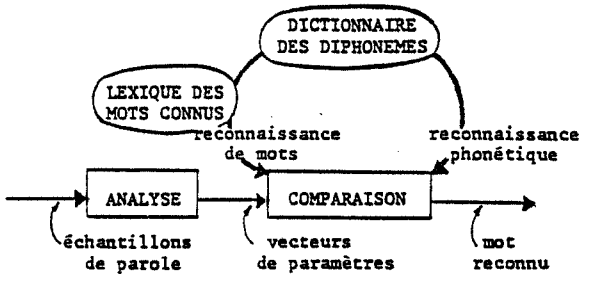
Un problème fondamental en reconnaissance automatique de la parole (RAP) est le choix de l'unité de décision. L'approche phonétique est séduisante en raison du nombre très faible de références à garder en mémoire, mais les résultats obtenus sont décevants jusqu'à présent. En effet plus la référence est courte, plus elle dépend du contexte dont elle est extraite. L'objectif à long terme de ce travail est le choix de l'unité de décision pour la reconnaissance de la parole continue [1],[2],[3],[4]. Dans le travail présenté ici, nous avons choisi d'utiliser comme unité de décision le diphonème. N'importe quel mot (ou phrase) sera donc reconnu comme suite de diphonèmes enchaînés les uns aux autres, ce qui sous-entend une segmentation. Plusieurs approches peuvent être adoptées. L'une consiste à segmenter la phrase en diphonèmes, avant d'effectuer la phase de reconnaissance. Ceci permet de se ramener à une reconnaissance d'unités isolées, mais toute erreur de segmentation est fatale pour la reconnaissance de la phrase. De plus les règles de segmentation peuvent être dépendantes du vocabulaire et la méthode doit donc être repensée lors de chaque modification du dictionnaire. Une autre approche consiste à segmenter la séquence de parole pendant la phase de reconnaissance; ceci nécessite un coût en temps de calcul plus élevé, mais évite les problèmes soulevés précédemment.

Le système présenté ici a les caractéristiques suivantes:

- l'unité de décision est le diphonème.
- il peut reconnaître n'importe quelle suite de parole constituée d'enchaînements consonne-voyelle / voyelle-consonne (CV-VC).
- il est monolocuteur.
- la segmentation est obtenue en fin de reconnaissance. Le module de reconnaissance retourne alors comme solution la suite de diphonèmes qui réalise la correspondance optimale au sens de la comparaison dynamique.

I. SYSTEME DE RECONNAISSANCE DE MOTS ISOLES PAR DIPHONEMES ENCHAINES

Il s'agit d'un système de reconnaissance globale qui peut être schématisé comme suit:



Ce système comporte donc les modules suivants:

I.1 un module d'analyse /paramétrisation

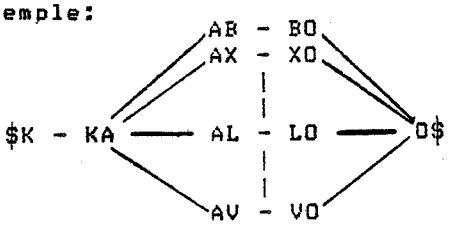
Les étapes successives de l'analyse sont les suivantes:
échantillonage à 10 kHz,
fenêtres de Hamming de 25.6 ms toutes les 12.8ms,
préaccentuation (6 db/octave),
FFT : 256 points
échelle de bark : 16 canaux (Log),
normalisation: passage aux pentes (i.e. différences entre les canaux j et j-1) : 15 paramètres.

I.2 un dictionnaire de références

Les références sont les diphonèmes, où par diphonème on entend la suite de vecteurs allant de la partie stable d'un premier phonème à la partie stable du phonème suivant.
Le dictionnaire a été construit à partir de la prononciation d'une série de logatomes contenant tous les diphonèmes (ex. apataka,ipitiki,.. abadaga....). Ces mots, après analyse, sont visualisés sous forme de sonagramme numérique, ce qui permet d'y localiser les frontières des diphonèmes, le découpage se faisant manuellement. Nous envisageons évidemment d'automatiser cette étape, c'est-à-dire segmenter automatiquement les logatomes en diphonèmes , puis effectuer une phase de tests pour détecter les références mal découpées.
Il faut ajouter que le dictionnaire est construit à partir d'une seule prononciation de chaque logatome. Une mauvaise élocution (i.e. une élocution non stable) d'un de ces mots entraîne forcément une mauvaise reconnaissance pour les références qui en proviennent.
Une grande partie de notre travail a consisté à mettre au point le dictionnaire de références. Les résultats obtenus en reconnaissance montrent clairement deux types d'erreurs:
- confusions systématiques entre phonèmes: il y a probablement un problème de référence,
- confusions, pas trop fréquentes, entre phonèmes acoustiquement voisins: c'est le problème classique de la RAP.

I.3 un lexique de mots connus du système

Ce module est optionnel. Sans lexique de mots on fait une reconnaissance de type phonétique, c'est-à-dire n'importe quelle suite de phonèmes CV-VC est possible. On peut faire la reconnaissance d'une séquence CV-VC quelconque (mot, phrase..), mais les calculs sont très longs et le taux d'erreurs est évidemment plus élevé du fait qu'on teste toutes les combinaisons possibles de toutes les références.
Le dictionnaire des mots fournit des contraintes lexicales qui, pour chaque événement de la phrase test, limite le choix des diphonèmes possibles. Ceci accélère le processus de reconnaissance et limite les possibilités d'erreurs. Sur un vocabulaire dit de paires minimales (p. ex. cabot, cachot...) l'approche par diphonèmes permet de faire de la reconnaissance discriminante. En effet le dictionnaire des mots peut être représenté sous forme de réseau où chaque mot est décrit comme suite de diphonèmes enchaînés.
exemple:



Le mot à reconnaître est comparé aux différentes suites de diphonèmes proposées par le réseau. Les différences entre les notes de reconnaissance ne proviennent que de la partie discriminante entre les mots.

I.4 un module de comparaison

La séquence de parole analysée doit être reconnue comme suite de diphonèmes enchaînés. L'algorithme employé est le "level-building DTW algorithm" de Myers et Rabiner (1981) [5],[6] que nous avons adapté au problème des diphonèmes.
Vue la définition retenue pour les diphonèmes (découpage dans la partie stable des phonèmes) il est clair qu'on ne peut pas former n'importe quelle séquence de diphonèmes: il faut que le phonème final d'un premier diphonème soit identique au phonème initial du diphonème suivant. C'est cette contrainte (appelée contrainte Domino) qu'il faut intégrer dans l'algorithme de Myers et Rabiner.
Ceci entraîne la création de niveaux "syntaxiques", ou plutôt de contraintes syntaxiques chargées d'empêcher la comparaison de deux suites de diphonèmes si celles-ci ne se terminent pas par le même phonème. Il y a donc autant de niveaux syntaxiques que de phonèmes différents (33 dans le cadre de ce travail). Le nombre de diphonèmes enchaînés i.e. la longueur de la phrase n'est pas fixé, mais borné supérieurement.

La solution obtenue par programmation dynamique est la séquence qui réalise la distance cumulée minimale, où le minimum est pris à la fois sur le contenu et la longueur de la phrase. Pour connaître cette séquence il faut faire un retour arrière à travers les différents niveaux (au sens de Myers et Rabiner) ce qui détermine les frontières des diphonèmes.

II RESULTATS/ DISCUSSION

Des tests ont été effectués sur un vocabulaire de 11 mots qui ne diffèrent que par le phonème du milieu:

cabot, cachot, cadeau, cageot, calot, k-mot, canot, capot, carreau, catho, caveau.

Ce vocabulaire a été enregistré 5 fois.

Test 1:

Les tests de reconnaissance ont été effectués sans contraintes lexicales (mais avec contrainte Domino). Chaque mot est comparé à n'importe quelle suite de diphonèmes CV-VC où le nombre maximal de diphonèmes enchaînés est fixé à 6. Sur les 55 mots à reconnaître il n'y a eu qu'une seule erreur sur la longueur de la chaîne: rajout du phonème R à la fin d'un mot. Il faut remarquer que dans le cadre CV-VC il n'y a pas beaucoup de risques d'erreur sur la longueur du mot à part des rajouts en début ou en fin de mot, vue la contrainte d'enchaînement des diphonèmes.
Le taux d'erreur est évalué en fonction des confusions phonétiques. On a obtenu les erreurs suivantes (pour le code phonétique cf [7]) :

PHONEME	K	A	O	N	V	L
NB OCCUR	55	55	55	5	5	5
CONFUSION	P 2	< 25	E 1	Z 1	R 1	J 4
	T 1	(3	+ 2	V 1		
		E 3	= 2			

Ces résultats exigent quelques commentaires:
Pour le A on a obtenu 25 confusions avec < [un].
Ce genre de résultat nous amène à mettre en cause, entre autre, le dictionnaire des références car si une erreur est aussi fréquente cela veut dire que la référence qu'on cherche à reconnaître et/ou celle qu'on a reconnue ne sont pas représentatives de l'événement acoustique correspondant.

Le taux de reconnaissance phonétique peut être évalué à 80%.

Test 2:

En ajoutant les contraintes lexicales le mot reconnu appartient forcément à la liste des mots prononcés. Il n'y a donc plus que les erreurs sur la partie discriminante des mots qui sont possibles.
Les tests ont été faits non seulement avec pour paramètres les pentes entre canaux adjacents (frames de 15 param.) mais aussi avec les différences entre les canaux j et j-2 (frames de 14 valeurs).
Avec les vecteurs de dimension 15 on obtient les résultats suivants:
- 2 confusions KAJO-KAXO,
les phonèmes J et X [chJ ne diffèrent que par le trait voisé/non-voisé, de plus le diphonème JO peut être considéré comme mauvaise référence car le J n'a pas assez d'énergie dans les hautes fréquences.
- 1 confusion KARO-KAVO,
(3 erreurs sur 50: 94%)

La reconnaissance avec les vecteurs de dimension 14 fournit de meilleurs résultats:
- 1 confusion KAJO-KAXO,
(1 erreur sur 50: 98%)

Pour les deux séries de test il y a eu erreur systématique entre KALO et KAJO. Il faudrait faire des tests sur les diphonèmes AL et LO pour décider s'il s'agit de mauvais représentants. Dans ce cas on peut améliorer les résultats en changeant le contenu du dictionnaire. Sinon il faut conclure que l'unité de décision est trop petite pour tenir compte des effets de coarticulation. En tenant compte de ces erreurs dans les pourcentages de reconnaissance on obtient pour le test1 8 erreurs sur 55 soit 85% et pour le test2 6 erreurs sur 55 soit 90%.

CONCLUSION/ PERSPECTIVES

Le diphonème semble être bien adapté à la reconnaissance de suites CVCV. Néanmoins il faudrait faire des tests plus poussés sur les liquides, les nasales et les fricatives voisées. Eventuellement il faudra avoir recours à des références multiples ou bien utiliser des triphonèmes pour mieux tenir compte du contexte. De plus nous devrons compléter le dictionnaire en y ajoutant les diphonèmes de type VV et CC.
L'intérêt de la méthode décrite ci-dessus est qu'elle permet de faire la reconnaissance diphonétique sur une

séquence de parole quelconque. Mais plus la phrase à reconnaître est longue, plus les moyens à mettre en oeuvre deviennent importants en place mémoire et en temps de calcul. Un moyen d'y remédier est de réaliser une segmentation grossière qui consiste à localiser des ilôts de confiance (ici des frontières de diphonèmes). La longueur des segments ainsi découpés permet de fixer le nombre maximal de niveaux permis.
Un problème important reste l'adaptation du système à un nouveau locuteur pour laquelle on envisage la technique de bootstrapping [8].

BIBLIOGRAPHIE

[1] J. MARIANI 'diphone recognition of continuous speech', ICA 81

[2] A.E.ROSENBERG, L.R.RABINER, J.G.WILPON, D.KAHN, 'Demi-syllable based isolated word recognition system', IEEE ICASSP 83.

[3] T.WATANABE, 'Segmentation-free syllable recognition in continuously spoken japanese', IEEE ICASSP 83.

[4] C.SCAGLIOLA, L.MARMI, 'Continuous speech recognition via diphone spotting, a preliminary implementation', IEEE ICASSP 82.

[5] C.MYERS, L.R.RABINER, 'A level-building dynamic time warping algorithm for connected word recognition', IEEE TASSP avril 81.

[6] C.S.MYERS, L.R.RABINER, 'Connected digit recognition using a level-building DTW algorithm', IEEE TASSP juin 81.

[7] B. PROUTS, 'Contributions à la synthèse de la parole à partir du texte', thèse de docteur-ingnieur.

[8] A.M.COLLA, D.SCIARRA, 'Automatic diphone bootstrapping for speaker-adaptive speech recognition', IEEE ICASSP 84.

DEUX UTILISATIONS DU CODAGE VECTORIEL AU TRAITEMENT AUTOMATIQUE DE LA PAROLE

H.BONNEAU, M.ESKENAZI,J.MARIANI

LIMSI/CNRS, BP.30, 91406 ORSAY CEDEX

ABSTRACT
We evaluate Vector Quantization for two applications in automatic speech processing: speaker-dependant isolated word recognition and speaker identification.
For each application, we discuss our choices of an automatic calssification method for the creation of the dictionary. For global recognition, we have chosen the covering method; whereas for speaker identification, clustering was used.
The results for global recognition show that there is no loss in performance and that memory space is reduced by a factor of 3.
First speaker identification results show 98% correct identification for any sentence pronounced by a speaker who is known to the system.

INTRODUCTION

Le codage vectoriel s'est avéré une technique efficace en matière de transmission de la parole. Il nous a donc semblé intéressant d'évaluer son utilité pour deux autres applications au traitement automatique de la parole: reconnaissance globale de mots isolés (dans le but de diminuer l'occupation mémoire) et identification du locuteur (dans l'optique de réaliser un système indépendant de la phrase prononcée). Pour chaque application, différentes méthodes ont été choisies et testées.

I LE CODAGE VECTORIEL

I.1 PRINCIPE DE LA METHODE

Une séquence de parole (séquence d'apprentissage) sert à la constitution du dictionnaire. Ce dictionnaire est construit automatiquement et se compose des formes spectrales de la séquence d'apprentissage qui sont les plus représentatives.

Quand un spectre inconnu de parole arrive, il est quantifié, c'est à dire qu'il est remplacé par le numéro de son plus proche voisin dans le dictionnaire. Les traitements ultérieurs se font alors non plus sur le spectre inconnu mais sur son représentant dans le dictionnaire, lui-même désigné par son numéro d'ordre. Par soucis de clarté nous appellerons CODVEC un vecteur du dictionnaire.

I.2 METHODES DE FORMATION DU DICTIONNAIRE

La qualité d'un quantificateur dépend donc essentiellement de son dictionnaire. Plusieurs méthodes existent dans la littérature faisant appel à des techniques de base de classification automatique. On peut distinguer deux grands types de méthodes: les méthodes dites de nuées dynamiques (Clustering) et celles employant un algorithme incrémental (Covering).

I.2.1 ALGORITHMES DES NUEES DYNAMIQUES

Les algorithmes utilisant des techniques de type nuées dynamiques génèrent un nombre de codvecs fixé.
Pour générer N codvecs la plupart de ces algorithmes demandent le choix de N vecteurs initiaux.
Il s'agit d'un processus itératif dont chaque itération opère en deux étapes:
-La première étape consiste à calculer les distances entre tous les vecteurs de la séquence d'apprentissage et les N codvecs issus de l'itération précédente. On peut alors former la partition engendrée par ce dictionnaire sur la séquence d'apprentissage. Chaque classe est associée à un codvec et constituée de tous les vecteurs de la séquence d'apprentissage dont il est le plus proche voisin.
-Dans l'étape suivante chaque codvec est remplacé par le centre de gravité de la classe qui lui est associée. Une distorsion est alors calculée entre le dictionnaire ainsi désigné et la séquence d'apprentissage.
Le processus cesse d'itérer quand la stabilité désirée est obtenue. Cette stabilité est évaluée à partir de deux distortions successives.

Par son principe, le codage vectoriel répond à ce genre de problème. En effet, les techniques de classification exposées en I.2 permettent de sélectionner un ensemble de spectres (les codvecs) représentatifs de l'élocution d'une personne, à partir d'une séquence de parole suffisamment longue et fournie phonétiquement.

- APPRENTISSAGE

Chaque locuteur prononce un texte relativement représentatif de la langue. On désigne ainsi pour chaque locuteur l'ensemble des codvecs qui le caractérisent: son dictionnaire. On obtient donc une base constituée d'autant de dictionnaires que de locuteurs à identifier.

- IDENTIFICATION

Le locuteur prononce une phrase ou un mot (suite de vecteurs).

Il faut déterminer une distance entre la suite prononcée et chaque dictionnaire de la base. Le locuteur dont le dictionnaire a la plus petite distorsion est celui qui est reconnu. Cette distorsion globale sur la phrase est égale à la somme des distances entre chaque vecteur de la phrase test et son codvec dans le dictionnaire en question

Ainsi, chaque dictionnaire étant constitué à la suite d'un apprentissage sur un texte relativement représentatif de la langue, la méthode choisie permet d'effectuer l'identification du locuteur, non plus sur un vocabulaire figé comme c'est le cas dans la majorité des systèmes existants, mais sur n'importe quelle phrase prononcée.

III CHOIX DES METHODES

III.1 ANALYSE ET PARAMETRISATION DU SIGNAL VOCAL

Nous avons employé des analyses du signal différentes selon la nature du traitement à effectuer.

Etapes successives de l'analyse

Partie commune	Reconnaissance	Identification
- Fenêtre de Hamming	512ms	256ms
- Préaccentuation	256 valeurs	128 valeurs
- Transformation en echelle de BARK	16 valeurs	32 valeurs
- Normalisation	Passage aux pentes: $x_{i+1} - x_i$ ->15 valeurs	Soustraction de la moyenne: $x_i - \bar{x}$ -> 32 valeurs

III.2 CONSTITUTION DU DICTIONNAIRE

-EN RECONNAISSANCE

Parce qu'elle ne nécessite aucun moyennage et qu'elle fournit un dictionnaire non figé, la méthode incrémentale se trouve bien adaptée au problème de la reconnaissance globale.

Elle offre la possibilité de pouvoir compléter le dictionnaire à partir d'une nouvelle séquence de parole, sans abandonner les vecteurs qui constituent déjà le dictionnaire (comme c'est le cas quand on utilise des algorithmes de nuées dynamiques). Cette qualité devra faciliter l'extension du système de reconnaissance au cas multilocuteur. Une première approche sera en effet de compléter le dictionnaire monolocuteur avec d'autres séquences de parole d'un ou de plusieurs autres locuteurs.

- IDENTIFICATION DU LOCUTEUR

Dans ce cas il n'y a à priori pas d'algorithme mieux adapté qu'un autre.

Cependant, nous avons constaté qu'avec la méthode incrémentale le nombre de codvecs désignés (à partir du même texte d'apprentissage) varie énormément d'un locuteur à l'autre (du simple au double). Nous avons donc testé un algorithme de nuées dynamiques (qui permet de fixer à priori le nombre de codvecs) de façon à pouvoir équilibrer les dictionnaires. Cette méthode a d'ailleur permis une amélioration des résultats (cf IV.2)

Nous avons choisi l'algorithme de SPLIT [1].

IV.TESTS ET RESULTATS

Les programmes ont été implantés sur VAX730(VMS) en FORTRAN.

IV.1 EN RECONNAISSANCE

Un programme de reconnaissance sans codage vectoriel, similaire, en ce qui concerne l'analyse spectrale, à celui utilisé pour la méthode avec codage vectoriel, a été mis au point pour fournir une norme de comparaison.

IV.1.1 TESTS

Vocabulaire CHIFFRES

-Un premie test a été effectué sur la reconnaissance des dix chiffres (de zéro à neuf). Les enregistrements utilisés pour les essais proviennent de la base de données du groupe de l'OTAN RSG10 pour l'évaluation des systèmes de mots isolés et enchaînés.

Parmi les enregistrements, ont été choisis ceux de trois locuteurs de langue française (deux hommes JM et JG, et une femme FN) et un locuteur de langue hollandaise (une femme LD), puisqu'il a été déterminé le plus difficile à reconnaitre par la majorité des systèmes testés.

Vocabulaire KABO

-Un second test a été effectué sur un vocabulaire plus difficile, dit de paires minimales, de 25 mots. Ces mots diffèrent:

- soit par leur première consonne:
  ex: Carreau, Garrot,
- soit par leur deuxième consonne:
  ex: Catho, Cadeau,
- soit par leur voyelle accentuée:
  ex: Cachan, Cachou.

Nous avons enregisté un locuteur de langue française (PL) prononçant 12 fois le vocabulaire.

Des variantes de ce type d'algorithme évitent de se fixer a priori un dictionnaire initial.

Les techniques de nuées dynamiques permettent donc de générer un nombre fixé de codvecs. Les dictionnaires vérifient en outre un certain nombre de propriétés mathématiques: distorsion minimale, bonne stabilité, et convergence [1][2]. Cependant il faut souligner qu'ils sont très couteux. De plus, les codvecs sont obtenus par moyennage sur les classes (si on veut assurer leurs propriétés et en particulier leur convergence) et ne sont donc plus des vecteurs réels.

I.2.2 ALGORITHME INCREMENTAL

Contrairement aux algorithmes de nuées dynamiques l'algorithme incrémental [3] ne génère pas un nombre donné de codvecs mais garantit, entre chaque couple de codvecs du dictionnaire, une distance supérieure à un seuil préalablement spécifié.

Le premier vecteur de la séquence d'apprentissage est pris comme premier codvec. Chaque vecteur de la séquence d'apprentissage subit alors le traitement suivant:

- Recherche de son plus proche voisin parmi tous les codvecs déjà désignés.
- Dans le cas où la distance avec son plus proche voisin est inférieure au seuil choisi, il est considéré comme non représentatif et donc rejeté.
- Sinon le vecteur est rajouté dans le dictionnaire.

Cet algorithme est très peu coûteux. Il présente de plus l'avantage de fournir un dictionnaire non figé: on peut facilement le compléter à partir d'une nouvelle séquence de parole.

Il faut cependant noter qu'il n'assure pas toutes les propriétés des dictionnaires issus des algorithmes de nuées dynamiques, notamment en ce qui concerne la distorsion.

II APPLICATIONS ETUDIEES

Nous avons choisi d'étudier deux types d'applications du codage vectoriel au traitement automatique de la parole: la reconnaissance globale de mots isolés monolocuteur et l'identification du locuteur.

II.1 RECONNAISSANCE DE MOTS ISOLES

Dans le cadre de la reconnaissance globale il nous a semblé intéressant d'évaluer l'utilité du codage vectoriel pour la réduction de l'espace mémoire en fonction du maintien des taux de reconnaissance.

- APPRENTISSAGE

L'apprentissage se fait en deux étapes, le locuteur ayant prononcé deux fois chaque mot du vocabulaire.

- Formation du dictionnaire des codvecs:

Celui-ci est obtenu à partir de la prononciation de tous les mots du vocabulaire par le locuteur. Afin d'éviter de recalculer plusieurs fois les mêmes distances lors de la phase de reconnaissance, on stocke, pendant l'apprentissage, les distances entre tous les codvecs dans un fichier.

- Formation du dictionnaire des références:

Il s'agit de constituer le dictionnaire qui contient pour chaque mot du vocabulaire, la liste des numéros des codvecs le représentant. Dans le cas d'un vocabulaire peu important et de la reconnaissance monolocuteur, ceci peut fort bien être réalisé lors de la première étape d'apprentissage. Nous avons choisi d'adopter tout de suite la méthode la plus générale. En effet, si on admet que le dictionnaire des vecteurs est suffisamment large, on peut alors envisager d'étendre le vocabulaire sans avoir à modifier le dictionnaire des vecteurs.

- RECONNAISSANCE

La reconnaissance d'un mot inconnu s'effectue en deux étapes:

- Quantification:

Pour chaque vecteur du mot inconnu, le plus proche voisin est déterminé par une recherche exhaustive, afin d'assurer le maximum de précision.

- Désignation du mot reconnu:

Le mot inconnu est donc représenté par la liste des numéros des vecteurs le constituant, de même pour les mots de référence. Les distances entre tous les codvecs ayant été calculées, il est alors possible, par un algorithme de programmation dynamique, et en utilisant l'équation symétrique de SAKOE et CHIBA, de déterminer rapidement la distance entre le mot inconnu et chaque mot de référence.

C'est le mot le plus proche du mot inconnu au sens de cette distance qui sera désigné comme le mot reconnu.

Le système peut être schématisé comme suit:

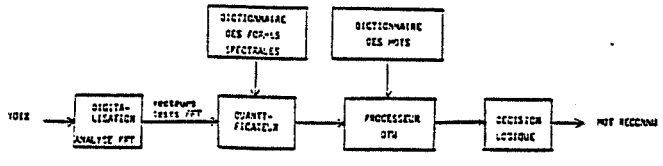


FIGURE 1

II.2 IDENTIFICATION DU LOCUTEUR

Un système d'identification du locuteur est basé sur le fait qu'à la suite d'un apprentissage donné, on dispose d'un nombre suffisant de traits pour caractériser l'élocution d'un individu.

V.1.2 RESULTATS

Pour l'apprentissage:

- Le gain de place mémoire obtenu par le codage vectoriel par rapport aux méthodes classiques est un résultat important.

Soit N le nombre de mots du vocabulaire,

L le nombre moyen de vecteurs par mot,

Q le nombre de vecteurs du dictionnaire,

D la dimension des vecteurs,

Sans codage vectoriel la place mémoire occupée est représentée par:

$$Mo = NxLxD \qquad (1)$$

Avec codage vectoriel:

$$Mo = QxD + NxL \qquad (2)$$

La figure 2 montre que la place mémoire nécessitée par le vocabulaire CHIFFRES est nettement inférieure dans le cas du codage vectoriel pour chaque locuteur testé.

Pour la reconnaissance:

Les taux de reconnaissance pour les deux vocabulaires résumés sur la figure 3 sont très encourageants, montrant que l'emploi du codage vectoriel n'entraîne pas de dégradation des performances.

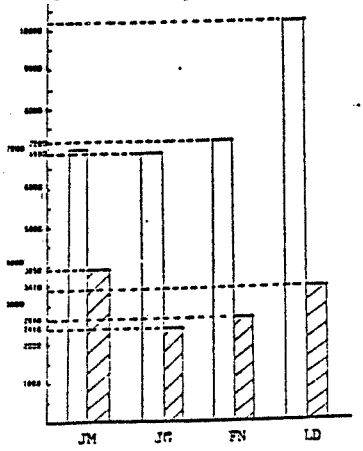


FIGURE 2

Comparaison de l'occupation mémoire pour chaque locuteur sans et avec CV

VOCABULAIRE	LOCUTEUR	NOMBRE DE MOTS TESTES	% DE RECONNAISSANCE
CHIFFRES	JM	100	100 %
	JG	50	100 %
	FN	50	100 %
	LD	100	95 %
KAIRO	PL	250	94.5%

FIGURE 3

TAUX DE RECONNAISSANCE

IV.2 IDENTIFICATION DU LOCUTEUR

IV.2.1 TESTS

Le texte choisi pour la constitution du dictionnaire caractéristique de chaque locuteur est 'LA BISE ET LE SOLEIL'.

Deux algorithmes ont été testés: l'algorithme incrémental et un algorithme de nuées dynamiques. Pour ce dernier, nous nous sommes fixé 64 vecteurs par dictionnaire.

Les tests d'identification ont été effectués sur cinq phrases différentes, prononcées par chaque locuteur, indépendantes du texte d'apprentissage.

IV.2.2 RESULTATS

Dix locuteurs (5H, 5F) constituent la base sur laquelle les systèmes ont été testés.

Pour chaque algorithme nous avons donc effectué 50 tests (10 locuteurs et 5 phrases par locuteur).

- Méthode incrémentale:

Comme il a été dit précédemment (cf III.2) , le nombre de vecteurs obtenus (pour un même seuil) varie énormément selon le locuteur. Les cas extrêmes sont obtenus pour un locuteur féminin (EM) et pour un locuteur masculin (GA):

- locuteur EM: 336 vecteurs
- locuteur GA: 127 vecteurs,

En ce qui concerne les taux d'identification, le résultat est de 85 % d'identification correcte. Les erreurs sont toujours des confusions vers un locuteur dont le dictionnaire contient plus de vecteurs.

- Algorithme des nuées dynamiques.

Les résultats sont très encourageants: sur les 50 tests effectués l'identification est de 98 %.

Le fait d'avoir des dictionnaires équilibrés a donc apparemment amélioré les résultats. Il faut cependant noter qu'une telle méthode nécessite un apprentissage très lourd: sur "LA BISE ET LE SOLEIL" il coûte environ 2h30 de CPU par locuteur.

CONCLUSION

Nous avons évalué un système de reconnaissance de mots isolés monolocuteur. Cette première étude a mis en évidence l'intérêt du codage vectoriel: il permet en effet de diminuer nettement la place mémoire, sans dégrader la qualité de la reconnaissance et ceci même pour des vocabulaires relativement difficiles.

La souplesse du système permet d'autre part d'envisager certaines extensions: reconnaissance de gros vocabulaires et reconnaissance multilocuteur.

En ce qui concerne l'identification du locuteur, les résultats obtenus laissent penser que le codage vectoriel permettra bientôt de réaliser des systèmes de vérification du locuteur indépendants de la phrase prononcée.

BIBLIOGRAPHIE

[1]Y.LINDE,A.BUZO,R.M.GRAY:
"An algorithm for Vector Quantizer Design", IEEE Trans.on comm.signal processing, janvier 80.

[2]A.BUZO,A.H.GRAY,R.M.GRAY,J.D.MARKEL:
"Speech coding based upon Vector Quantization",IEEE Trans.on communication signal processing, Août 80.

[3]D.P.LANDELL,J.A.NAYLOR,R.E.WOHLFORD:
"Effect of Vector Quantization on a Speech Recognition System", IEEE, ICASSP 84.

H.BONNEAU,M.ESKENAZI,J.MARIANI:
"Utilisation du Codage Vectoriel en Reconnaissance de Mots Isolés", Séminaire GALF, 15 février 85.

RECONNAISSANCE GLOBALE DISCRIMINANTE

O.DIOURI, JL.GAUVAIN, J.MARIANI

L.I.M.S.I. - CNRS, B.P.30 91406 ORSAY Cedex

Résumé

Les techniques de reconnaissance globale basées sur la comparaison dynamique et utilisées pour les mots isolés et les mots enchainés, donnent des résultats médiocres sur des vocabulaires contenant des mots acoustiquement voisins et difficiles à distinguer comme *carreau-garrot* et *stalactite-stalagmite*. La principale source d'erreur vient du fait que, toutes les différences entre deux mots voisins sont traitées de façon équiprobable, à cause de la mauvaise représentation des références au niveau de l'apprentissage.

Deux approches complémentaires sont proposées. La première consiste à extraire du corpus d'apprentissage le "bon" ensemble de références, les résultats obtenus montrent l'effet du critère de sélection sur les performances, la seconde approche consiste à construire un réseau discriminant pour les mots voisins, qui permet de focaliser le processus de comparaison sur les régions les plus significatives. Une amélioration est observée pour certains vocabulaires; les résultats sont discutés pour les deux approches.

I. Introduction

Les systèmes de reconnaissance globale basés sur la comparaison dynamique produisent des confusions sur des vocabulaires contenant des mots acoustiquement voisins comme *carreau-garrot* et *stalactite-stalagmite*.
Il existe plusieurs raisons pour lesquelles on aboutit à de tels résultats. La principale source d'erreur vient du fait que, toutes les différences entre deux mots voisins sont traitées de façon équiprobable, bien que les différences des régions phonétiquement identiques soient moins significatives que celle des régions phonétiquement différentes. Un autre problème essentiel, est la mauvaise représentation des mots références au niveau de l'apprentissage. Les systèmes de reconnaissance ont besoin d'un traitement préalable sur le corpus d'apprentissage afin d'en extraire les références plus représentatives. Pour les mots se différenciant par l'initiale ou la finale un autre problème peut intervenir, celui de détection des frontières et d'ajustement temporel. En effet, des améliorations peuvent être apportées en ajustant certains paramètres, dans les algorithmes de détection des frontières et d'ajustement temporel, selon le locuteur et le vocabulaire étudié. Nous présentons dans cet article trois expériences décrivant différentes techniques d'apprentissage qui améliorent les performances de reconnaissance.

II. Conditions expérimentales

Pour les différentes expériences réalisées au cours de cette étude, l'analyse acoustique est réalisée par un banc de filtres couvrant la bande de fréquence 100-5000Hz et dont les fréquences centrales sont réparties selon une échelle de Bark. Pour les paragraphes III et V, nous avons utilisé un banc de 16 canaux, alors que pour le paragraphe IV, nous n'avons utilisé que 8 canaux. L'algorithme de comparaison dynamique est utilisé pour calculer les distances entre les mots et les ajuster temporellement [1]. Soient $R_1 = \{R_1(i); 1 \leq i \leq I\}$ et $R_2 = \{R_2(j); 1 \leq j \leq J\}$, deux références de l'ensemble d'apprentissage, où $R_1(i)$ et $R_2(j)$ sont les spectres caractérisant les signaux aux instants i et j respectivement. La distance cumulée entre R_1 et R_2 est déterminée par programmation dynamique en utilisant l'équation récursive symétrique suivante:

$$g(i,j) = \min \left\{ \begin{array}{l} g(i-2,j-1) + 2d(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-1,j-2) + 2d(i,j-1) + d(i,j) \end{array} \right\} \quad (1)$$

où d est une métrique sur l'ensemble de tous les vecteurs $R_1(i)$ et $R_2(j)$. La distance cumulée $g(I,J)$ entre les deux références est la somme des distances locales $d(i,j)$ pondérées le long du chemin optimal, et la distance entre les deux références est finalement:

$$D(R_1,R_2) = \frac{g(I,J)}{I+J} \quad (2)$$

III. Moyenne et techniques d'apprentissage: expérience 1

Le processus d'élaboration des références à partir de différentes prononciations des mots du vocabulaire, a une grande importance sur la qualité de reconnaissance[2]. Etant donné un vocabulaire de w mots avec P prononciations pour chaque mot, il existe P^w combinaisons possibles pour extraire un ensemble de références.

• Monoréférence aléatoire: L'apprentissage consiste à prendre, de façon aléatoire, un énoncé de chaque mot du vocabulaire pour construire l'ensemble de références.
• Multiréférence: On peut envisager de garder l'ensemble des données comme références, mais cette technique s'avère très coûteuse en espace mémoire. En effet la variabilité inhérente intra-locuteur peut nécessiter un grand nombre de références pour chaque mot.
• Centroïde [3] : si on considère la classe des différents énoncés d'un même mot comme un nuage de points dans un espace, le critère de sélection consiste à choisir la référence située au centre du nuage. Soit $C = (r_1, r_2, ..., r_p)$ la classe de différents énoncés du même mot r du corpus d'apprentissage, on note par $DIM(r)$ la distance intra-classe moyenne entre toutes les autres prononciations appartenant à C. La référence r_m représentante de la classe C de r est choisie de façon que:

$$m = \underset{1 \le i \le p}{\operatorname{argmin}} [DIM(r_i)] \tag{3}$$

où $\operatorname{argmin}_x f(x)$ est la valeur de x qui minimise $f(x)$ et

$$DIM(r_i) = \frac{1}{p} \sum_{\substack{j=1 \\ j \neq i}}^{p} D(r_i, r_j) \tag{4}$$

• Moyenne 1: Le premier algorithme consiste à choisir comme référence la moyenne des différentes prononciations du mot par rapport à celle qui a la longueur la plus proche de la longueur moyenne. Afin de moyenner ensemble, à l'intérieur d'une classe, des prononciations de différentes longueurs, la longueur moyenne intra-classe est calculée. La référence dont la longueur est la plus proche de la moyenne est identifiée. Ainsi toutes les autres répétitions sont comparées avec celle-ci, et une moyenne est obtenue pour chaque spectre du mot pour former la référence représentante de la classe. Cette technique est intéressante parce qu'on représente plusieurs énoncés d'un même mot par une seule référence, la reconnaissance bénéficie ainsi des avantages de la technique multi-référence tout en réduisant la taille mémoire nécéssaire à l'ensemble de références. Cependant, lorsqu'il s'agit de moyenner ensemble les énoncés d'un même mot articulés différemment, cette technique donne des résultats médiocres.
• Moyenne 2: Le second algorithme que nous avons étudié, est analogue à celui de Rabiner [4] basé sur un apprentissage robuste. Il consiste à construire la référence moyenne des deux énoncés les plus proches parmi les différentes prononciations d'un même mot dans une classe. Pour chaque mot r_i d'une classe C, on définit $DI(r_i, r_j)$ pour $1 \le i, j \le p$ et $i \neq j$, la distance intra-classe entre r_i et un autre énoncé r_j appartenant à C. Notre critère consiste à choisir les deux meilleures références r_{m_1} et r_{m_2} de la classe C de façon que :

$$(m_1, m_2) = \underset{\substack{1 \le i \le p \\ 1 \le j \le p \\ i \neq j}}{\operatorname{argmin}} DI(r_i, r_j) \tag{5}$$

La référence représentante de la classe C est obtenue en moyennant ensemble ces deux références le long du chemin optimal obtenu par comparaison dynamique.
Toutes les techniques développées ci-dessus sont évaluées dans un contexte monolocuteur en mots isolés, et sur un vocabulaire V_1 que nous qualifions "difficile", contenant quatre sous-ensembles de mots voisins:

carreau - garrot - calot
poids - toit
bois - doigt
laine - lemme
sectateur - spectateur

Six locuteurs dont deux féminins et quatre masculins ont participé aux tests. Quinze séries d'enregistrements ont été réalisées pour chaque locuteur, les cinq premières séries sont utilisées pour l'apprentissage et les dix autres pour les tests de reconnaissance, soit 660 tests.

méthode classique	multi-références	centroïde	moyenne 1	moyenne 2
83.5	93.5	89	89.5	90.5

Tableau 1: Taux de reconnaissance en fonction du type d'apprentissage.

Le tableau 1 montre les différents résultats obtenus pour chaque technique présentée plus haut. Par rapport à une méthode classique basée sur un apprentissage aléatoire mono-référence, les techniques de la moyenne et du centroïde donnent des performances nettement meilleures. Parmi celle-ci la seconde classe de moyenne est choisie comme meilleur critère de sélection d'ensemble de références pour notre vocabulaire. Il est clair que la qualité de reconnaissance dépend énormément du critère de sélection des références à partir du corpus d'apprentissage.

IV. Réseau discriminant automatique: expérience 2

L'algorithme c_1 présenté dans cette expérience focalise la reconnaissance sur les régions différentes de deux mots et donc particulièrement adapté à des vocabulaires de mots acoustiquement voisins. Rabiner et Wilpon[5] ont proposé une reconnaissance en deux passes, et subdivisent le vocabulaire en sous-ensembles de mots voisins. Brièvement, la première passe consiste à classer le mot à identifier dans l'un des sous-ensembles par comparaison dynamique. A l'intérieur du sous-ensemble identifié, la seconde passe utilise des coefficients de pondération calculés en phase d'apprentissage pour chaque paire de références pour priviligier les régions discriminantes des deux mots. La technique que nous proposons ici, s'inspire de celle de Moore [6][7][8]. C'est une solution à une seule passe, qui consiste à construire un réseau discriminant à

partir des références issues de la phase d'apprentissage. Le réseau est construit de telle façon que les régions ne contribuant pas à la différence soient intégrées sur des chemins identiques, comme le montre la figure 1. Lors de la phase de reconnaissance, la dissemblance est évaluée pour chaque chemin du réseau et les différences ne dépendent que des sections du réseau constituées de plusieurs chemins. Le processus de comparaison est ainsi focalisé sur les régions discriminantes. Si deux mots différents mais acoustiquement voisins sont comparés, les distances locales des régions semblables sont très petites, et celles des régions différentes sont grandes. Autrement dit, les distances locales peuvent être utilisées pour guider la formation du réseau.

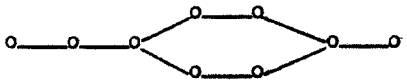


Figure 1: réseau discriminant (classe c_1)

L'algorithme c_1 repose sur deux seuils:
- Un seuil local d'intégration ρ qui est déterminé d'une façon automatique en fonction de la distance entre les deux références. Les distances locales $d(i,j)$, le long du chemin optimal sont comparées à ce seuil. Au dessous du seuil, les deux zones comparées sont moyennées et représentées sur des chemins identiques du réseau, et au dessus du seuil elles sont représentées sur des chemins différents.
- Un seuil global qui est déterminé expérimentalement, en fonction des distances intra-classe pour chaque sous-ensemble du vocabulaire, si la distance entre deux mots est au dessus du seuil, ces derniers ne subissent pas de traitement discriminant.

Des tests ont été réalisés en mode monolocuteur (pour deux locuteurs) et en mots isolés sur un vocabulaire V_2 contenant 13 mots voisins distincts par un phonème au centre du mot:

cabot, cachot, cadeau, cageot, cahot, caillot, calot, kmot, canot, capot, carreau, catho, caveau.

Chacun des deux locuteurs a prononcé 10 fois ce vocabulaire dont une série constitue l'ensemble de références et les neufs autres sont utilisées pour les tests de reconnaissance, soit 1170 tests pour chaque locuteur.

	méthode classique	réseau discriminant
loc1	88	97
loc2	87	88

Tableau 2

Le tableau 2 montre des résultats concluants surtout pour le locuteur 1. Pour le locuteur 2 qui ne s'exprime pas dans sa langue maternelle, l'amélioration est réduite.

V. Réseau discriminant: expérience 3

Un deuxième algorithme c_2 est présenté pour construire le réseau discriminant. Contrairement à c_1, l'intégration locale se fait tout en conservant, séparemment les deux spectres des deux mots au même niveau du réseau, de telle façon qu'en phase de reconnaissance on puisse choisir celui qui est plus proche du spectre à identifier (figure 2).

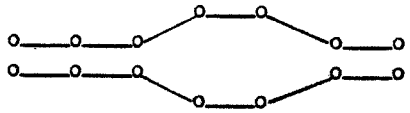


Figure 2: réseau discriminant (classe c_2)

Les deux algorithmes c_1 et c_2 ont été testés sur le vocabulaire V_1 avec les mêmes locuteurs que dans l'expérience 1. Les résultats montrent globalement une petite différence(1%) sur le taux de reconnaissance entre les deux algorithmes, en faveur de c_1. Cependant cette différence varie en fonction du locuteur et du sous-ensemble étudié. Pour la suite de l'expérience, nous avons opté pour l'algorithme c_1.
La seconde étape de cette expérience consiste à appliquer le réseau discriminant à des sous-ensembles de références obtenues par l'une des techniques d'apprentissage, décrites dans l'expérience 1. Nous avons utilisé la moyenne 2 avec laquelle, nous avons obtenu le meilleur taux de reconnaissance. Le tableau 3 montre les résultats obtenus pour six locuteurs sur un corpus de 660 tests dans chaque sous-ensemble.

	carreau garrot calot	poids toit	bois doigt	lemme laine	sectateur spectateur	taux global
méthode classique	93.0	76.0	89.5	89.8	66.0	83.5
c1	93.5	78.0	86.5	91.0	69.0	84.0
c2	93.5	75.8	86.6	89.1	64.1	83.0
moyenne 2 + c1	94.0	93.0	93.3	92.0	76.5	91.0

Tableau 3: Taux de reconnaissance en fonction des différents sous-ensembles de V_1.

L'utilisation d'un seuil local automatique pour construire le réseau pose des problèmes pour localiser les différences de deux mots; nous avons essayé une construction manuelle, pour le locuteur2 pour lequel le nombre d'erreurs est le plus élevé, de façon à bien intégrer les régions non pertinentes ensemble, et séparer les régions discriminantes, ceci en ajustant manuellement le seuil local d'intégration en fonction du locuteur et des deux mots à séparer. Une amélioration de 10% est obtenue pour les sous-ensembles *poids/toit*, *bois/doigt*.

Discussions

Les taux de reconnaissance dépendent énormèment de l'apprentissage utilisé. Les résultats du tableau 1 montrent un gain de 10% pour l'apprentissage multi-référence et 7% pour la technique de la moyenne 2 par rapport à l'apprentissage aléatoire mono-référence. Il apparait que la technique de la moyenne 2 constitue le meilleur compromis performance/coût-mémoire. Si le réseau discriminant donne des bons résultats sur les vocabulaires de mots distincts par leurs centres comme celui du paragraphe IV, il n'en est pas de même pour d'autres vocabulaires. En effet, lorsqu'il s'agit de discriminer entre les mots distincts par leurs attaques (*poids*, *toit*) ou par leurs finales (*lemme*, *laine*), on rencontre quelques difficultés pour construire des réseaux convenables à partir d'un ensemble d'apprentissage, et pour chaque locuteur. Le réseau discriminant automatique s'avère moins intéréssant pour le vocabulaire V_1: légère amélioration de 1% par rapport à une méthode classique, et 0.5% par rapport à la moyenne 2 (voir tableaux 1 et 3). Une analyse manuelle plus fine des paires *bois*/*doigt* et *poids*/*toit* (sur lesquelles le réseau crée des erreurs), a montré que les différences se situent plutôt dans les voyelles qui accompagnent les deux attaques et non aux attaques des deux mots comme on l'aurait imaginé. Une telle information n'est pas portée par les distances locales qui guident la construction du réseau. Un autre problème peut intervenir dans le cas où l'un des mots a une durée plus longue que celle de son voisin. En effet le processus de comparaison dynamique impose avec ses contraintes locales, un chemin qui n'est pas toujours optimal. Ainsi le réseau discriminant n'est pas très fiable puisqu'il n'associe pas exactement les régions discriminantes des deux mots entre elles. Un processus de compression temporelle serait nécessaire dans ce cas, pour réduire les parties stables du mot[9].

Conclusion

Les expériences décrites dans le paragraphe III montrent que la qualité de reconnaissance dépend énormèment du choix de l'ensemble de références, donc du critère de sélection. Aussi, le réseau discriminant développé dans le paragraphe IV peut apporter une amélioration considérable pour les mots acoustiquement voisins distincts par leurs parties centrales, mais son interêt s'avère moindre pour les mots distincts par leurs attaques ou finales. Cependant, l'expérience 3 montre qu'on peut améliorer la qualité de reconnaissance en ajustant manuellement le seuil d'intégration en fonction du locuteur et du sous-ensemble étudié.

Références

[1] H. Sakoe et S. Chiba, "Dynamic programming optimisation for spoken word recognition", *IEEE Trans. Acoust., Speech, and Signal Proc.*, ASSP 78.

[2] G.L. Bradshaw, R. Cole, Z. Li "A Comparison of Learning Techniques Speech Recognition". *Proc. IEEE ICASSP-82.*

[3] M.J. Roussell, J.C.A. Deacon, R.K. Moore "Some Implications of the Effect of Template Choice on the Performance of an Automatic Speech Recognition". *Proc. IEEE ICASSP-84.*

[4] L.R. Rabiner, A.E. Rosenberg, S.E. Levinson, J.G. Wilpon "Speaker Independant Recognition of Isolated Words Using Clustering Techniques", *IEEE Trans. Acoust., Speech, and Signal Proc.*, ASSP 79.

[5] L.R. Rabiner and J.G. Wilpon "Isolated Word Recognition Using a two Pass Pattern Recognition Approch", *Proc. IEEE ICASSP-81.*

[6] R.K. Moore, M.J. Russell, M.J. Tomlinson "The Discriminative network; A Mechanism for focusing Recognition in Whole-Word Pattern Matching", *Proc. IEEE ICASSP-83.*

[7] R.K. Moore, D. Beardsley, M.J. Russell, M.J. Tomlinson "Towards An Integrated Discriminative Network for Automatic Speech Recognition", Proceedings of Institute of Acoustics 82.

[8] R.K. Moore, M.J. Russel, M.J. Tomlinson "Automatic Speech Recognition Using Local Timescale Variability Information", *Proc. IEEE ICASSP-82.*

[9] J.L. Gauvain, J.J. Mariani et J.S. Liénard, "On the use of time compression for word-based recognition", *Proc. IEEE ICASSP-83.*

APPLICATION DES λ^2 - DISTRIBUTIONS A LA RECONNAISSANCE DE LA PAROLE

C. DEMARS [1] - *J.L. GAUVAIN* [2]

[1] UER de Physique Université de Paris VII Place Jussieu 75005 Paris
[2] L.I.M.S.I. (C.N.R.S.) BP.30 91406 Orsay Cedex

Abstract

The purpose of this paper is to present an attempt at applying instantaneous frequency methods to word recognition. This method is used to produce a new type of time versus frequency representation corresponding to the distribution of the energie of the signal along the instantaneous frequency axis. After a preliminary processing of temporal normalisation the representation of the words to be recognize are matched to the references by means of dynamic programming. Features chosen for this approach are presented along with the results obtained in speaker dependent isolated word recognition.

I. Introduction

Il existe en analyse de parole plusieurs méthodes désormais classiques: analyse spectrale (FFT, banc de filtres), analyse cepstrale, codage predictif, etc... Malgré le perfectionnement constant apporte à leur utilisation en reconnaissance de parole, certains problèmes d'analyse restent à résoudre. Ceux-ci sont d'autant plus importants que la qualité de toute la chaîne de reconnaissance est fortement tributaire de l'analyse acoustique [1]. Par ailleurs on peut penser que les problèmes de reconnaissance ne sont pas des problèmes de mesure au sens physique du terme. Les algorithmes ont seulement besoin au départ d'une représentation définie de façon univoque à partir du signal, dans la mesure où la correspondance entre le "signe" et le "sens" est arbitraire. C'est ce qui nous a amené à proposer un nouveau type d'analyse acoustique liée à la notion de fréquence instantanée et donc un nouveau type de représentation temps-fréquence correspondant à la distribution d'énergie sur l'axe des fréquences instantanées [2][3].

On rappelle quelques propriétés de ce type d'analyse
i) localisation: Alors que l'analyse de Fourier permet de regarder les propriétés moyennes d'une portion de signal, l'analyse proposée autorise l'examen des variations locales
ii) séparation naturelle de l'information sur la phase et l'amplitude
iii) grande précision et résolution simultanément sur l 'amplitude et la fréquence. En effet alors que l'analyse spectrale utilise une notion de fréquence issue de l'analyse harmonique, l'analyse proposée est fondée sur la notion de fréquence instantanée conçue comme la vitesse angulaire d'un point dans un plan, une des coordonnées étant le signal, l'autre le signal en quadrature [4][5]. Elle n'est donc pas sujette aux mêmes contraintes que l'analyse spectrale en particulier à l'inégalité de GABOR : $\Delta f \times \Delta T \simeq 1$. Dans une réalisation analogique [6] une variation de fréquence peut être détectée sur une seule oscillation, la fréquence d'un signal de fréquence constante (1000Hz) et de durée très brève (5ms) peut être déterminée avec une erreur inférieure à 7%, il est possible de suivre une variation rapide de fréquence sur un temps très court (de 250 à 2000 Hz en 50ms). De plus il y a continuité: Le système fournit en sortie 2 fonctions du temps $\mathcal{E}(t)$ enveloppe et $\varpi(t)$ fréquence instantanée du signal qui sont, dans le cas d'une réalisation analogique, des fonctions continues du temps. Dans le cas d'un système numérique, après la réponse impulsionelle à la mise en route *initiale*, pour chaque échantillon d'entrée le système fournit, à la fréquence d'échantillonnage, 2 valeurs $\mathcal{E}(t)$, $\varpi(t)$, avec un délai de 11 échantillons dans la réalisation actuelle. Lorsque dans une bande de fréquence il n'y a qu'un seul harmonique le sytème fournit un excellent détecteur de mélodie [7][8]. Cette méthode d'analyse a été utilisée par ailleurs en géophysique [9][10]. En conclusion les propriétés de ce type d'analyse nous ont paru suffisament intéréssantes pour chercher à savoir si une telle approche pouvait être appliquée à l'analyse de la parole et plus particulièrement en reconnaissance.

II. Définition des distributions λ^2

La notion de fréquence instantanée est liée à celle de signal analytique introduite par J. Ville consistant à associer au signal $x(t)$ le signal en quadrature $y(t)$. On construit ainsi un signal complexe $x(t) + iy(t)$ dont le carré du module est appellé enveloppe et l'argument $\theta(t)$ est appellé la phase. La dérivée de la phase

$$\varpi(t) = \frac{d\theta}{dt} = d(Arctg\frac{y}{x}) = \frac{xy' - yx'}{x^2 + y^2} \qquad (1)$$

est la fréquence instantanée du signal. A tout signal $x(t)$ on associe 2 fonctions du temps $\mathcal{E}(t)$ son enveloppe et $\varpi(t)$ sa fréquence instantanée. On lui fait correspondre une distribution $T(\varpi, t)$ où $T(\varpi, t)$ est l'enveloppe i.e. l'énergie du signal à l'instant t pour la fréquence ϖ. Par intégration par rapport au temps on obtient une fonction $\lambda^2(\varpi)$ de la seule variable ϖ qui représente une distribution de l'énergie sur l' axe des fréquences instantanées. On peut effectuer cette intégration sur tout le signal ou sur une tranche de durée fixe qui est déplacée progressivement. On obtient ainsi dans le plan temps-fréquence instantanée une représentation de type analogue à celle des sonogrammes. La figure 1 présente la fréquence instantanée, l'enveloppe, la distribution $\lambda^2(\varpi)$ d'un signal artificiel: sur la partie gauche de haut en bas le signal, sa fréquence instantanée, son enveloppe. Le signal est engendré par la stimulation périodique d'une cellule du 2ème ordre dont le pôle et la fonction de transfert sont représentés sur la partie droite de la figure (graphiques supérieurs). Le

3ème graphique en bas à droite représente en échelle logarithmique la distribution d'énergie sur l'axe des fréquences instantanées. On remarquera que la fréquence de résonance est bien mis en évidence; la largeur de la distribution correspond à l'excursion en fréquence dans le graphique donnant la fréquence instantanée en fonction du temps. La figure 2 présente sur la partie gauche de haut en bas un signal de parole, attaque du mot "DEUX", son enveloppe, sa fréquence instantanée et sur la partie droite de haut en bas les distributions λ^2 de 6 tranches du signal prises successivement, avec recouvrement. La figure 3 est pour sa partie inférieure une représentation de type sonographique des distributions $\lambda^2(\varpi)$ d'une suite de tranches, de 128 points, prises successivement, du mot "SIX". L'énergie dans chacun des canaux (de largeur 25 t/s) est codée en niveaux de gris suivant une échelle sensiblement logarithmique avec interpolation d'une tranche à l'autre pour lisser l'image. On peut calculer sur ces distributions toutes sortes de paramètres secondaires par exemple le pourcentage d'énergie sur les fréquences négatives ou encore le moment d'ordre 1 qui s'interprète comme la fréquence moyenne. Le graphique supérieur représente l'évolution au cours du temps de ce moment d'ordre 1.

III. Application à la reconnaissance

1. *Le prétraitement acoustique*

Après avoir été échantillonné à 10kHz et digitalisé sur 12 bits par un convertisseur A/D le signal a été préaccentué à l'aide d'un différentiateur récursif à une section [11]. Il est ensuite appliqué à l'entrée d'un réseau déphaseur numérique et à large bande dont les sorties sont déphasées entre elles de 90 degrés dans la bande 75-5000Hz avec une erreur inférieure à 1 degré. Chaque branche du réseau est constituée de 4 filtres passe-tout en cascade dont la fonction de transfert est donnée [12] par

$$H(z) = \prod_{i=1}^{4} \frac{z_i - z^{-1}}{1 - z_i z^{-1}}$$

La détermination effective des coefficients z_i en fonction des spécifications du réseau se trouve dans la référence [13]. Les sorties du système ne sont pas en réalité l'enveloppe et la fréquence instantanée du signal d'entrée parce que les deux sorties d'un réseau déphaseur sont fondamentalement distordues en phase. On pourrait tourner la difficulté en utilisant des transformées de Fourier en revenant au domaine temporel après avoir mis à zéro les composantes supérieures à la moitié de la fréquence d'échantillonage; on obtient alors le signal en quadrature par rapport au signal d'entrée. Il se trouve que l'oreille n'est pratiquement pas sensible aux distorsions de phase (loi d'Ohm pour l'audition) et l'expérience montre qu'après traversée du réseau les signaux de sortie ne sont pas discernables du signal d'entrée. Les signaux de sortie $x(t)$ et $y(t)$ sont appliqués ensuite à une paire de filtres différentiateurs optimaux, à phase linéaire, à réponse impulsionelle finie de 15 coefficients dont la fréquence maximum est 4000Hz avec un maximum d'erreur de 0.00014 [14]. Le calcul de l'enveloppe et de la fréquence instantanée ont été faits sur des tranches de 300 points décalées de 100 points, sans fenêtre de pondération. Pour les fréquences positives 8 canaux de fréquence instantanée ont été définis selon une échelle de Barks [15], pour les fréquences négatives un seul canal. L'énergie présente dans chaque canal est calculée par sommation. Ces valeurs sont codées selon une échelle logarithmique. On obtient alors dans le plan temps-fréquence instantanée une représentation de la distribution d'énergie du signal dans chaque canal en fonction du temps.

2. *La reconnaissance*

Il s'agit de reconnaissance globale c'est-à-dire acoustique dans un contexte monolocuteur effectuée en trois étapes

i) une détection de frontières

Elle est faite par seuil sur l'enveloppe par un algorithme qui prend aussi en compte les valeurs suivantes afin d'éviter une détection érronée sur un bruit parasite local.

ii) une compression temporelle

Elle est destinée à résoudre les problèmes d'alignement temporel entre un mot inconnu et une référence, un locuteur ne pouvant prononcer plusieurs fois un même mot avec exactement le même rythme et la même durée. Ces problèmes peuvent être résolus de manière très efficace par un algorithme de comparaison dynamique(16) qui toutefois présente les inconvénients de nécessiter une importante quantité de calculs et d'imposer de sévères limites sur les distorsions temporelles acceptables si on veut utiliser un algorithme optimisé. On a utilisé ici un processus de normalisation temporelle basé sur le fait que les distorsions temporelles affectent surtout les zones stables du signal et qu'un son stable peut être représenté par un nombre réduit d'évènements [17].

iii) un algorithme de comparaison dynamique

Il est destiné à mettre en correspondance les échelles temporelles des séquences à comparer par une transformation non linéaire afin d'obtenir une coïncidence optimale

3. *Résultats et discussion*

Le vocabulaire est constitué des 10 chiffres 0 à 9 prononcés en Français. On a enregistré 120 mots prononcés de façon aléatoire par un même locuteur. Les tests ont été effectués en deux phases, dans la première phase les 10 premiers mots sont pris pour référence, dans la deuxième les 10 suivants, soit en tout 200 reconnaissances. Dans une première application on obtient 92% de taux de reconnaissance. Dans une deuxième application une interpolation linéaire a été effectuée simultanément sur l'enveloppe et la fréquence instantanée afin que ces deux fonctions soient mieux définies: Les variations de ces fonctions, plus particulièrement celles de la fréquence instantanée peuvent être très rapides. Il faut en effet se rappeler que dans le cas d'une réalisation digitale du système on n'obtient qu'un échantillonage des deux fonctions (qui sont dans la réalité continues), et ce, à la fréquence d'échantillonnage du signal, fréquence qui n'est pas forcément suffisante pour représenter correctement les variations des fonctions. On obtient alors 99% de reconnaissance. On remarquera qu'améliorer l'interpolation a augmenté le pourcentage de reconnaissance. D'autre part un problème relativement important et délicat pour la reconnaissance est celui de la détection des frontières. Il n'est pas sûr que l'algorithme actuellement retenu soit bien adapté à la nouvelle méthode d'analyse. Enfin le choix des canaux, issu de l'analyse spectrale, n'est pas forcément transposable à une analyse basée sur une autre notion de fréquence; ce qui pose le problème général du choix des paramètres à fournir à l'algorithme de reconnaissance.

IV. Comparaison avec des approches classiques

1. *Relation avec une analyse spectrale*

Les calculs effectués ont permis de vérifier que, conformément à la théorie [2], le moment d'ordre 1 de la distribution λ^2 et du spectre sont égaux et que le moment d'ordre 2 de la distribution λ^2 est inférieur au moment correspondant du spectre. Il s'agit d'une vérification pratique et expérimentale: les calculs de l'enveloppe et de la fréquence instantanée sont effectués en assembleur sur 16 bits, les intégrations sur une tranche du signal et non sur tout le signal comme le demande la théorie. Il existe entre les moments d'ordre supérieur des deux distributions des relations

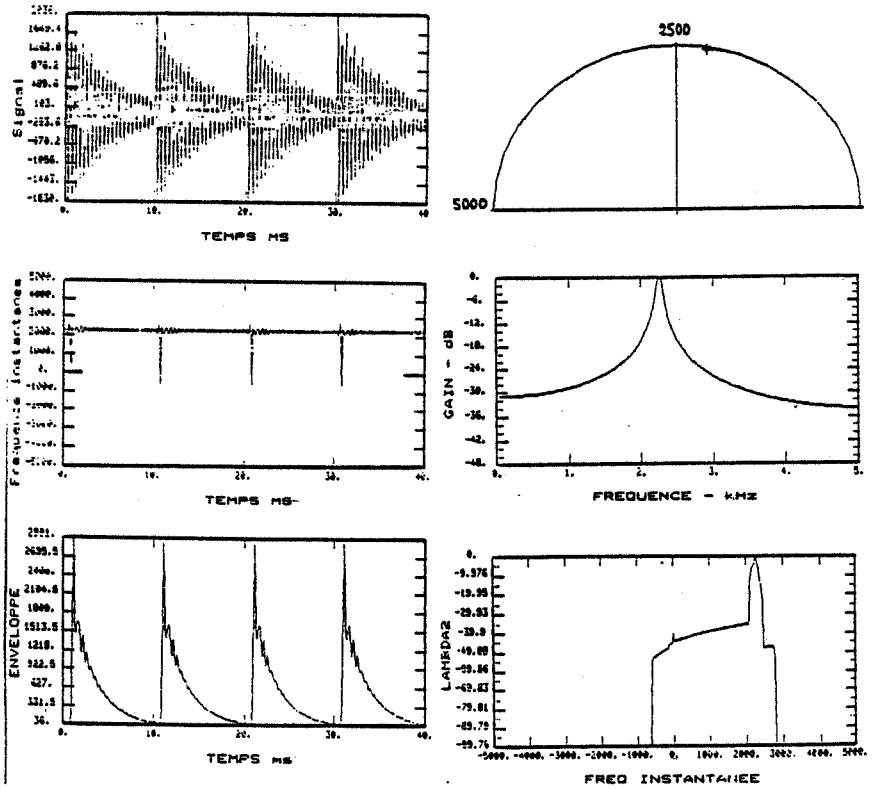


Fig.1 Partie gauche: de haut en bas, le signal, sa fréquence instantanée son enveloppe. Partie droite: de haut en bas, pôle de la fonction de transfert, spectre et distribution $\lambda^2(=)$ du signal

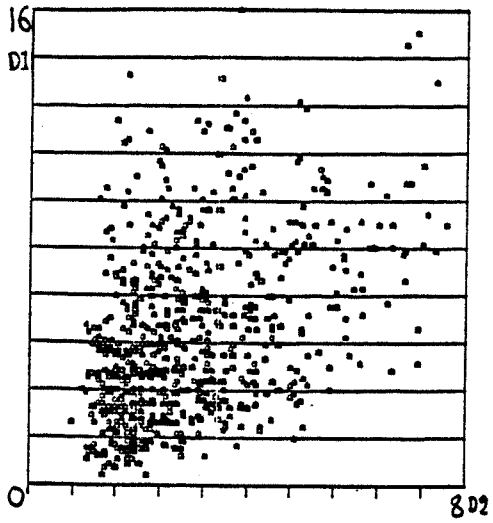


Fig.5 Nuage de points dans le plan $D1/D2$ des distances sur les λ^2 distributions et sur les coefficients cepstraux.

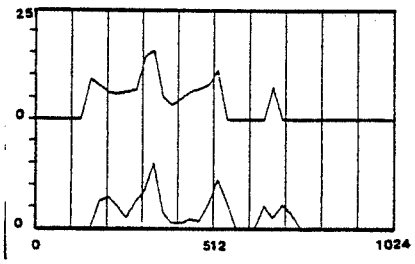


Fig.4 Distance entre deux tranches successives du mot "CINQ" en abscisse temps en ms, en ordonnée distance en dB en haut analyse cepstrale, en bas analyse en fréquence instantanée

de plus en plus complexes avec l'ordre [18] que nous n'avons pas cherché à vérifier.

2. *Relation avec une analyse cepstrale*

On a cherché à mettre en relation la représentation proposée avec une représentation de type cepstrale: Pour chaque mot on calcule, pour deux tranches de 256 points, successives, et sans recouvrement, d'une part la distance $D1$ entre les deux distributions λ^2 correspondantes:

$$D1 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ln P(i) - \ln PP(i))^2} \qquad (2)$$

d'autre part la distance $D2$ calculée par une formule analogue sur les coefficients cepstraux 2 à 31, lissés par une demi fenêtre de Hamming. La fig.4 présente en fonction du temps les distances entre deux tranches successives du mot "CINQ" respectivement pour les distributions λ^2 et les coefficients cepstraux: Les principaux évènements phonétiques semblent convenablement détectés. La fig.5 présente le nuage de points dans le plan $D1/D2$ pour un corpus comprenant deux fois les chiffres de 0 à 9 et 24 autres mots entièrement voisés, en abscisse distance $D2$ sur les cepstres, en ordonnées distance $D1$ sur les distributions λ^2, en décibels. Il ne semble pas y avoir de corrélation significative entre les deux distances

3. *Relation avec la représentation de Wigner-Ville*

Le fait que la fréquence instantanée n'est autre que la fréquence moyenne de la distribution de Wigner-Ville [19] serait intéressant à exploiter.

V. Conclusion

Une application à la reconnaissance de mots isolés d'une méthode basée sur la notion de fréquence instantanée a été présentée. Le problème principal est celui de la détermination, à partir de l'analyse, de paramètres pertinents. Un autre problème intéressant est de chercher à préciser les relations théoriques ou pratiques que peut entretenir la représentation proposée avec celles proposées par d'autres méthodes (analyse spectrale, analyse cepstrale, transformées de Wigner-Ville) de façon à mieux cerner le domaine spécifique à chacune des méthodes.

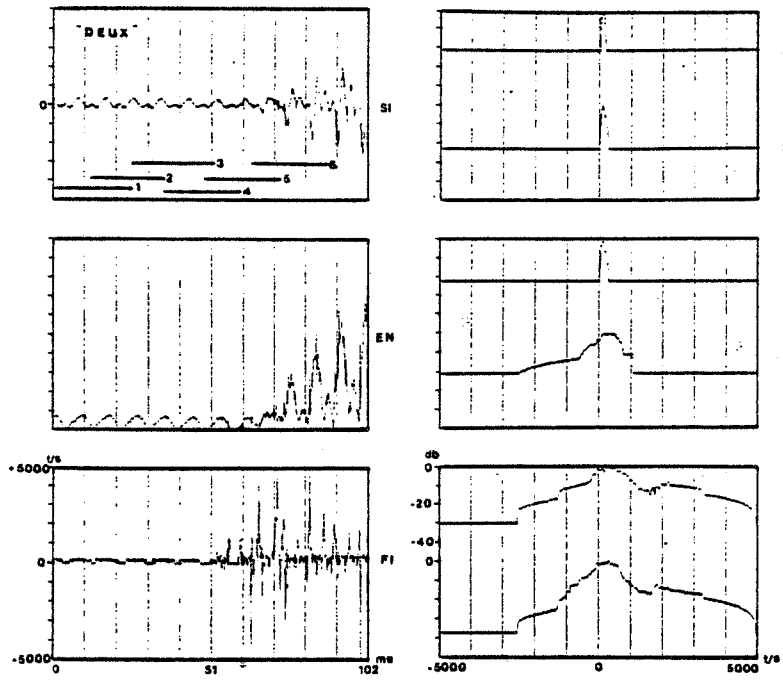


Fig.2 Partie gauche: de haut en bas. le signal de parole, son enveloppe sa fréquence instantanée. Partie droite: distributions $\lambda^2(\varpi)$ des 6 tranches indiquées

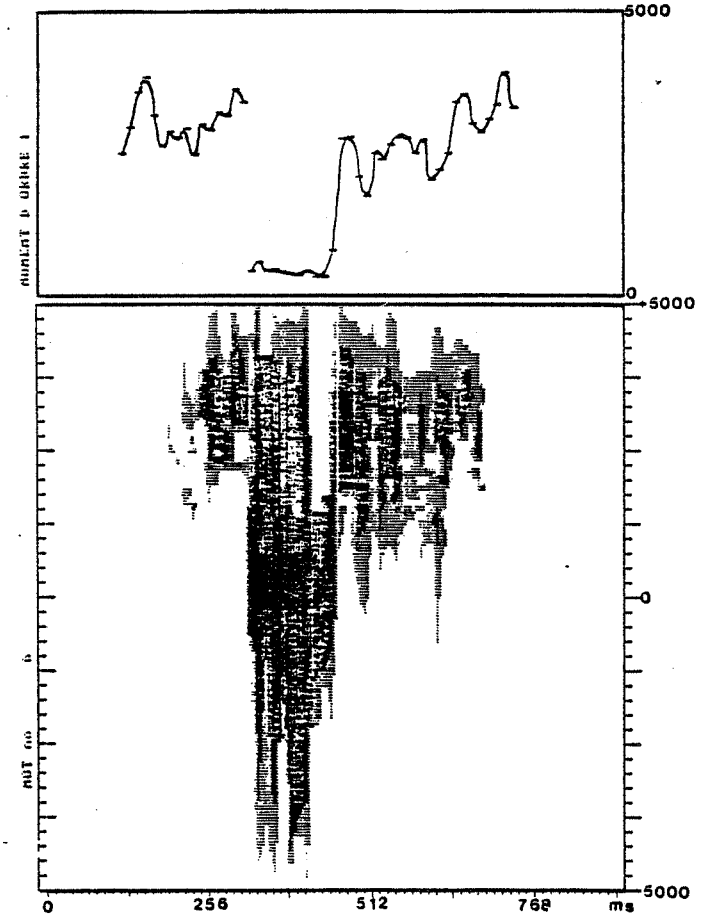


Fig.3 Représentation de type sonographique du mot "SIX": a) Distribution $\lambda^2(\varpi)$ b) Moment d'ordre 1 de cette distribution

VI. Bibliographie

[1] J. Mariani. "ESOPE: Un système de compréhension de la parole continue". Thèse d'état, Université de Paris 6, 1982.

[2] C. Berthomier, "Représentation d'un signal dans un plan fréquence instantanée-temps". Thèse doctorat d'état Université Paris 7 1976.

[3] C. Berthomier, "Instantaneous frequency and energy distribution of a signal". Signal processing 5(1983).

[4] D. Gabor, "Theory of communication" Inst. Elect. Engrs.93 PartIII. No.26. 1946.

[5] J. Ville, "Théorie et applications de la notion de signal analytique". Cables et Transmissions, F(Janvier), 1948 (hors commerce).

[6] R. Riguet, "Calculateur analogique de l'enveloppe, de la fréquence et de la phase instantanée de signaux naturels". Thèse d'université, Paris. 1974.

[7] C. Demars, C. Berthomier, M. Goustard. "The ontogenesis of the 'Great Call' of gibbons (*Hylobates Concolor*)" Proceedings of the 6th congress of the Internatinal Primatological Society Cambridge(U.K) 1976 Recent advances in primatology Vol I. Academic press London 1978.

[8] C. Demars, C. Berthomier, M. Goustard, "The great call of *Hylobates Concolor Hainanus*, comparison with the homologous emission of H. *Concolor Gabriellae* and *H.Concolor Leucogenys*". Perspectives in Primates Biology Ed. P.K. Seth, 1983.

[9] M. T. Taner, F. Koehler, R. E. Sheriff, "Complex seismic trace analysis" Geophysics, 44, No. 6 (June 1979).

[10] C. Berthomier, N. Cornilleau-Wehrlin, "Application de la notion de signal analytique à la détermination de l'amplitude et de la fréquence instantanée" Ann.Telecom. , 30, No. 7-8, 1975.

[11] L. R. Rabiner, Stieglitz, "The design of wide-band recursive and non recursive digital differentiators". IEEE Trans.on Audio and electroacoustics vol. AU-18, no. 2, June 1970.

[12] B. Gold, C. Rader, "Digital processing of signals". Mac Graw Hill 1969.

[13] C. Demars, "Détermination pratique d'un réseau déphaseur à 90 degrés numérique et à large bande". Onde Electrique, avril 1984.

[14] L.R. Rabiner, R. W. Shafer. "On the behavior of minimax relative error FIR digital differentiors" The Bell System Technical Journal, vol. 53, no. 2, Feb. 1974.

[15] J.L. Gauvain, "Reconnaissance de mots enchainés et détection de mots dans la parole continue" Thèse 3ème cycle, Université de Paris XI, Juin 1982.

[16] H. Sakoe, S. Chiba "Dynamic programming algoritm optimisation for spoken word recognition". IEEE-ASSP, vol. ASSP-26, no. 1, Feb. 1978

[17] J.L. Gauvain, J. Mariani, J. S. Liénard, "On the use of time-compreession for word-based recognition". ICASSP-83, Boston, 1983.

[18] L. Mandel, "Interpretation of Instantaneous Frequencies". American Journal of Physics, vol. 42, october 1974.

[19] T.A. C.M. Claasen, W.F.G. Mecklenbrauker, "The Wigner distribution - A tool for time-frequency signal analysis". Philips-Journal of Research Vol. 35, No. 3, 1980.

UN ALGORITHME DE RECONNAISSANCE DE MOTS ENCHAINES AVEC CONTRAINTES SYNTAXIQUES

A.Boyer, J.di Martino, J.P.Haton

C.R.I.N., Equipe "Reconnaissance des Formes et Intelligence Artificielle"
BP 239 54506 Vandoeuvre.

ABSTRACT

This paper deals with an algorithm for connected word recognition, using syntactic constraints, and a new formulation of dynamic programming.

1. INTRODUCTION

L'une des principales difficultés rencontrées en reconnaissance de mots isolés ou enchaînés est de déterminer correctement les débuts et fin d'élocution. La plus simple, mais aussi la plus fastidieuse des méthodes, parmi celles qui sont utilisées à l'heure actuelle, consiste à isoler les formes vocales manuellement à l'aide d'un logiciel de visualisation du signal de parole. Cette technique nécessite un investissement énorme dans le cas de corpus de plusieurs milliers de mots, ou phrases et ne saurait de toute facon être retenue dans des systèmes de reconnaissance automatique de la parole. Une deuxième approche consiste à utiliser un algorithme de détection parole-non parole plus ou moins heuristique. Malheureusement, ce type d'algorithme n'est guère fiable lorsque les débuts ou fins de mots sont faiblement énergétiques ou lorsque le signal est bruité. Toutefois, c'est cette méthode avec correction manuelle des erreurs de segmentation les plus importantes dans la phase d'apprentissage qui prévaut à l'heure actuelle. Une autre approche envisageable, et que nous allons expliciter dans ce papier, est de modifier l'algorithme de comparaison dynamique afin qu'il puisse à la fois trouver le meilleur recalage temporel entre les deux formes comparées, mais aussi déterminer la meilleure segmentation compte tenu de certaines contraintes temporelles.

Dans la première partie de ce papier, après un bref rappel de la technique de programmation dynamique, nous exposons la méthode de comparaison dynamique avec relâchement des contraintes aux frontières, qui permet une segmentation implicite des formes vocales comparées, à la fois dans le cadre de la reconnaissance de mots isolés et de mots enchaînés.

Dans une deuxième partie, nous appliquons cette technique à un algorithme de reconnaissance de mots enchaînés, en introduisant de plus des contraintes syntaxiques. De telles contraintes, puisqu'elles permettent de grouper les mots en classe, limitent les choix possibles à chaque étape du processus de comparaison. Dès lors, nous espérons augmenter considérablement la taille du vocabulaire à taux de reconnaissance égal avec des méthodes sans contraintes syntaxiques. Toutefois, ces améliorations seront obtenues au prix de contraintes supplémentaires pour le locuteur qui devra obligatoirement respecter la syntaxe définie.

2. METHODE DE COMPARAISON DYNAMIQUE AVEC RELACHEMENT AUX CONTRAINTES

2.1. Quelques rappels sur la programmation dynamique

Soient T=t(1)...t(I) et R=r(1)...r(J) deux formes où t(i) et r(i) sont des vecteurs caractérisant le signal de parole, obtenus par exemple par une technique de paramétrisation classique du type vocoder, lpc, cepstre, etc...., lors de la ième unité de temps(l'unité de temps généralement adoptée varie entre 10 et 20 ms) et où I et J sont les longueurs en prélèvements des formes T et R. Le but de la programmation dynamique est de trouver le meilleur chemin de recalage entre les formes T et R, afin de compenser les distorsions dues aux variations non linéaires du rythme d'élocution.

Afin que les chemins de recalage respectent l'évolution dans le temps du signal de parole, ceux-ci sont soumis à des conditions de monotonie ainsi qu'à des contraintes locales. De plus, pour que tous les échantillons soient pris en compte dans le processus de comparaison dynamique, les conditions aux frontières suivantes sont imposées:

(1) $i(1)=1,\ j(1)=1,\ i(K)=I,\ j(K)=J$

où K est le nombre de points du chemin de recalage. Ces contraintes aux frontières sont très sévères car elles sont fondées sur l'hypothèse implicite que les formes vocales ont été correctement segmentées. Or, comme nous l'avons signalé en introduction, la détection parole-non parole est loin d'être encore parfaite, le taux d'erreur variant entre 0 et 10% suivant le locuteur et l'algorithme utilisé. Ce taux d'erreur se retrouve généralement intégralement dans le taux d'erreur de reconnaissance. L'approche que nous proposons est de relâcher les contraintes (1) tout en maintenant le caractère optimal de l'algorithme de programmation dynamique, afin que la segmentation soit réalisée par ce dernier.

Le problème du recalage temporel est résolu par la minimisation d'une métrique D sur l'ensemble des chemins de recalage, que l'on définit classiquement comme étant un cumul normalisé des distances entre les formes mises en comparaison: (2)$\hat{w}=\text{argmin } D(w)$. La relation (2) peut être résolue par programmation dynamique grâce au principe d'optimalité de Bellman[1]. Désignons par D(i,j) la distance associée au chemin de recalage partiel optimal aboutissant au point (i,j), et par dp((i',j'),(i,j)) la distance pondérée entre le point (i',j') et (i,j) en suivant un chemin local imposé par la contrainte locale utilisée.Le principe d'optimalité ainsi que la relation (1) permet de résoudre (2) par la relation récurrente suivante:

$$(3)\quad D(i,j)=\min_{(i',j')\in V(i,j)} D(i',j')+dp((i',j'),(i,j))$$

avec (i',j') appartenant au voisinage V(i,j) du point (i,j) défini par la contrainte locale utilisée.

2.2. Programmation dynamique avec relâchement aux contraintes

1ère phase:il s'agit de déterminer le point par lequel passe le chemin optimal partiel aboutissant au point (i,j), en tenant compte des différences de longueur des chemins de recalage pouvant aboutir au point (i,j):

$$(4)\quad (\hat{i},\hat{j})=\text{argmin}_{(i',j')\in V(i,j)}\left(di(i',j')/li(i',j')\right)$$

avec $di(i',j')=D(i',j')+dp((i',j'),(i,j))$
et $li(i',j')=L(i',j')+l((i',j'),(i,j))$
(argmin f donne l'argument qui rend f minimale).

où L(i',j') est la longueur du chemin optimal aboutissant au point (i',j') et l((i',j'),(i,j)) est la longueur du chemin local entre les points (i',j') et (i,j).

2ème phase:le chemin optimal étant déterminé grâce à $(\hat{i},\hat{j})$,les relations (5), (6) permettent d'évaluer D(i,j) et L(i,j):

$$(5)\quad D(i,j)=D(\hat{i},\hat{j})+dp((\hat{i},\hat{j}),(i,j))$$
$$(6)\quad L(i,j)=L(\hat{i},\hat{j})+l((\hat{i},\hat{j}),(i,j)).$$

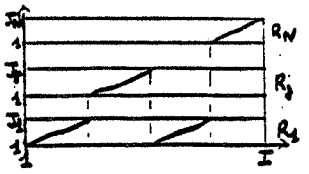
2.3. TECHNIQUE DE RELACHEMENT AUX CONTRAINTES APPLIQUEE A LA RECONNAISSANCE DE MOTS ENCHAINES

L'un des algorithmes les plus performants à l'heure actuelle en reconnaissance de mots enchaînés est celui de Bridle-Nakakagawa [3][4].Il s'agit d'un algorithme en une passe, donc parfaitement adapté au temps réel. Le principe de l'algorithme consiste à déterminer le chemin de recalage optimal dans un espace à trois dimensions, comme l'illustre la figure 1, où un point est repéré par ses trois coordonnées:

-n définissant le sous espace relatif à la forme de référence Rn auquel appartient le point

-i étant l'indice au niveau de la forme inconnue

-j étant l'indice relatif à Rn



-figure 1-Exemple de chemin de recalage d'après Bridle-Nakagawa

Grâce à ce formalisme, le problème de la reconnaissance de mots enchaînés est résolu dans l'algorithme de Bridle-Nakagawa par programmation dynamique à l'aide de l'équation récursive:

$$(7)\quad D(n,i,j)=\min_{(n',i',j')\in V(n,i,j)} D(n',i',j')+dp((n',i',j'),(n,i,j))$$

avec (n',i',j') appartenant au voisinage V(n,i,j) du point (n,i,j).

Dans le cas de la contrainte locale sans condition de pente, Ney[5] a montré que deux types de voisinage doivent être définis:

-si j=1, on parle alors de contrainte "externe" car elle intervient à la frontière de deux formes de référence. Le voisinage de (n,i,j) s'écrit alors :

$V(n,i,j)=\{(n,i-1,j),(n',i-1,Jn')\}$

où n' varie de 1 à N (avec N:nombre de formes de référence) et ou Jn' représente la longueur en prélevements de la forme n'.

-si j>1, on parle cette fois de contrainte "interne" car elle n'intervient qu'à l'interieur d'une forme de référence. Alors:

$V(n,i,j)=\{(n,i-1,j),(n,i-1,j-1),(n,i,j-1)\}$.

Du fait que la relation (7) peut être considérée comme une simple généralisation de l'équation (3), les relations récursives dans le cas d'un relâchement aux contraintes s'obtiennent immédiatement par extension des relations (4)(5)(6):

-relation de première phase:

(8) $(\hat{n},\hat{i},\hat{j})=\mathbf{argmin}\ di(n',i',j')/li(n',i',j')$
$(n',i',j')\in V(n,i,j)$

avec
$di(n',i',j')=D(n',i',j')+dp((n',i',j'),(n,i,j))$ et
$li(n',i',j')=L(n',i',j')+l((n',i',j'),(n,i,j))$ où (n',i',j') appartient à V(n,i,j).

-relations de deuxième phase:

(9) $D(n,i,j)=D(\hat{n},\hat{i},\hat{j})+dp((\hat{n},\hat{i},\hat{j}),(n,i,j))$
(10) $L(n,i,j)=L(\hat{n},\hat{i},\hat{j})+l((\hat{n},\hat{i},\hat{j}),(n,i,j))$.

3. UN ALGORITHME DE RECONNAISSANCE DE MOTS ENCHAINES AVEC CONTRAINTES SYNTAXIQUES:

Nous présentons un algorithme de reconnaissance de mots enchaînés, qui tout en utilisant les techniques de la programmation dynamique avec relâchement aux contraintes, prend en considération des contraintes syntaxiques. Dans un premier temps, nous allons décrire les contraintes syntaxiques dans les algorithmes de programmation dynamiques avant d'expliciter un algorithme les utilisant.

3.1. Les contraintes syntaxiques dans un algorithme de programmation dynamique:

Soient $S=\{s1,....,Sn\}$ un ensemble fini d'états et $V=\{R1,....,RN\}$ le vocabulaire de l'application considérée.

Une transition t, ou contexte d'une forme de référence, définie sur V et S est un élément de SxVxS: t=(x,n,y) où (x,y) appartient à SxS et n à V.

Un réseau N défini par S et V est un sous ensemble de SxVxS.

Soient t=(x,n,y) et t'=(x',n',y') deux transitions. Nous dirons que t<=t' si y=x'.

Nous définissons trois fonctions p1,p2,p3: à toute transition (x,n,y),p1 associe x,p2 n,p3 y.

Nous appelons chemin du réseau N joignant les états a et b toute suite t1,....,tz de transitions de N telles que p1(t1)=a,p2(tz)=b, et ti<=ti+1 pour i variant de 1 à z-1. Nous notons P[a,b] l'ensemble de tous les chemins joignant a et b.

Le réseau N est dit complet si il existe deux états particuliers xs et xe, appelés respectivement noeud de départ et noeud d'arrivée, tels que: pour tout y de S, il existe p appartenant à P[xs,xe] et t appartenant à p tels que p2(t)=y (t n'est pas nécessairement unique).

Nous dirons que N est un réseau syntaxique fini s'il est complet.

Une forme de référence n est dite dans l'état x s'il existe un contexte t tel que p2(t)=n et p1(t)=x.
Pour exprimer les contraintes syntaxiques, nous avons besoin de trois fonctions:Gamma qui donne tous les contextes aboutissant à un état particulier,Omega qui donne tous les contextes issus d'un état particulier,Phi qui donne l'ensemble des états qu'une forme de référence peut prendre.

3.2. Presentation de l'algorithme de reconnaissance de mots enchaînés avec relâchement aux contraintes et contraintes syntaxiques:

Nous avons vu précédemment que le plan de comparaison proposé par Bridle-Nakagawa[3][4] est de dimension 3. Dans ce plan de comparaison, il s'agit de trouver la "superforme" de référence optimale, obtenue par la concaténation d'un certain nombre de formes de référence, réalisant la meilleure coincidence au sens d'une certaine métrique avec la forme test. Pour résoudre ce problème, Bridle et Nakagawa [3][4] utilisent la programmation dynamique. Si l'on introduit des contraintes syntaxiques, il faut ajouter une nouvelle dimension à l'espace de comparaison pour tenir compte des différents états que peut prendre une forme de référence au cours du processus de comparaison. Cette nouvelle dimension est la dimension d'état. Tout point de l'espace de comparaison est alors représenté par un quadruplet (n,i,j,s). En ce point, le ième prélèvement de la

forme test est en coincidence avec le " jème prélèvement de la forme de rèfèrence n supposèe dans l'état s. Dans ce cas, et toujours dans le cadre du formalisme de Ney[5], le voisinage du point (n,i,j,s) s'écrit pour la contrainte sans condition de pente:(11)j>1:

V(n,i,j,s)={(n,i-1,j,s),(n,i-1,j-1,s),(n,i,j-1,s)}
(12)j=1:

V(n,i,1,s)=U {(n',i-1,Jn',s')} U {(n,i-1,1,s)}

avec (n',s') tel que (s',n',s) appartienne à Gamma(s).

Ces voisinages se rèècrivent très facilement dans le cadre du relâchement aux contraintes. Le voisinage V(n,i,j,s) d'un point quelconque (n,i,j,s) de l'espace de comparaison ayant été défini, la relation de programmation dynamique s'écrit dans le cas des contraintes syntaxiques, généralisant (5):

$$(13)\ D(n,i,j,s)=\min_{A\in V(n,i,j,s)} D(A)+dp((A),(n,i,j,s))$$

où A=(n',i',j',s') appartient à V(n,i,j,s).

di Martino [2] a montrè que pour dèterminer la superforme de rèfèrence optimale, il était indispensable d'introduire des pointeurs qui permettent de remonter le chemin optimal à la fin de la comparaison. Ces fonctions, au nombre de 4 dans le cas de contraintes syntaxiques, doivent être dèfinies en tout point du plan de comparaison (n,i,j,s) et donnent les composantes de l'origine du chemin partiel aboutissant en ce point. L'algorithme de reconnaissance de mots enchaînès avec relâchement aux contraintes et contraintes syntaxiques se décompose de la facon suivante:

1. Initialisation de D, L et des fonctions pointeurs.

2. Calcul en tout point (n,i,j,s) de l'espace de comparaison de D(n,i,j,s) et L(n,i,j,s) de la facon suivante:

première phase:

$$(14)\ (\hat{n},\hat{i},\hat{j},\hat{s})=\operatorname{argmin}_{A\in V(n,i,j,s)}(di(A)/li(A))$$

avec $di(A)=D(A)+dp((A),(n,i,j,s))$ et $li(A)=L(A)+1((A),(n,i,j,s))$ avec A=(n',i',j',s') appartenant à V(n,i,j,s). deuxième phase:

(15)D(n,i,j,s)=D(n̂,î,ĵ,ŝ)+dp((n̂,î,ĵ,ŝ),(n,i,j,s))
(16)L(n,i,j,s)=L(n̂,î,ĵ,ŝ)+1((n̂,î,ĵ,ŝ),(n,i,j,s))
et évaluation des fonctions pointeurs.

3. détermination de la meilleure sèquence de mots en minimisant D(n,I,Jn,s) pour (s,n) tel que (s,n,sm) appartient à Gamma(sm),où sm est le noeud final.

4. CONCLUSION

Nous avons prèsentè dans cet article le principe d'un algorithme de reconnaissance de mots enchaînès permettant d'une part le relâchement des contraintes aux frontières de mots,et d'autre part la prise en compte de contraintes syntaxiques. Cet algorithme est actuellement implantè sur un Exormax (Motorola) équipè d'un microprocesseur 68000 avec carte "array processor". Nous le testons sur un corpus de formes rèelles incluant les dix chiffres. Les rèsultats de ces tests feront l'objet d'une prochaine publication.

5. BIBLIOGRAPHIE

[1] R.Bellman "Dynamic Programming" Princeton, N.F:Princeton, Univ.Press, 1957

[2] J.di Martino "Contribution à la reconnaissance globale de la parole:mots isolès, mots enchaînès" Thèse de docteur ingènieur, Universitè de Nancy 1, 1984

[3]J.S.Bridle, N.D.Brown, R.M.Chamberlain "An algorithm for connectd word recognition" IEEE ICASSP, Paris, France, Mars 1982, pp.899-902

[4]S.I.Nakagawa "A connected spoken word recognition method by o(n) dynamic programming pattern matching algorithm" in Proc.1983, IEEE ICASSP, Boston, April, 1983, pp.296-299.

[5]H.Ney "The use of a one stage dynamic programming algorithm for connected word recognition" IEEE Transactions ASSP, vol.ASSP-32, april 1984, pp.263-271.

[6] A.Boyer "Etude d'algorithmes de reconnaissance de mots enchaînès" Rapport de DEA, 1984.

UTILISATION DES TECHNOLOGIES VOCALES DANS UNE APPLICATION MULTI-CANAUX

A. AOUATI, F. CHUPEAU(*), J. J. MARIANI, D. MEMMI

LIMSI CNRS BP 30 91406 ORSAY CEDEX.

(*)CENA ORLY SUD 94542 ORLY AEROGARE CEDEX.

ABSTRACT

An application of voice-based technologies to the specific area of Air-traffic control is presented. The important components of the system are the recognition device, the application-dependent syntax and semantic software, the speech understanding software, the voice data collection procedures and the feedback.

1.0 INTRODUCTION

Il s'agit, d'une part, de faire dialoguer un élève contrôleur avec le simulateur de trafic aérien, dans un langage opératif utilisé pour les échanges verbaux entre les contrôleurs et les pilotes, dans des exercices de formation. Et d'autre part, de retrouver dans une communication verbale, en situation de travail, l'information "INDICATIF", i.e le destinataire du message.

Notre préoccupation majeure est d'alléger, autant que possible, les contraintes d'élocution imposées aux utilisateurs et de satisfaire au mieux les éxigences que comporte l'utilisation de la reconnaissance globale de mots enchainés. En effet, le locuteur respecte pour la prononciation, des contraintes de débit, et pour la formulation des messages, les règles de la phraséologie utilisée.

Les résultats obtenus à ce jour en reconnaissance, montrent les améliorations apportées par la constitution de références acoustiques robustes pour l'ensemble des mots du lexique prononcés isolément d'une part, et d'autre part, par l'apprentissage effectué dans le contexte. Des raffinements sont obtenus si l'apprentissage tient compte de la spécificité du langage utilisé, en particulier pour guider le choix des séquences de mots à apprendre.

L'interprétation des assertions du locuteur est guidée par la connaissance que l'on a de la structure des messages émis, d'une part, et d'autre part, par la définition du contenu sémantique attaché aux entités syntaxiques du langage utilisé. Les énoncés reconnus sont transformés en une suite d'instructions d'un langage formel de commande, utilisé dans la communication avec le simulateur.

2.0 LE TERMINAL VOCAL

2.1 Constitution

Le terminal vocal est constitué d'une carte de reconnaissances de mots enchainés et de détection de mots dans la parole continue MOZART, d'une carte de synthèse de la parole à partir du texte ICOLOG, et de la carte intelligente sur laquelle est implanté un interpreteur VLISP.
Les carte sont organisées dans une architecture multiprocesseur. Les processeurs communiquent via un interface multibus, et la carte intelligente est dans la configuration MAITRE. Chaque carte ESCLAVE partage une zone mémoire RAM avec la carte intelligente, pour les besoins du dialogue.

3.0 LA CONSTITUTION DU LANGAGE

Nous avons implanté le système ARBUS qui permet à des utilisateurs non-informaticiens de décrire facilement le langage qu'ils désirent utiliser, de le modifier, et de le completer au fur et à mesure de l'utilisation du système. Le langage construit peut être sauvegardé sur fichier, en mémoire de masse. Le langage est représenté par des arborescences. L'analyseur syntaxique associé au système est descendant

gauche-droite. Il peut analyser des grammaires hors-contexte, avec branchements vers des sous-arbres. Il n'accepte pas les grammaires ambiguës, et effectue un retour arrière dès qu'est trouvée une analyse correcte pour le mot à analyser.

4.0 L'APPRENTISSAGE ET LA SPECIFICITE DU LANGAGE UTILISE

4.1 Le lexique

Pour les exercices visés, nous avons constitué un lexique suffisant (60 mots) pour exprimer la plupart des messages à émettre. Le lexique comporte, outre les mots et les chiffres, des expressions telles que: tournez à gauche, tourner à droite, taux de descente, taux de montée, qui peuvent être considérées comme entités lexicales. Par conséquent, à chaque expression il correspond une image acoustique dans le dictionnaire des références au lieu de trois.

4.2 La phraséologie

La multiplicité des formes d'expression, comme elle apparait dans la formulation des paramètres - 410 s'énonce "quatre unité zéro" ou "quatre cent dix" - tend à accroitre la perplexité du langage d'une part, et d'autre part, le nombre des références acoustiques.

4.3 La syntaxe

Les contraintes syntaxiques sont utilisées, par le système de reconnaissance, pour prédire, donc pour restreindre l'univers de recherche aux seuls mots du lexique pouvant se présenter à un niveau de l'arbre syntaxique. En outre, elles déterminent le nombre maximum de mots dans la phrase prononcée.
Nous avons établi la liaison entre le système ARBUS et le système de reconnaissance de mots enchainés MOZART. Ainsi le premier apporte au second les contraintes syntaxiques lors de la reconnaissance.

5.0 LA CONSTITUTION DU DICTIONNAIRE DES REFERENCES

L'unité de décision étant le mot, Le soin apporté à cette opération conditionne les performances du système.
La liaison entre la carte intelligente et MOZART permet la sauvegarde et le chargement des références acoustiques par transfert de fichiers entre la mémoire de masse et MOZART. En outre, tous les paramètres de travail peuvent être modifiés en interactif, donc sur le site.

5.1 La méthode utilisée

Pour chaque entité du lexique nous constituons une image acoustique robuste. A cet effet le locuteur prononce chaque mot du vocabulaire jusqu'à ce que deux références d'un même mot soient suffisamment proches (au sens d'un critère de dissemblance ajustable en intéractif). C'est la dernière référence qui est conservée. Au cours de cette opération les occurrences d'un même mot ne sont pas consécutives.

Pour la plupart des mots les références acoustiques ainsi obtenues donnent des résultats satisfaisants.

Pour les chiffres et les mots courts, la coarticulation prend une importance considérable. Par conséquent, pour chaque mot, plusieurs références sont nécessaires. A cet effet l'apprentissage est effectué en deux phases:
- Une première phase au cours de laquelle une référence robuste est constituée.
- Une deuxième phase où le locuteur énonce des séquences de mots qui seront ségmentées par comparaison aux références robustes.

Ainsi nous constituons pour les mots qui engendrent des risques d'ambiguité:
- Une référence isolée
- Une référence en milieu de phrase
- Une référence en fin de phrase

5.2 Le choix des séquences de mots à apprendre

Le choix des séquences à apprendre est guidée par la spécificité du langage utilisé. Les paramètres des instructions de contrôle font intervenir des séquences de chiffres prononcés en fin de phrase et souvent à un rythme rapide. Cette élocution rapide en conjonction avec un facteur de branchement de la grammaire élevé rend leur reconnaissance particulièrement difficile.
Une première amélioration est obtenue par un apprentissage des suites de mots utilisés pour exprimer les paramètres afin de tenir compte des effets de coarticulation et de prosodie. On garde ainsi une image acoustique apprise dans le contexte pour chacun des mots de la suite prononcée.

5.2.1 Les mots acoustiquement voisins -

On observe aussi des erreurs de reconnaissance dues aux mots semblables qui se présentent au même niveau de l'arbre syntaxique. Ces mots acoustiquement voisins engendrent des risques d'erreur par un branchement erroné. Une réduction sensible de ces erreurs est obtenue en ajustant le seuil de ségmentation pendant l'apprentissage des séquences de mots et le seuil de sélectivité au cours de la reconnaissance.

5.2.2 Influence des seuils de sélectivité - L'ajustement des seuils reste une tâche délicate. Si des seuils sévères éliminent ce type d'erreurs, ils augmentent le taux de rejets des phrases prononcées.
Le rejet est de loin préférable à une mauvaise reconnaissance d'autant plus qu'un branchement erroné pourrait engendrer une phrase sémantiquement acceptable qui passerait le filtre de l'analyse conceptuelle.

5.3 Resultats

Les résultats des tests conduits avec un échantillon de phrases et de clauses usuellement prononcées par les locuteurs pour exprimer des instructions de contrôle et les paramètres associés ont permi une mesure des ambiguités acoustiques restantes.

Pour les essais effectués en local, cent phrases ont été constituées. Les résultats obtenus sont de 98 pour cent pour les mots et de 95 pour cent pour les phrases pour trois locuteurs masculins, avec des contraintes syntaxiques rigides et un facteur de branchement statique moyen de 10.

6.0 LA GéNéRATION DES CLAUSES POUR LE SIMULATEUR DE TRAFIC

La communication avec le simulateur de trafic met en jeu des concepts peu nombreux. Les échanges s'effectuent au moyen d'un langage formel de commande.
La commande "CAP/AF93/180/D" signifie:

Désignation du Cap 180 à l'avion Air France 93. Il s'agit de transformer les assertions du locuteur en instructions simulateur cohérentes et directement exécutables.
Les phrases reconnues vont être représentées par une structure suffisament expressive qui permet d'une part, un test de cohérence, et d'autre part, la génération du code exécutable.

6.1 Interface avec le simulateur

L'interface avec le simulateur de trafic, pendant la reconnaissance en ligne, est réalisé au moyen d'un descriptif qui est une représentation du sens de la phrase.
Donnons un exemple:
Phrase reconnue: °Air France 9 3 Prenez Cap 2 6 0§
Voici comment apparaissent les clauses simulateur dans la représentation utilisée

```
° (INDICATIF (AF93 (Air France 9 3)))
(INSTRUCTION (CAP (Prenez Cap)))
(PARAM (260 (2 6 0))) §
```

Le descriptif comme il apparait dans l'exemple est un schéma instancié. Ses slots sont initialement vides.
Le travail d'instantiation est effectué par des requêtes lexicales auxquelles sont associées des requêtes sémantiques.
Les requêtes lexicales ont pour objet la recherche d'une unité lexicale particulière. Cette recherche est effectuée par exploration de l'arbre d'analyse obtenu par une analyse syntaxique préalable de la phrase reconnue.
Les requêtes sémantiques effectuent le travail d'instantiation proprement dit des slots du descriptif.

6.2 La rétroaction

Le formalisme utilisé pour représenter la réponse du système à un énoncé est trés expressif, et la détection d'une phrase incorrecte ou incomplète s'effectue à peu de frais.
Pour les énoncés incomplèts, les slots qui n'auront pas été instanciés apparaitront dans le descriptif avec une valeur indéfinie. On voit donc la possibilité offerte pour réaliser une rétroaction "intelligente", en suggérant au locuteur, par synthèse de messages appropriés, la répétition des clauses manquantes ou mal reconnues.

6.3 Situation

Le travail d'évaluation qui est mené permettra une mesure des ambiguités sémantiques. L'utilisation du système dans l'univers de l'application et la constitution d'un recueil de données adaptées à une analyse statistique va permettre une évaluation objective des performances réelles.

BIBLIOGRAPHIE

1 DANIEL G.BOBROW, RONALD M.KAPLAN, MARTIN KAY,DONALD A.NORMAN, HENRY THOMPSON AND TERRY WINOGRAD. GUS, A frame-driven dialog system -Artificial intelligence, 8, 1977, 155-173.

2 F.CHUPEAU ET J.GOUBERT- Indicatifs d'appel radio -CENA/R 84-04 FEV 1984-

3 P. FALZON - Les communications verbales en situation de travail: Analysé des restrictions du language naturel. Rapport technique INRIA 1982.

4 J. L. GAUVAIN -Reconnaissance de mots enchainés et détection de mots dans la parole continue. Thèse de docteur ingénieur, juin 1982.

5 MICHAEL W.GRADY -Air intercepter controller prototype - A system engeneering view of advanced speech technologie performance standards.

6 STAN C.KWASNY - Relaxation techniques for parsing grammatically ill-formed input in natural language understanding systems. American journal of computational linguistics, VOLUME 7, NUMBER 2, APRIL-JUNE 1981

7 D. MEMMI, J. MARIANI - ARBUS: A tool for developping application grammars- COLING, PRAGUE, 1982.

8 J. MARIANI -ESOPE: Un système de compréhension de la parole continue- Thèse de doctorat es-sciences PARIS VI, 1982.

9 P.R. MICHAELIS, A. CHAPANIS, G.D. WEEKS and M.J.KELLY Word usage in interactive dialog with restricted and unrestricted vocabularies- IEEE Transactions on professional communication VLO. PC-20, NO.4 DEC 1977

ELABORATION D'UN SYSTEME EXPERT EN LECTURE DE SONAGRAMMES

P-E. STERN, M. ESKENAZI, D. MEMMI

LIMSI-CNRS, B.P 30 91406 ORSAY Cedex

ABSTRACT

We describe the knowledge-based system we have designed to integrate a phonetician's expertise in speech spectrogram reading. This knowledge might prove useful within a continuous speech recognition system.

Description of the spectrogram is carried out through an interactive menu-driven program which includes specific knowledge about the image itself.

A forward-chaining control structure (inference engine) has been written in PROLOG 2. It uses an anchor point progressive strategy which manages certainty factors.

INTRODUCTION

Pour répondre à la complexité des problèmes de reconnaissance et de synthèse de la parole, on a maintenant recours à des techniques d'intelligence artificielle et particulièrement aux systèmes experts . Cette approche est d'autant plus séduisante qu'elle permet de formaliser des connaissances sur la parole qui ne pouvaient pas être prises en compte par un système plus classique.

Un système expert se décompose en trois parties:
- La base de faits contient la description du problème.
- Un ensemble de règles de production traduisant la connaissance de l'expert.
- Un moteur d'inférence qui est la structure de contrôle du système.

L'intérêt d'une telle technique est de bien séparer l'ensemble des connaissances utilisées de l'usage que l'on peut en faire, ce qui permet d'affiner progressivement cet ensemble.

Nous nous intéressons ici plus particulièrement aux connaissances nécessaires au décodage phonétique. L'expertise utilisée est celle de la lecture du sonagramme, une représentation spectrale de la parole. Certains experts phonéticiens obtiennent en effet, à partir du sonagramme, une transcription phonétique de la phrase prononcée avec des résultats meilleurs que ceux des systèmes actuels de décodage phonétique [1]. La connaissance utilisée par ces experts constitue une source d'information nouvelle pour les systèmes de reconnaissance. Différentes études montrent l'intérêt naissant de cette approche [2] [3][4].

Pour permettre au système expert de décoder un sonagramme, on doit lui décrire l'image sonagraphique et définir une structure informatique apte à accueillir l'expertise et à l'utiliser en fonction de cette description.

I DESCRIPTION DU SONAGRAMME ET BASE DE FAITS

Lorsqu'on présente un sonagramme inconnu à notre expert (ME), celui-ci effectue tout d'abord une segmentation de l'image suivant l'axe temporel. Les segments correspondent, à quelques exceptions près, à un découpage phonémique. L'expert propose ensuite pour chaque segment un ou plusieurs phonèmes. Nous supposons ici la segmentation réalisée.

1) Description du sonagramme

Durant la tâche de décodage du sonagramme, l'expert ne relève dans l'image que les motifs qui lui semblent pertinents. Cependant, les éléments de description qu'il fournit comportent déjà une part d'interprétation. Par exemple, le terme qu'il utilise pour décrire un "formant" d'une consonne nasale est différent de celui qu'il emploie pour le formant d'une voyelle. Si l'on tenait compte de cette part d'interprétation, seul l'expert serait vraiment capable de décrire le sonagramme au système.

Nous avons donc répertorié tous les éléments de description utilisés par l'expert en termes pseudo-phonétiques et recherché leur signification dans l'image, ce qui nous a permis de dégager un ensemble de primitives de description. L'image peut ainsi être décrite d'une manière moins subjective par un non expert. Une description parfaitement objective du sonagramme pourrait être obtenue directement par un module de traitement d'image, en utilisant une définition structurelle en termes de textures uniformes. Nous avons, sur ce sujet, débuté une collaboration avec une spécialiste en reconnaissance des formes (C.Faure ENST).

La description de l'image comporte un aspect global et un aspect segmental.

a) Aspect global

L'expert effectue une observation globale avant de décoder le sonagramme.

Il évalue la durée moyenne (accentuée et non accentuée) d'un segment en fonction de la durée de tous les segments dans l'image. Pour chaque segment, il juge alors sa durée par rapport à cette durée moyenne. L'on peut ainsi détecter deux

plosives sourdes consécutives au milieu du groupe rythmique.
L'intensité de chaque segment est également jugée par rapport à l'ensemble des autres segments, ce qui permet, par exemple, de différencier certaines fricatives.

b) Description segmentale

Les textures du sonagramme correspondent à un certain nombre d'indices de description (formant, bruit, silence). On distingue donc trois textures dans l'image : régulière, irrégulière et les zones blanches.

L'intensité des indices est en outre décrite de manière relative à l'intérieur d'un même segment (exemple: le troisième formant est plus marqué que le second).

- Texture régulière

Pour chaque formant on indique :

- sa position: sa fréquence centrale, sa largeur.
- son intensité: peu marqué, marqué, très marqué.
- la direction et l'étendue de ses transitions.

De plus, pour décrire le bas du spectre on distingue trois types de formes: haut voisement, bas voisement, irrégularité dans les basses fréquences.

- Texture irrégulière

Le bruit est d'abord décrit de manière globale: par son intensité, son étendue fréquentielle. Les concentrations particulières d'énergie dans le segment bruité (rappel des formants voisins) sont décrites par leur fréquence centrale et leur intensité relative. Les explosions sont décrites d'une manière analogue au bruit.

- Zones blanches

Le terme silence permet de décrire les parties sans énergie correspondant par exemple aux occlusives sourdes.

Ces notions,assez éloignées de termes purement phonétiques, ont été introduites, en collaboration avec notre expert, pour permettre une description plus proche de l'image.

2) Représentation des faits.

Il s'agit maintenant, à partir des éléments de description du sonagramme, de constituer la base de faits qui sera utilisée par le moteur d'inférence. Nous avons tenu compte de la structure de ces éléments et cherché à obtenir une représentation de type objet. L'emploi de PROLOG nous a contraint à éclater celle-ci sous une forme logique (relationnelle). Les objets sont les indices de description (bruit, formant...). Le premier argument de chaque indice est le numéro du segment considéré. Les descripteurs de cet indice (position, intensité...) se réfèrent à celui-ci par son nom en premier argument d'une manière semblable à SNARK [5]. Chaque fait est affecté d'un poids fonction de la confiance qu'on lui porte (cf II) :

```
fait(indice(num-segment,nom),poids)->;
fait(descripteur(nom,valeur),poids)->;
```

Cette représentation est transparente pour l'utilisateur, toute la description étant réalisée à l'aide d'un éditeur qui connait la structure des faits et propose une description arborescente sous forme de menu. Cet éditeur est totalement déclaratif, ce qui permet de modifier et de compléter aisément la représentation interne des faits. Il assure également la consistance de la base de faits. De plus, comme la description n'est jamais exhaustive, le moteur interroge l'utilisateur sur les faits manquants au cours du déclenchement des règles.

On peut ainsi effectuer une description rapide de l'image, s'approchant du langage naturel et donc accessible à un utilisateur ne connaissant pas PROLOG.

II CONNAISSANCES ET STRUCTURE DE CONTROLE

Examinons à présent les stratégies et connaissances utilisées.

1) Description de la structure de contrôle.

Dans un problème d'identification, comme la lecture de sonagrammes, le raisonnement n'aboutit pas forcément à un stade terminal (un seul phonème candidat) et il serait difficile de formuler des buts à la manière d'un système de diagnostic. L'identification appelle donc un raisonnement inductif, c'est à dire, en termes de contrôle, du chaînage avant.

Le langage PROLOG, utilisé pour écrire le système, réalise lui-même un certain nombre de tâches (unification et résolution sur des clauses de Horn). Nous avons pu ainsi implanter rapidement un moteur avec variables. Afin de contrôler la stratégie de résolution, l'interpréteur PROLOG est utilisé non comme moteur d'inférence mais comme langage de programmation.

La structure de contrôle de notre système est une simple clause récursive:

```
moteur ->
   definir-probleme(x) regle(x,d,g)
   tester-conditions(d) conclure(g)      (1)
   / moteur ;
moteur ->;
```

2) Stratégies.

Au cours du décodage, l'expert utilise de nombreuses stratégies. Nous décrivons ici celles qui nous ont paru les plus intéressantes et que nous avons intégrées dans notre système.

a) Ilôts de confiance

Notre expert ne lit pas le sonagramme de gauche à droite mais recherche des segments qu'il qualifie de plus faciles à identifier. C'est en s'appuyant sur les premières informations déduites qu'il poursuit son raisonnement sur les segments voisins. Il reviendra éventuellement sur les segments qui ont constitué les ilôts de confiance, pour approfondir leur identification.

Ce comportement peut sembler artificiel par rapport à une lecture linéaire dans le sens de production de la parole, car il est indépendant de la structure de l'image que l'on cherche à reconnaître. Cependant cette approche souligne l'importance d'une vue globale de l'image pour reconnaître dans les segments des caractéristiques qui les distinguent nettement des autres (silence, bruit). Pour réaliser cette approche globale, il faut décrire l'intégralité du sonagramme (ou au moins tout un groupe rythmique) avant de chercher à l'identifier. Nous avons, de plus, rangé les règles par paquets, chacun traitant d'un même sujet. Au niveau de la structure de contrôle, la clause "definir-probleme" utilisée dans (1) permet de choisir le sujet du problème à résoudre. Les problèmes sont ordonnés en fonction de la démarche par ilôts de confiance. Chaque problème est étudié

en profondeur c.a.d que l'on cherche à déclencher toutes les règles qui le concernent. On reviendra sur un problème ancien si de nouveaux faits permettent de déclencher des règles sur celui-ci.

Cela nous a également incité à utiliser des plausibilités car, sur ces ilôts, l'expert fournit des informations plus sûres que sur les autres segments.

b) Raisonnement progressif.

L'expert raisonne de manière progressive à partir de traits acoustiques ou phonétiques. Lorsqu'il n'est pas en mesure de fournir un ou deux candidats, l'expert peut généralement préciser si le segment est, par exemple, occlusif ou fricatif. Parallèlement, il détermine le point d'articulation, en grande partie à l'aide du mouvement des formants, et peut qualifier un segment comme dental, palatal, ou labial. Ces traits ne reflètent pas toujours une terminologie classique de phonétique mais correspondent à un ensemble cohérent par rapport à l'image.

c) Prise en compte d'éléments contextuels.

- indices contextuels.

L'expert recherche des indices acoustiques dans les segments voisins du segment analysé. Cette observation est souvent indispensable pour effectuer l'identification d'un segment.Elle peut porter sur un contexte très large, jusqu' à trois segments plus loin dans certains cas.

- contraintes phonologiques.

Dès qu'un segment est identifié de manière certaine, l'on peut parfois réduire l'espace de recherche pour un segment adjacent en tenant compte de contraintes de types phonologiques. Ainsi, la suite /f/ /s/ /f/ n'est pas possible à l'intérieur d'un groupe rythmique. L'intégration de ces contraintes avec le niveau décodage phonétique permet d'identifier un segment de manière plus souple que le passage d'un treillis dans une matrice de précédence phonologique.

La prise en compte du contexte nécessite l'utilisation de variables dans les règles notamment pour représenter le numéro du segment. Des fonctions spéciales permettent d'instancier les contextes voisins (adjacent, a-droite, a-gauche).

d) Emission d'hypothèses.

Il arrive également à l'expert de formuler des hypothèses. Il cherche alors à les confirmer ou les rejeter en regardant si elles sont compatibles avec certains indices présents dans l'image.

Dans une première version de notre système la gestion des hypothèses était explicite [6]. Il fallait spécifier pour chaque règle si elle concluait sur une (ou plusieurs) hypothèses ou s'il s'agissait d' une solution certaine. Nous avons adopté à présent la solution des coefficients de vraisemblance qui nous paraît plus souple et plus conforme aux différentes stratégies exprimées par notre expert.

e) Gestion des plausibilités.

Les lois de combinaisons choisies sont inspirées de celles du système expert MYCIN [7]. Lorsque plusieurs règles ont des conclusions communes, le poids de ces conclusions se trouve renforcé. On simule ainsi à la fois la pondération des conclusions exprimées par l'expert et la notion de renforcement des hypothèses.

3) Les connaissances utilisées.

Nous avons axé notre étude, pour l'instant, sur un corpus de phonèmes restreint: 12 phonèmes représentant des oppositions de classes phonétiques: sourd/sonore, dentale/vélaire... Nous avons utilisé 200 sonagrammes numériques, prononcés par un même locuteur, dont 100 ont servis à évaluer la fiabilité de notre expert. Son taux de reconnaissance sur ceux-ci est de 95% de phonèmes corrects sur un treillis de 2,2 phonèmes en moyenne.

On présente sur la fig 1 un exemple de déclenchement de règles.

Les caractéristiques des connaissances utilisées sont liées aux stratégies déjà évoquées (cf II 2). On distingue deux types de connaissances en fonction de leurs sources et de l'utilisation qui en est faite.

a)Les règles d'identification.

Elles ont pour la plupart une forme du type:

si (ensemble de primitives de description)
et (contexte particulier) alors
(ensemble de phonèmes ou classes phonétiques)

Elles sont pondérées par des adverbes de plausibilité (sûrement peut-être..),
et s'appuient sur des connaissances phonétiques, mais également contextuelles (phonologie, coarticulation).

Ces règles sont des clauses spéciales:

regle(sujet, conditions, actions)->;

- sujet : on précise ici le fait prédominant dans la règle.
- conditions :

La syntaxe de la liste des conditions est très souple:

conditions = conditions et conditions
conditions = conditions ou conditions
conditions = non(conditions)
conditions = fait
conditions = procédure

Les procédures sont les fonctions arithmétiques ainsi que des fonctions plus spécifiques (adjacent, compris-entre). La condition est satisfaite pour un fait s'il est présent dans la base et si son poids est supérieur à un seuil. La négation correspond inversement à un poids inférieur à un seuil négatif.

- actions :

Il s'agit d'une liste de couples (fait,poids). Si la règle est déclenchée, ce fait sera ajouté dans la base ou remis à jour en fonction de son poids précédent.

Le choix de ces coefficients pose évidemment un problème. Dans la pratique nous avons tenu compte des adverbes exprimés par l'expert en langage naturel. Les valeurs exactes sont peu significatives, seuls le comportement du système et la cohérence de l'ensemble importent.

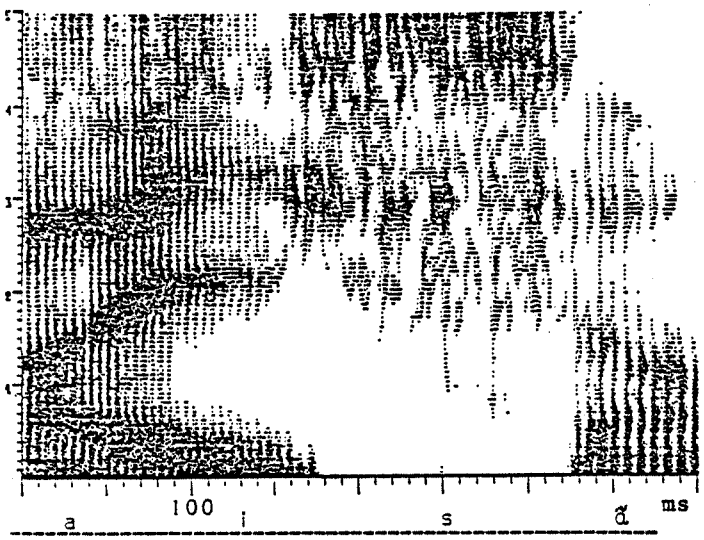
b) La description des classes phonétiques.

L'expert fournit l'ensemble des connaissances générales sur les différentes classes qu'il utilise. Cette connaissance est figée, elle s'exprime naturellement sous une forme arborescente:

occlusive ->occlusive-sourde
occlusive ->occlusive-sonore
occlusive-sourde -> /p/,/t/,/k/ e.t.c.

Cette connaissance particulière est donnée au système sous forme déclarative.

```
is-a(occlusive(c),occlusive-sourde(c)) ->;
is-a(occlusive-sourd(c),phoneme(c,"T"))->;
```




```
HYPOTHESE  : BRUIT(5) OU(ASPECT(BRUIT(5),MARQUE),
                         ASPECT(BRUIT(5),BIEN-MARQUE))
             NON(BAS-VOISEMENT(5))
CONCLUSION : <FRICATIVE-SOURD(5),PLUS(50)>
==> FRICATIVE-SOURD(5) DEDUIT AVEC LE POIDS PLUS(50)
==> PHONEME(5,"S") DEDUIT  AVEC LE POIDS PLUS(16)
==> PHONEME(5,"F") DEDUIT  AVEC LE POIDS PLUS(16)
-----------------------------------------------------
HYPOTHESE  : BRUIT(5) ZONE(BRUIT(5),<2500,5000>)
             COMPRIS(<2500,5000>,2500,5000)
             RIEN-DESSOUS(5,2000)
CONCLUSION : <PHONEME(5,"S"),PLUS(60)>
==> PHONEME(5,"S") DEDUIT  AVEC LE POIDS PLUS(67)
-----------------------------------------------------
HYPOTHESE  : BRUIT(5)  CONCENTRATION-BF(BRUIT(5))
             ZONE(CONCENTRATION-BF(5),3200)
             COMPRIS(3200,3050,3400)
CONCLUSION : <PHONEME(5,"S"),PLUS(80)>
==> PHONEME(5,"S") DEDUIT  AVEC LE POIDS PLUS(94)
-----------------------------------------------------
HYPOTHESE  : FORMANT-1(4)
             NOMBRE-FORMANT(4,4)  SUPEG(4,3)
CONCLUSION : <VOCALIQUE(4),PLUS(50)>
==> VOCALIQUE(4) DEDUIT  AVEC LE POIDS PLUS(50)
-----------------------------------------------------
HYPOTHESE  : FORMANT-1(4) FORMANT-2(4)
             FORMANT-3(4) FORMANT-4(4)
CONCLUSION : <VOCALIQUE(4),PLUS(50)>
             <VOYELLE(4),PLUS(80)>
==> VOCALIQUE(4) DEDUIT  AVEC LE POIDS PLUS(75)
==> VOYELLE(4) DEDUIT  AVEC LE POIDS PLUS(80)
-----------------------------------------------------
HYPOTHESE  : FORMANT-1(4)  ZONE(FORMANT-1(4),400)
             FORMANT-2(4)  ZONE(FORMANT-2(4),2100)
             ECART-SUP(400,2100,1500)
CONCLUSION : <VOYELLE-ECARTEE(4),PLUS(70)>
==> VOYELLE-ECARTEE(4) DEDUIT  AVEC LE POIDS PLUS(70)
==> VOYELLE-COMPACTE(4) DEDUIT AVEC LE POIDS MOINS(70)
==> PHONEME(4,"I") DEDUIT  AVEC LE POIDS PLUS(35)
-----------------------------------------------------
HYPOTHESE  : VOYELLE-ECARTEE(4) FORMANT-2(4) SEQ(4,5)
             PARTIE-DROITE(FORMANT-2(4),HORIZONTALE)
CONCLUSION : <TRANSITION(5,DENTALE),PLUS(50)>
==>TRANSITION(5,DENTALE) DEDUIT AVEC LE POIDS PLUS(50)
-----------------------------------------------------
HYPOTHESE  : VOYELLE(4) FORMANT-2(4)
             ZONE(FORMANT-2(4),2100) SUPEG(2100,1900)
             ASPECT(FORMANT-2(4),PEU-MARQUE)
CONCLUSION : <PHONEME(4,"I"),PLUS(70)>
==> PHONEME(4,"I") DEDUIT  AVEC LE POIDS PLUS(81)
-----------------------------------------------------
HYPOTHESE  : FRICATIVE-SOURD(5) TRANSITION(5,DENTALE)
CONCLUSION : <PHONEME(5,"S"),PLUS(70)>
==> PHONEME(5,"S") DEDUIT  AVEC LE POIDS PLUS(99)
-----------------------------------------------------
```

FIG 1: UN EXEMPLE DE DECLENCHEMENT DE REGLES SUR LES SEGMENTS 4 ET 5 (PORTION DU SONAGRAMME CI-DESSUS).

Le moteur utilise ces informations spécifiques à deux niveaux:

- Dans la partie condition des règles:

Exemple: Si on a conclu précédemment sur /t/, on peut utiliser le fait "occlusive", plus général, sans qu'aucune règle ne l'ait conclu. L'information est ainsi propagée vers le haut.

- Dans la partie conclusion des règles:

A chaque fois que l'on déduit une classe, l'ensemble des éléments de cette classe se trouve affecté d'un poids, fonction de la pondération de la règle et du facteur de branchement de cette classe (voir fig 1).

4) Interface.

Outre l'éditeur de faits, de nombreuses fonctions sont disponibles sous forme de menu et facilitent la mise au point de l'ensemble du système: trace, sauvegarde, chargement d'une base de faits, liste des règles, liste des faits, accès à l'éditeur PROLOG pour modifier ou ajouter des règles, résumé d'une session.

CONCLUSION

Le système réalisé permet de prendre en compte de nombreuses stratégies et d'intégrer progressivement les connaissances de notre expert tout en restant le plus proche possible de l'image dans la description du sonagramme. Ce travail souligne, en outre, l'importance d'une lecture globale de l'image et de la prise en compte du contexte pour la reconnaissance phonétique. Il nous reste à affiner l'ensemble des règles obtenues sur le corpus déjà étudié. Nous envisageons également d'étendre notre étude à d'autres phonèmes et à plus de locuteurs.

BIBLIOGRAPHIE

[1] R.A.COLE, A.I.RUDNICKY, V.W.ZUE, D.R.REDDY, "Speech as patterns on paper", Perception and Production of Fluent Speech, R.A.COLE ed, Lawrence Erlbaum, 1980.

[2] N.CARBONELL, D.FOHR, J-P.HATON, F.LONGCHAMP, J-M.PIERREL, "An expert-system for the automatic reading of french spectrogram", IEEE ICASSP 1984.

[3] J.JOHANSSEN, J.MACALLISTER, T.MICHALELE, S.ROSS, "A Speech Spectrogram Expert", IEEE ICASSP 1983.

[4] S.R.JOHNSON, J.H.CONNOLLY, E.A.EDMONDS, "Spectrogram Analysis: A Knowledge-Based Approach to Automatic Speech Recognition", to appear in "Proceedings of Expert Systems '84", C.U.P.

[5] J-L.LAURIERE, "SNARK: Un moteur d'inférences pour systèmes experts en logique du premier ordre", Rapport Institut de Programmation num. 430, Nov. 1983.

[6] D.MEMMI, M.ESKENAZI, J.MARIANI, P-E.STERN, "SONEX: Système expert en lecture de sonagrammes", Rapport d'ATP Intelligence Artificielle, Oct. 1984.

[7] E.H.SHORTLIFFE, MYCIN: Computer Based Medical Consultations, American Elsevier, New-York, 1976.

TECHNIQUES D'INTELLIGENCE ARTIFICIELLE EN DECODAGE ACOUSTICO-PHONETIQUE

N. Carbonell, J.P. Damestoy, D. Fohr,
J.P. Haton, F. Lonchamp, J.M. Pierrel

CRIN - Université Nancy I
Campus scientifique, B.P. 239
54506 Vandoeuvre les Nancy Cedex

ABSTRACT

The acoustic-phonetic decoding of a sentence is a major bottle-neck in continuous speech recognition. Our group has been working in this area for the past twelve years and has developed several systems relying on synchronous centi-second pattern-matching. In order to improve the performances of automatic phonetic-decoding, we have been developing for the past two years an expert production rule system that implements the competence of an expert spectrogram reader.

In the first part of this paper we describe the method used in order to acquire the expert's knowledge and strategies, together with results and the present state of our experimental system.

Complementary to and in close relationship with this appoach we are trying to improve the efficientness of acoustic-decoding systems, by using knowledge representation based on the notion of frame. This formalism makes it possible to use different types of acoustic analyses and phonetic features. It also allows to control the recognition process by a planning system. The second part of this paper gives an overall description of a system that represents the phonetic decoding process in terms of a frame language.

INTRODUCTION

La compréhension de la parole continue est, pour l'essentiel, une activité humaine qui résiste à l'analyse, ce qui explique en grande partie les performances médiocres des systèmes mono-locuteurs actuels de compréhension d'énoncés prononcés de manière naturelle, ainsi que les difficultés supplémentaires rencontrées lorsqu'on impose au système d'accepter en entrée des locuteurs différents, moyennant une phase d'apprentissage limitée.

Toutefois, en ce qui concerne le décodage acoustico-phonétique du processus de compréhension, on peut accéder plus facilement à la compétence humaine. En effet, la lecture de spectrogrammes de parole a donné lieu au développement d'une véritable expertise. Or il est plus facile d'analyser l'activité d'un expert dont le savoir-faire a été développé après la première enfance (ou cas particulier, à l'âge adulte), grâce à une expérimentation répétée (entraînement) et, parallèlement, à l'acquisition de connaissances acoustico-phonétique explicites (celles du phonéticien) que d'élucider l'activité en grande partie non-con-ciente que représentent la la perception et l'interprétation d'un énoncé.

Pour améliorer les algorithmes existants de décodage acoustico-phonétique, nous avons donc entrepris l'analyse et la modélisation du savoir-faire d'un expert en lecture de spectrogrammes ; ce qui nous a conduits naturellement à utiliser une approche de type système expert que nous présentons dans une première partie.

Mais, dans l'état actuel de la technique, il n'est pas certain - pour des raisons d'efficacité essentiellement - qu'il faille donner à la composante acoustico-phonétique d'un système de compréhension de parole continue la forme définitive d'un système expert. Parallèlement à l'approche système expert, nous avons développé un système de décodage phonétique à bases de connaissances qui utilise, pour représenter la compétence de l'expert, un formalisme fondé sur la notion de frame (prototype). Nous présentons les grandes lignes de ce travail dans la seconde partie.

L'APPROCHE SYSTEME EXPERT

Nous rappelons brièvement, dans un premier temps, la méthodologie que nous avons utilisée pour acquérir l'expertise ; on trouvera dans [1] et [3] une description détaillée de notre démarche. Nous présentons ensuite les développements récents de nos travaux ainsi que l'état actuel du projet.

Méthodologie d'acquisition de l'expertise et résultats

A. Première étape -

Nous avons conjointement :

- analysé l'activité de l'expert au cours de la lecture d'un spectrogramme correspondant à l'énoncé d'une phrase empruntée au corpus de Combescure [4], à partir des commentaires dont il accompagne son décodage, de ses remarques a posteriori sur les erreurs qu'il a commises et de discussions plus générales où il tente d'expliciter sa démarche,

- recueilli ses connaissances de phonéticien sur les divers sons du français, tout au moins celles qu'il jugeait utiles pour le décodage phonétique.

Nous disposons à l'heure actuelle :

- de descriptions précises de son activité :
  . compte-rendus des discussions qui ont accompagné le décodage en notre présence d'une quinzaine de spectrogrammes,

. analyse fine, d'une part, de ses commentaires (enregistrés) relatifs au décodage de vingt spectrogrammes, d'autre part, de ses remarques (enregistrées également) sur les erreurs commises au cours du décodage des vingt spectrogrammes évoqués précédemment et de quarante autres spectrogrammes, décodés également en dehors de notre présence,

- de règles d'identification des sons du français que nous avons déduites de l'analyse des commentaires de l'expert au cours du décodage des spectrogrammes, ou qu'il nous a fournies explicitement dans le cadre, soit de son activité, soit des discussions générales que nous avons eues avec lui ; toutefois, aucune règle n'a été introduite dans le système sans avoir été soumise, au préalable, à son jugement,

- d'une étude exhaustive des performances de l'expert pour chacun des sons du français ; cette analyse porte sur cinquante des soixante spectrogrammes évoqués plus haut, c'est-à-dire sur 650 réalisations de phonèmes (cf. [2]).

B. Deuxième étape -

Pour valider les règles obtenues au cours de l'étape A, plus généralement, pour construire de manière incrémentale un système expert qui mette en oeuvre les connaissances et la démarche de l'expert, nous avons

- numérisé (échantillonnage à 12 kHz) le signal correspondant aux phrases traitées par l'expert,

- segmenté et étiqueté avec l'aide de l'expert ce corpus ; deux étiquetages (et segmentations) ont été réalisés : l'un par sons, l'autre par événements phonétiques, le second décrivant le signal de manière plus fine que le premier.

Précisons que notre corpus comprend soixante phrases différentes (empruntées au corpus de Combescure) prononcées à un rythme naturel par six locuteurs masculins non entraînés. Nous envisageons de porter la taille de ce corpus à cent phrases (soit dix locuteurs). L'enregistrement des quarante phrases supplémentaires est terminé, leur décodage par l'expert en voie d'achèvement.

Actuellement, nous testons sur l'ensemble du corpus numérisé les règles obtenues au cours de la première étape à l'aide d'un système expérimental qui comprend, outre des procédures d'extraction de paramètres acoustiques, un moteur d'inférences très simple qui permet d'entrer ou de modifier conversationnellement une règle et fournit, en plus des conclusions de son raisonnement, une trace des règles appliquées successivement. Les écarts entre les résultats fournis par le moteur et ceux de l'expert sont analysés afin de combler les lacunes du système et améliorer ses performances (construction incrémentale). Nous précisons dans le paragraphe suivant le fonctionnement de ce moteur d'inférences.

Description du système dans sa version actuelle

Ce système, élaboré à des fins essentiellement expérimentales, comprend :

- un ensemble de modules de pré-traitement dont le rôle est d'extraire du signal, pour chacun des segments de l'énoncé, des informations qui constituent la base de faits associée au segment,

- un ensemble de règles et un moteur d'inférences qui, à partir de la base de faits relative à un segment, des règles dont il dispose et, éventuellement, de procédures d'analyse acoustique fine qu'il a la possibilité d'activer au cours du raisonnement, détermine la nature exacte d'un segment.

Ce découpage du traitement correspond à la démarche de l'expert qui, dans un premier temps, jette un regard d'ensemble sur le spectrogramme avant de commencer le décodage proprement dit de l'énoncé, segment par segment, de la gauche vers la droite. A noter que, de son propre avis, il pourrait aussi bien effectuer le décodage en progressant vers la droite et vers la gauche, à partir de points d'ancrage constitués par des segments facilement identifiables.

A. Le pré-traitement -

Les indices à partir desquels l'expert raisonne et construit la représentation d'un segment sont, pour la plupart, détectés par des algorithmes activés séquentiellement.

Ce sont des algorithmes globaux au sens où ils opèrent sur l'ensemble de la représentation d'un énoncé, non contextuels, dans la mesure où ils ne tiennent pas compte des altérations dues aux phénomènes de co-articulation.

Ils assument l'essentiel du découpage du continuum sonore en unités significatives car, pour l'expert, la segmentation est évidente dans la plupart des cas ; nous ne disposons donc que d'un nombre relativement restreint de règles de segmentation que l'expert utilise uniquement dans les cas difficiles : voyelles en hiatus, groupes sonante(s) et voyelle(s) ou groupes consonantiques comprenant une sonante. Ces règles permettent d'affiner la segmentation fournie par les algorithmes de pré-traitement dans les cas où identification et segmentation doivent s'effectuer de manière interactive, ou lorsque la segmentation nécessite des analyses complémentaires fines du signal.

Nous avons mis au point trois algorithmes qui effectuent une segmentation grossière de l'énoncé et une classification phonétique sommaire des segments trouvés ; ils s'exécutent dans l'ordre suivant qui nous paraît satisfaisant pour l'instant :

- détection des noyaux vocaliques des différentes syllabes de l'énoncé :
cet algorithme, à partir de l'analyse des variations temporelles de l'énergie comprise dans une bande de fréquences particulière (250 à 2350 Hz) :
  . détermine les noyaux vocaliques (maxima relatifs d'énergie) ainsi que leurs frontières ; 96 % des voyelles de notre corpus sont ainsi détectées ;
  . estime, à partir de la durée des différents noyaux trouvés, la durée moyenne (médiane) d'un segment ; cette valeur est utilisée comme référence au cours de l'analyse segment par segment de l'énoncé,

- détection des plosives :
cette procédure, en fait, met en évidence les zones de silence relatif dans l'énoncé, en comparant les valeurs de l'énergie obtenues (dans la bande de fréquence 700-6000 Hz) pour les zones comprises entre deux noyaux vocaliques avec celles trouvées, dans la même bande, pour ces deux noyaux ; plus de 80 % des occlusives sont détectées par cet algorithme en cours de mise au point ; nous espérons atteindre un taux voisin de 90 %,

- détection des fricatives :
on utilise plusieurs bandes de fréquences, en particulier 4000-6000 Hz et 250-1500 Hz ; la décision finale est prise après analyse par FFT du segment ; les performances atteintes sont comparables à celles que l'on obtient pour les plosives.

A noter que le découpage en segments fourni par ces algorithmes est grossier : en particulier, les sonantes sont soit détectées en tant que segments (vocalique ou fricatif), soit amalgamées aux segments vocaliques et fricatifs.

Outre ces trois algorithmes, nous avons mis au point des procédures qui extraient, pour chaque segment trouvé, les informations nécessaires à son identification ultérieure par le système expert :
. nombre et position des formants (pour un segment vocalique),
. répartition fréquentielle de l'énergie au sein du burst (pour un segment occlusif suivi d'une voyelle),
. etc.

B. Les règles et le moteur d'inférences -
Le pré-traitement fournit :

- pour chaque segment détecté, un ensemble de mesures acoustiques qui ne seront pas remises en cause au cours du raisonnement,

- pour certains d'entre eux, des informations qui pourront être remises en cause :
. l'indication d'une classe phonétique,
. et même, pour un petit nombre de segments, une identification complète ; en effet, les réalisations de certains phonèmes dans des contextes particuliers sont aisément identifiables à l'aide de critères simples non-contextuels ou bien par comparaison avec des formes de référence (cf entre autres /s/ en contexte non-labial).

Rappelons que les marques de segmentation résultant du pré-traitement constituent des informations sûres mais incomplètes, que la composante système expert pourra, éventuellement, enrichir au cours du raisonnement.

Les informations acquises au cours du pré-traitement sur chaque segment sont regroupées dans la base de faits relative au segment.

Le rôle de la composante système expert est d'associer à chaque segment défini par le pré-traitement, une liste de suites de phonèmes. Comme les règles d'identification sont souvent contextuelles, elle construit, pour l'ensemble de l'énoncé, un treillis de phonèmes.

a) les règles :

Il s'agit de règles de production du type :
prémisses → conclusions
où les prémisses sont construites soit à partir des faits initiaux (mesures, classe phonétique, liste de phonèmes), soit à partir des conclusions obtenues au terme d'une étape antérieure du raisonnement ; à noter que ces conclusions sont intégrées au fur et à mesure à la base de faits.

b) le moteur d'inférences :

* fonctionnement -
il dispose, en entrée :
. des résultats du pré-traitement via la base de faits associée au segment qu'il analyse,
. du signal de parole numérisé, par l'intermédiaire de procédures d'analyse acoustique, activables au cours du raisonnement.

Il analyse l'énoncé de la gauche vers la droite et fonctionne en chaînage avant ou arrière. Le chaînage arrière est utilisé par l'expert pour valider des hypothèses dont il n'est pas entièrement sûr ou lorsque, le signal comportant peu d'informations pertinentes, il essaie de réduire le nombre des interprétations possibles en les passant l'une après l'autre en revue.

* stratégie -
actuellement, les règles sont sélectionnées en fonction de la classe phonétique allouée au segment par le pré-traitement ; elles sont appliquées dans un ordre quelconque.
Seules les règles faisant intervenir le contexte droit d'un segment dans leurs prémisses nécessitent des retours-arrière. Pour résoudre ce problème avec efficacité, nous avons adopté une solution que nous présentons à l'aide d'un exemple qui donnera aussi une idée d'ensemble du fonctionnement du système qui est décrit de manière plus détaillée dans [6].

Soit une suite de quatre segments associés aux bases de faits suivantes :

S1 : noyau vocalique / F1 = 800 Hz / F2 = 1000 Hz
S2 : occlusive / fréquence_de_l'énergie_maximum_du_burst = 3000 Hz
S3 : noyau vocalique / F1 = 200 Hz / F2 = 1900 Hz
S4 : fricative / limite_inférieure_du_bruit ≫3000 Hz / limite_inférieure_du_bruit_descendante = vrai (de la droite vers la gauche).

L'analyse s'effectue comme suit :

. étape 1 : S1 est interprété comme /a/ grâce à la règle : R0 : si formant 2 - formant 1 < 250 Hz
alors (/a/)

. étape 2 : pour identifier S2, on a le choix entre les règles R1 et R2 :

R1 : si occlusive = vrai
et 3000 Hz < fréquence_de_l'énergie_maximum_du_burst < 4500 Hz
et contexte_droit (/y/, /u/, /w/)
alors (/t/)

R2 : si occlusive = vrai
et 2500 Hz < fréquence_de_l'énergie_maximum_du_burst < 3500 Hz
et contexte droit (/i/, /e/)
alors (/k/) .

Comme les informations dont on dispose sur S3 à ce stade de l'analyse ne permettent pas de lever l'ambiguïté, on génère, à partir de /a/, une arborescence à deux branches dont l'une représente les conclusions de R1, l'autre celles de R2 :

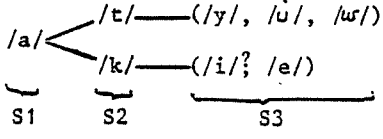


Figure 1

. étape 3 : la règle :

R3 : si 200 Hz < formant_1 < 300 Hz
et 1800 Hz < formant_2 < 2200 Hz
alors (/i/, /e/, /y/)

est la seule qui puisse s'appliquer ; elle ,e permet donc pas de lever l'ambiguïté.

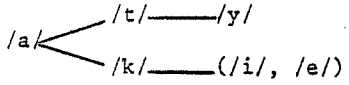


Figure 2

. étape 4 : on a le choix entre les règles R4 et R5 :

R4 : si fricative : vrai
et 3000 Hz <limite_inférieure_du_bruit < 4000 Hz
et contexte (/y/, /u/, /w/)
alors (/s/)

R5 : si fricative = vrai
et 2000 Hz < limite_inférieure_du_bruit < 3000 Hz
et contexte (/i/, /e/)
alors (/ʃ/)

mais aucune de ces deux règles ne permet de lever l'ambiguïté.

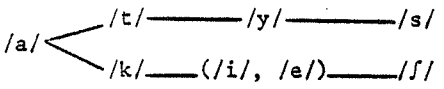


Figure 3

En revanche, si on dispose d'une règle qui précise que la limite inférieure du bruit d'une fricative est abaissée dans un contexte labial, on pourra supprimer la branche (/k/, /i/, /e/, /ʃ/) ; le treillis se réduira donc à :

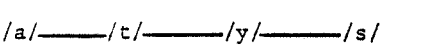


Figure 4

Quand la phase de validation des règles obtenues à partir de l'analyse de la compétence de l'expert sera terminée nous affinerons les stratégies du système, par exemple :

- en assouplissant les conditions d'application d'une règle (nombre et nature des prémisses dont la validation est indispensable, assouplissement des critères de validation d'une prémisse),
- en associant un score ou un coefficient de plausibilité aux conclusions d'une règle, prenant en compte le degré de validité des prémisses,
- en améliorant le processus de sélection des règles applicables à un segment.

MODELISATION DU DECODAGE PHONETIQUE PAR UN LANGAGE DEFINI SUR DES PROTOTYPES (FRAMES)

Reconnaissance de la parole et intelligence artificielle

L'implantation d'un système expert en reconnaissance de la parole pose des problèmes d'efficacité dans l'état actuel de la technique. En particulier les contraintes du fonctionnement en temps réel sont quasi-impossibles à respecter, à la fois pour des raisons de puissance intrinsèque des machines (nombre d'inférences logiques par seconde) et pour des raisons plus profondes concernant les principes mêmes de tels systèmes. Des études sont nécessaires sur ce dernier point : structures de contrôle adaptées à l'évolution temporelle de la base de faits, activation de connaissances déclaratives ou procédurales sur événement (démons), etc.

L'indéterminisme important qui caractérise la reconnaissance de la parole continue implique l'utilisation de raisonnements approximatifs faisant largement appel au contexte et fonctionnant à plusieurs niveaux, ainsi que la possibilité de suivre plusieurs lignes de raisonnement simultanément et de les abandonner si nécessaire.

Enfin, une liaison efficace est indispensable entre le système de raisonnement et le monde physique dans lequel sont prélevées les données. Les faits sur lesquels s'appuie le raisonnement sont déduits de ces données par des procédures éventuellement complexes qu'il est souhaitable d'attacher aux connaissances.

Tout cela conduit au développement de solutions en partie originales : structures de contrôle manipulant des arborescences de solutions, représentations objets. Un système à bases de connaissances fondé sur une représentation par prototype [10] permet d'utiliser, en ce qui concerne le décodage phonétique, différents types d'analyses acoustiques et de contrôler le processus de reconnaissance par un système de plans d'action. Il facilite le raisonnement à différents niveaux (i.e. la mise en oeuvre de connaissances et de méta-connaissances).

En complément à notre système expert, nous avons développé un processus de décodage phonétique qui utilise ce formalisme à deux fins :

- pour représenter les connaissances de l'expert et la détection des indices acoustiques ; les premières sont décrites par des prototypes, la seconde par des procédures de calcul "attachées" aux "cases" de prototypes ;

- pour formaliser les stratégies complexes de décodage mises en oeuvre par l'expert ; ces méta-connaissances sont regroupées dans une grammaire de prototypes constituée de règles de réécriture.

Le décodage phonétique est donc représenté par un langage défini sur des prototypes.

Des approches similaires ont été adoptées par De Mori [5] pour l'extraction de traits acoustiques et, dans un cadre plus général, par Green et Wood [7] pour le décodage phonétique.

Les résultats acquis au cours du développement du projet système expert évoqué dans les pages précédentes ont servi de base à l'élaboration du système que nous présentons brièvement dans les paragraphes suivants.

Structure et fonctionnement du système

Un prototype est un ensemble de "cases" (slot(s)) pouvant contenir des informations (slot-filler(s))

de nature différente. Nous représentons sous la forme d'un prototype l'ensemble des connaissances concernant un segment phonétique ; les différentes cases d'un prototype contiennent les indices qui caractérisent ce segment. Par exemple :

prototype /p/ (explosion de la consonne plosive p)
voisement : 0
durée : < 40 ms
maximum d'énergie dans (750-2500 Hz) : < 100

Après la phase de segmentation, l'identification d'un segment s'effectue de la manière suivante :

- d'abord, le prototype correspondant au segment est instancié à l'aide des procédures attachées aux différentes cases ;

- ensuite, le segment est identifié par comparaison du prototype instancié avec un sous-ensemble de prototypes de référence ; un score de compatibilité est calculé pour chaque prototype de référence comparé.

Extraction des indices, sélection des prototypes et comparaison (matching) sont contrôlés par les méta-connaissances inclues dans la grammaire de prototypes qui permet de mettre en oeuvre des raisonnements très élaborés. Par exemple, si aucun des scores de compatibilité n'est acceptable, le système pourra recalculer certains indices ou modifier le poids relatif des indices dans le calcul des scores de compatibilité (cf. la notion d'indice plus ou moins net, d'une part, plus ou moins significatif, d'autre part).

La grammaire de prototypes contient donc trois types de règles :

- des règles de construction d'instances de prototypes : extraction de paramètres acoustiques, sélection d'indices phonétiques pertinents,

- des règles de décision : par exemple, des règles de choix des prototypes de référence qui interviendrons dans la phase d'identification du segment (matching),

- des méta-règles définissant la stratégie globale de décodage phonétique.

L'intérêt essentiel de cette modélisation, qui met en oeuvre une méthode ascendante d'analyse, approche classique en décodage acoustico-phonétique, réside dans le fait qu'elle autorise les interactions indispensables entre les processus de segmentation et d'identification phonétique, grâce aux relations qui existent entre les notions de règle, de prototype et de procédure.

Plus généralement, la mise en oeuvre de langages définis sur des prototypes permet une imbrication étroite entre deux modes de représentation : règles et objets.

Nous poursuivons actuellement nos recherches dans deux directions parallèles :

- enrichissement de la base de connaissances acoustico-phonétiques à partir de la compétence de l'expert (informations utilisées, démarche), dans le cadre du développement de notre projet système expert,

- développement de structure de contrôle élaborées, en particulier amélioration de l'interaction segmentation-identification, en vue de l'implantation d'un système de décodage phonétique efficace, fondé sur une représentation des connaissances nécessaires à l'aide de prototypes et d'une grammaire définie sur ces objets.

REFERENCES

[1] N. Carbonell, J.P. Haton, F. Lonchamp, J.M. Pierrel, "Elaboration d'un système expert pour le décodage phonétique automatique de la parole", Speech Communication, 2, N° 2-3, p. 231-233, 1983.

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

[3] N. Carbonell, M.O. Cordier, D. Fohr, J.P. Haton F. Lonchamp, J.M. Pierrel, "Acquisition et formalisation du raisonnement dans un système expert de lecture de spectrogrammes vocaux", Colloque ARC, Orsay 1984.

[4] P. Combescure, "Vingt listes de dix phrases phonétiquement équilibrées", Revue d'Acoustique, 14, N° 56, 1981.

[5] R. De Mori, "Extraction of Acoustic Cues Using a Grammar of Frames", Speech Communication, N° 2-3, 1983.

[6] D. Fohr, N. Carbonell, J.P. Haton, "SYSTEXP : un système expert pour le décodage acoustico-phonétique", 5èmes journées internationales sur les systèmes experts et leurs applications, Avignon, mai 1985.

[7] P.D. Green and A.R. Wood, "Knowledge-Based Speech Understanding : Towards a Representation Approach", Proc. ECAI, 1984.

[8] J.P. Haton, J.P. Damestoy, "A Frame Language for the Control of Phonetic Decoding in Continuous Speech Recognition", IEEE ICASSP-85, Tampa, Florida.

[9] J.P. Haton, "Techniques d'intelligence artificielle en compréhension de la parole et en vision : état des recherches", à paraître dans T.S.I.

[10] M. Minsky, "A Framework for Representing Knowledge", in P. Winston ed. "The Psychology of Computer Vision", Mc Graw-Hill, 1975.

ETUDE DES VARIATIONS ALLOPHONIQUES DE LA VOYELLE /a/ ET SES CONSEQUENCES POUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Jacqueline Vaissière

Dpt RCP, CNET, Lannion et Speech Group, MIT, Cambridge

RESUME

L'effet du contexte sur 708 /a/ segmentés automatiquement dans des phrases prononcées par 4 locuteurs a été etudie. Il a été trouvé que le début de /a/ est influencé par l'identité de la consonne précédente et par des phénomèmes de "carry-over" avant la voyelle précédente ou d'anticipation avec la voyelle suivante. Le centre de /a/ est influencé par sa durée, par la place d'articulation des consonnes environnantes, et par la hauteur de la voyelle suivante. La fin de /a/ est sous la dépendance de la consonne et de la voyelle suivantes. La construction de contextes maximilisant les influences, soit ouvrantes, soit fermantes, sur la voyelle /a/ est utile dans la conception de base de données représentatives et elle permet une réduction importante de la taille des corpus d'entrainement pour l'adaptation des systèmes de reconnaissance de parole continue à un nouvel locuteur.

INTRODUCTION

L'objet de cette communication est la presentation des résultats concernant une étude sur les variations dans la distance entre les deux premiers formants, F1 et F2, dans la voyelle /a/, en francais, dans de la parole continue. Le but ultime de cette recherche est de mieux comprendre les facteurs de variabilité dans la parole afin d'introduire cette connaissance dans les systèmes de reconnaissance automatiques. La voyelle /a/ a ete choisie comme première voyelle à etudier, car c'est la voyelle la plus fréquente en Francais, et sans doute celle dont les réalisations acoustiques sont les plus variées.

CORPUS ET PROCEDURE

708 /a/ ont été extraits de 344 phrases prononcées par 4 locuteurs: un nombre total de 304 phrases lues et 40 phrases semi-spontanées (10 phrases par locuteur, simulant des questions sur les Pages Roses de l'annuaire). Les données ont ete enregistrées au CNET, à Lannion, en utilisant le protocole GRECO, et numerisées à 12.8 kHz. Les phrases ont été alors analysées avec l'aide de Spire, un ensemble intégré de programmes d'analyse et de reconnaissance sur machine LISP, à MIT (Zue et Cyphers, 1985). La construction d'algorithmes directement utilisables par un système reconnaissance automatique étant notre but, nous avons preferé segmenter automatiquement (et non manuellement) les phrases. Nous avons donc utilisé un programme de segmentation implémenté sur Spire, initialement concu pour la langue américaine (en cours de finition). 89% des 795 /a/ contenus dans les 344 phrases ont été segmentés correctement, et selectionnés pour notre etude. Nous référons le lecteur à la these récente de Leung (Leung, 1985) pour de plus amples détails sur ce programme.

La valeur des pics LPC toutes les 5 msec a été calculée (voir Fig.1). Nous avons mesuré la valeur du premier pic (F1) et du deuxième pic (F2) au début, au milieu et à la fin de chaque voyelle. Une étude préliminaire ayant permis de démontrer que les variations de F1 et F2 etaient corrélées (plus F1 est élevé, plus F2 est bas), et que la distance entre F1 et F2 était une bonne mesure des variations subies par /a/ dans différents contextes, nous avons retenu la difference (F2 - F1) pour la suite.

L'ensemble de cette étude a montre que la distance entre F1 et F2 à différents points de la voyelle est fonction de la durée de la voyelle, de l'identité des consonnes environnantes, et de la hauteur des voyelles environnantes, c'est-a-dire que son interpretation fine nécessite une référence aux deux phonèmes qui précedent /a/ et aux deux phonèmes qui la suivent (la pause peut être considere comme un phonème). Nous illustrerons dans la presente communication que quelques unes de ces influences (par manque de place): la durée de la voyelle et la place d'articulation des consonnes environnantes sur le centre de la voyelle, d'une part, et la place et la manière d'articulation de la consonne précédente sur le début de la voyelle, d'autre part. Ensuite, nous résumerons les autres influences et discuterons des conséquences de ces résultats pour la reconnaissance automatique de la parole.

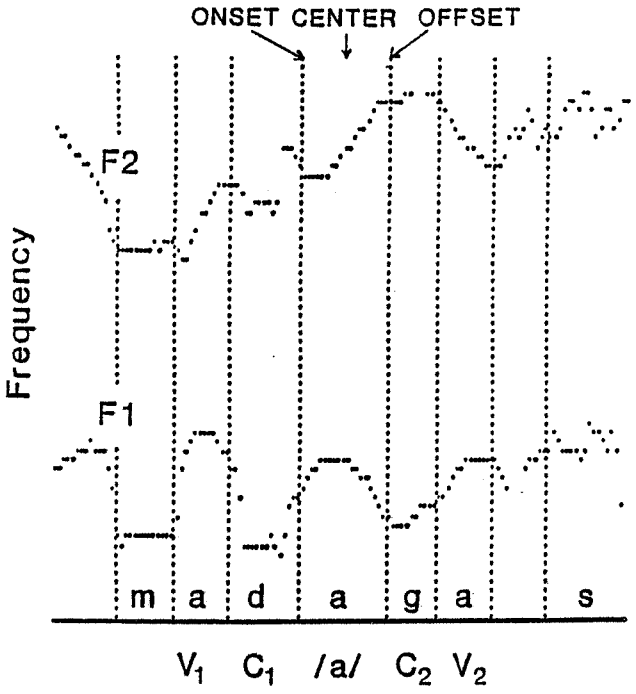


Fig 1: Visualisation des résultats de segmentation automatique (ligne verticale) et du suivi des deux premiers pics LPC (points), toutes les 5 msec, dans la séquence /adaga/, extrait du mot "Madagascar". Les flèches indiquent les trois positions sélectionnées pour l'étude: le début, le centre, et la fin de la voyelle.

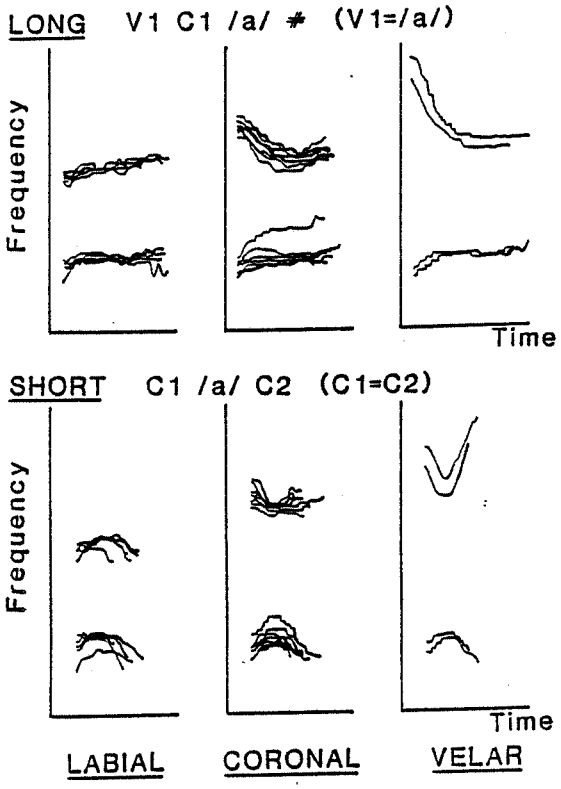


Figure 2: Superposition des tracés automatiques de F1 et F2 durant la voyelle /a/, pour un locuteur.

RESULTATS

1) INFLUENCE DE LA DUREE ET DE L'IDENTITE DES CONSONNES ENVIRONNANTES SUR LE CENTRE LA VOYELLE:

La figure 2 illustre la superposition du tracé de F1 et F2 dans un certain nombre de voyelles /a/, en fonction du temps, pour un locuteur. Les voyelles ont été séparées en six groupes en fonction de leur durée, et de la place d'articulation de la consonne précédente. Les traces du haut de la figure illustre des occurences longues de la voyelle /a/. L'allongement est un allongement prépausal (avant une pause, notée #). Les tracés du bas correspondent à des versions plus courtes de la voyelle. La première colonne correspond au cas ou la voyelle est précédée par une consonne labiale ou labio-dentale (/p,b,m,f,v/). Celle du milieu et celle de droite regroupent les voyelles /a/ précédée d'une consonne coronale (/t,d,n,s,z,ch,j,l/) ou par une velaire (/k,g/), respectivement.

Plusieurs observations peuvent être faites à partir de cette figure:

i) il existe une nette tendance pour les valeurs de F1 et F2 (et pour la différence (F2 - F1)) de se diriger de facon asymptotique vers les mêmes valeurs cibles. Ces valeurs cibles sont indépendantes du contexte consonantique et de la durée de la voyelle, et peuvent donc en consequence être considérées comme des attributs invariants de la voyelle. Au centre, la différence (F2 -F1) est invariante, bien que le début de la voyelle subisse d'amples variations en fonction du contexte. Cette première observation est en accord les résultats de Lindblom (Lindblom, 1963) dans son étude spectographique sur la réduction des voyelles.

ii) la seconde observation concerne les voyelles de plus courte durée. Dans ce cas, l'influence consonantique est claire. La distance entre F1 et F2 augmente de facon dépendente avec la place d'articulation de la consonne, des labiales aux velaires.

L'amplitude de la réduction de la voyelle /a/ est donc fonction de sa durée, et de la place d'articulation des consonnes environnantes. De tels resultats concordent avec les travaux de House et Stevens, publiés en 1963, sur les perturbations subies par les voyelles à cause de leur environnement consonantique. Les deux auteurs concluaient que le degré avec lequel la configuration idéale pour une voyelle est atteinte dépend de la distance que les articulateurs doivent

parcourir pendant la syllable, entre la consonne initiale, le nucleus vocalique et la consonne finale (House et Stevens, 1963).

2) INFLUENCE DE LA PLACE ET MANIERE D'ARTICULATION DE LA CONSONNE SUR LE DEBUT DE LA VOYELLE /a/ SUIVANTE:

La figure 3 représente les valeurs extrèmes (maximales et minimales) trouvées pour (F2 -F1), au debut de la voyelle, pour les 4 locuteurs, en fonction de l'identité de la consonne précédente. On peut observer sur cette figure que le paramètre le plus important, est ,comme attendu, la PLACE d'articulation de la consonne precédente (cf la notion de locus). La différence est plus petite lorsque la consonne est labiale, et plus grande lorsque la consonne est velaire. La MANIERE d'articulation joue un rôle non négligable: pour une place d'articulation donnée, la différence est plus petite dans le cas des occlusives et plus grande dans le cas des nasales. On peut remarquer que le voisement excerce également une influence: la différence est plus petite dans le cas des non voisees.

Il n'est pas possible de décrire, une par une, toutes les influences observées. Les influences relevées sont les suivantes:

1) le début de la voyelle est influencé par l'identite de la consonne précédente, et la place, la manière et le voisement jouent un rôle, dans cet ordre. Il est aussi influencé soit par un effet d'anticipation vocalique venant de la voyelle suivante soit par un effet de "carry-over" de la voyelle précédente.

2) au centre de la voyelle, la durée joue un rôle primordial. La place d'articulation des consonnes environnantes et la hauteur de la voyelle suivante doivent également etre prises en compte.

3) à la fin de la voyelle, quand la voyelle n'est pas suivie par une pause longue, la place d'articulation de la consonne suivante et la hauteur de la voyelle suivante jouent un rôle dominant.

La position de la langue aura donc tendance a anticiper, pendant la voyelle /a/, la position requise pour la voyelle suivante. Un /a/ localisé avant un /i/ sera plus fermé (F1 plus bas et F2 plus élevé), ceteris paribus, qu'un /a/ précédant un autre /a/, et cet effet, qui est quantitativement le plus important en fin de voyelle et au centre de la voyelle (dans cet ordre), peut également se faire sentir au début de la voyelle. L'importante de la hauteur de la voyelle suivante, V2, sur les transitions entre V1 et C avait déjà été notée par Ohman, dans son étude sur les phénomènes de coarticulation dans les séquences V1CV2 (Ohman, 1966).

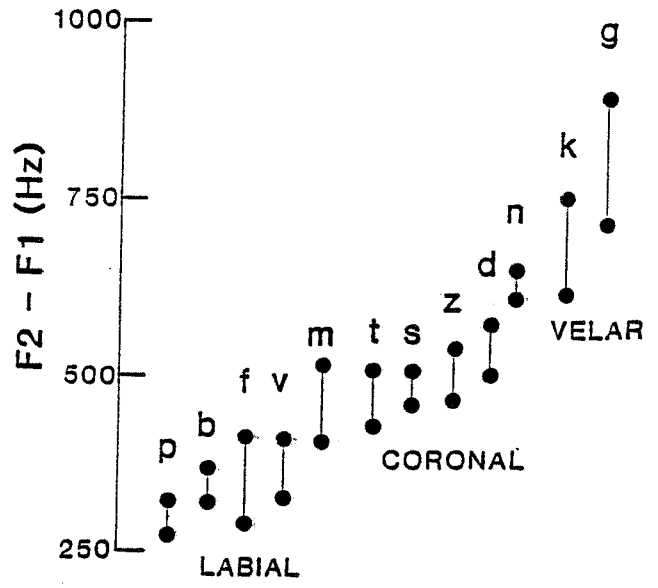


Figure 3: Valeur maximale et minimale trouvée pour la moyenne des distances (F2 - F1) au début de la voyelle, pour tous les locuteurs. (chaque moyenne pour chaque locuteur a été calculée sur 4 occurences de /a/).

CONSEQUENCES

Grâce à la connaissance des divers facteurs influencant la voyelle /a/, il est possible de creer des contextes pour maximaliser ou minimaliser les effets afin de couvrir par un nombre réduit d'occurences l'evantail possible des variations.

Donnons deux exemples.

1) Au centre de la voyelle, la distance sera minimale ou proche du minimum si la voyelle /a/ est située dans un contexte labial (/p/ étant le meilleur choix, voir Fig.3), précédee par une autre voyelle ouverte (/a/ par exemple), et suivie soit par une pause, soit par la syllable ouvrante /pa/. Le second /a/ dans la sequence "Papa part..." est un choix excellent. La même distance sera maximale si la même voyelle est placée après une vélaire (/g/ est le meilleur choix, voir Figure 3), précédée et suivie d'une voyelle fermée (la semi-voyelle J étant le meilleur choix), avec une consonne coronale entre les deux.

2) Au début d'une voyelle précedée par une labiale, la distance doit être mininum si cette labiale est /p/, précedée de la voyelle /a/ et suivie par une labiale, elle même suivie d'une voyelle ouverte. Le deuxième /a/ dans la séquence "Papa part" est un bon choix. La distance sera maximale si la consonne est nasale (/m/), dans un contexte de voyelles hautes (dans "dix magnifiques..." par exemple).

Cette connaissance des effets permet, grâce a un nombre reduit de /a/ de trouver la plage possible de variations. On peut donc l'utiliser efficacement soit pour la construction de bases de donnees, soit pour réduire de facon importante la corpus d'entrainement necessaire pour adapter un système de reconnaissance à un nouvel locuteur, soit encore pour introduire des contraintes pour la vérification des hypothèses lexicales. (il n'y a pas d'intersection, par exemple, entre les valeurs possibles de (F2 - F1) entre /p/, /t/, ou /k/, ou entre /l/ ou /r/).

La figure 4 illustre un exemple. Les cercles representent les valeurs minimales et maximales de (F2 - F1) au début de la voyelle trouvées pour chacun de nos 4 locuteurs dans les 344 phrases, en fonction de la place d'articulation de la consonne precedente. Les pointillés indiquent les mêmes valeurs trouvées sur 6 /a/ extraits du vaste corpus, les /a/ etant choisis en fonction de leur contexte. On peut voir que dans 16 cas sur 24, les valeurs extrèmes ont été trouvées, bien que les contextes ne soient pas idéaux du point de vue theorique.

DISCUSSION

La plage importante de variations des valeurs attestées pour les deux premiers formants de la voyelle /a/ en parole continue pose des problèmes interessants concernant l'identification de cette voyelle en Francais. L'auditeur utilise-t-il des informations venant du contexte de la voyelle /a/ pour "corriger" les valeurs trouvées dans la partie la plus stable de la voyelle /a/ elle même, ou alors ces variations, bien que quantitativement importantes, ne sont elles pas suffisantes pour provoquer une confusion éventuelle de /a/ avec celles des voyelles environnantes (o ouvert et e ouvert en Francais). Nous avons entrepris une étude systematique des voyelles proches de /a/ chez les mêmes locuteurs, en parole continue pour repondre à cette question.

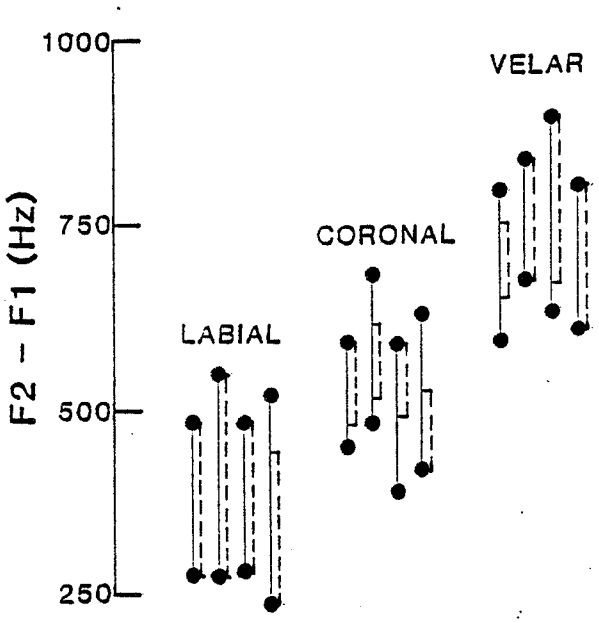


Figure 4: Valeurs maximales et minimales attestés dans le corpus pour la distance (F2 - F1) au début de la voyelle, pour chacun des 4 locuteurs, lorsque la voyelle est précédée par une labiale (à droite), une coronale (au milieu) et une velaire (à droite): les ronds représentent les valeurs obtenues sur les 344 phrases (minimum de 42 phrases par locuteur), et les pointillés représentent la plage de variations prédites par des calculs faits sur 6 /a/ selectionnés (dans les mêmes phrases pour chaque locuteur).

REMERCIEMENTS

Ce travail a ete réalisé à MIT, avec le support partiel du contrat N00014-82-K-0727. L'auteur remercie l'administration du CNET et Professeur Victor Zue pour avoir rendu possible un séjour dans le Speech Group de MIT.

REFERENCES

House, A.S. and Stevens, K. N., (1963), "Perturbations of vowel articulations by consonantal contexts: an acoustical study", JSHR, 6, 11-128.

Leung, H. C., (1985), "A procedure for automatic alignement of phonetic transcriptions with continuous speech", Master of Sciences, Massachussetts Institute of Technology.

Lindblom, B., (1963), "Spectographic study of vowel reduction", JASA, 35,11,1773-1781.

Ohman, S. E. G., (1966), "Coarticulation in VCV utterances: spectoagraphic measurements", JASA, 39, 1, 151-168.

Zue, v. W., (1985), "The MIT Spire system", Proceedings of the 1985 Speech-Tech Conf., New York.

CLASSIFICATION DE PHONEMES PAR METHODES NON-PARAMETRIQUES

S. Soudoplatoff

Centre Scientifique IBM
36, Av Raymond Poincare 75116 Paris

About what is going to be said

It is common in speech recognition, or in speech transmission, to map the acoustic vectors, which are usually in a high dimensional space (typically a few tens) into a small size vocabulary, (typically a few hundreds), procedure known as vectorial quantification. It is also common, in phonetic recognition, to output a phonetic lattice without any other information than the phonemes.

This paper shows that an improvement can be done by using a probabilistic vector giving the probability of an acoustic vector given a class, instead of the class itself, using for this purpose simple non-parametric density-estimate.

De l'introduction des probabilites en reconnaissance phonetique

Pourquoi ?

La reconnaissance phonétique, bien que n'étant souvent qu'un module parmis d'autres dans le graphe explicatif d'un système de reconnaissance de la parole, n'en représente pas moins un élément encore tres mal maitrisé. Le produit usuel d'un tel module est un treillis phonétique, dont la seule information supplémentaire aux phonèmes eux-même est leur ordre.

Dans le même ordre d'idée, un quantificateur vectoriel consiste à transformer un vecteur dans un espace de dimension élévéede l'ordre de quelques dizaines, en un élément d'un vocabulaire de taille limitée, de l'ordre de quelques centaines. Bien souvent, seule l'information entre le vecteur et le centre de classe est comprimée, lorsqu'elle n'est pas purement mise à la poubelle.

Ces constatations etant faites, il ne reste plus qu'à ameliorer le système en remplaçant le dictionnaire par un espace de vecteurs de probabilités donnant, pour chaque point x, des probabilités relatives à chacun des éléments du dictionnaire. Ainsi, le treillis phonétique donnerait non pas les phonemes, mais leur probabilité.

Exemple : affectation Bayesienne

Une méthode de reconnaissance phonétique est la classification de type Bayesien, en maximisant le produit $P(\phi) \times P(x \mid \phi)$, ϕ etant le phonème candidat, et x un vecteur acoustique à classifier. L'estimation du terme $P(\phi)$ peut se faire par des méthodes statistiques classiques, et ne rentre pas dans le cadre de cet exposé. Le second terme est typiquement une probabilité que l'on doit estimer.

De l'estimation de ces probabilites

Du classissisme

Parmi toutes les techniques d'estimation de densité, nous avons rejeté les méthodes parametriques, bien que la Gaussienne soit un objet assez repandu, malgre le calcul que cela impose. En effet, des visualisations d'analyses en composantes principales nous ont montréque les nuages ne présentaient pas de forme suffisamment regulières pour pouvoir faire des hypothèses simples de fonction de densité. Ces analyses montrent clairement, dans le cas de coéfficients d'auto-correlation, que les vecteurs sont "collés" contre une limite mathematique, que l'on montre facilement être une parabole (Ceci se montre facilement en considérant l'aspect defini positif de la matrice de Toeplitz associée, et ne rentre pas dans le cadre de cette etude).

Du noyautage

Nous nous sommes alors penchés sur les méthodes non-paramétriques d'estimation de densité. Ce problême a vu une approche intéressante dans le cas mono-dimensionnel, avec les noyaux de Parzen [1], qui ont été étendus facilement au cas multi-dimensionnel [2].

Rappelons la forme d'un estimateur de Parzen. Nous connaissons n tirages d'une variable aleatoire, $X_1, \ldots, X_n$, de loi de probabilité f(x), d'un espace E^p de dimension p. L'estimation se fait à l'aide d'une fonction K, de E^p dans R, par

$$f_r(x) = \frac{1}{n\, r^p(n)} \sum_{j=1}^{n} K\left(\frac{x - X_j}{r(n)}\right).$$

Pour que cet estimateur soit sans biais, il faut que

$$\int |K(y)|\, dy = 1 .$$

A l'évidence, l'éstimation de la probabilité en un point x par un comptage des points du nuage contenus dans une boule (nous supposons l'espace muni d'une metrique convenable) de rayon r, centrée autour du point x, est un cas particulier des estimateurs de Parzen. Il suffit en effet de choisir une fonction K constante à l'interieur de la boule unité centrée a l'origine, et nulle en dehors. Pour que la condition (2) soit verifiee, il faut que la constante soit égale à l'inverse du volume de la boule unité.

Du rayonnage

A ce stade, nous sommes confrontés au probleme du choix du rayon d'agrégation r. En effet, on perçoit bien qu'un rayon trop petit conduirait à des probabilités partout nulles, sauf aux points du nuage, et qu'un rayon trop grand amenerait des probabilités toutes égales.

Cette question a été abordée dans les deux articles cités, mais au prix d'une part d'une formulation mathematique complexe, d'autre part d'une hypothese a priori sur la loi de probabilité du nuage de points.

Notre propos est d'essayer d'extraire un optimum par la seule analyse du nuage de points, sans effectuer aucune hypothèse sur la distribution sous-jacente. A cette etape, nous raisonnerons au niveau de chaque nuage de phonemes.

De la hierarchie

Lorsqu'il s'agit de trouver un rayon optimum, a partir du nuage de points, il n'est pas interdit de penser à une classification hiérarchique. L'analogie entre les bornes extrêmes de r, decrites

plus haut, et qui conduisent à des probabilites égales, et aux limites d'une classification hiérarchique, qui sont une classe avec tout, et N classes à un element, a seule guidé notre demarche.

Pour poursuivre l'analogie, il fallait nous fixer sur des méthodes de densité. Nous en avons selectionné deux : la methode de percolation de Tremolieres [3] et la methode pseudo-hiérarchique de J.F. Marcotorchino [4].Sans rentrer dans les details, nous avons pu, en faisant varier le rayon d'agregation, trouver un parametre dont le maximum nous a fourni dans tous les cas, a un rayon pas trop idiot. Pour plus de details, voir [5].

De l'application

Il est toujours possible, lorsqu'on possède un vecteur de probabilité, de revenir à un etiquetage classique en affectant le point a la classe qui maximise $P(x \mid C_k)$. Ce test a ete effectué au sein du groupe des consonnes P, T et K. Ce test ne permet que de verifier la consistance des estimateurs de Parzen, car en perdant l'information des probabilités, on perd la saveur du modèle. Il a tout de meme permis de recuperer un certain nombre d'erreurs commises par une methode d'affectation au centre de classe le plus proche.

Un certain nombre de problèmes demeurent cependant, le plus serieux etant que certains points ont des probabilités partout nulles. Suivant les choix de rayons, nous avons trouvé entre 0.1% et 10% des points dans ce cas. La seule solution consiste à créer une classe supplementaire, qui contient tout ces points, avec des probabilites nulles pour toutes les autres classes.

Ou l'on bouquine

[1] E. Parzen, "On estimation of a probability density function and mode", Ann. Math. Stat., vol 33, 1962.

[2] T. Cacoullos, "Estimation of a multivariate density", Ann. Inst. Stat. Math., vol 18, 1966.

[3] Chandon et Pinson, "Analyse typologique, theorie et applications", Masson, 1981.

[4] J.F. Marcotorchino, "Une methode pseudo-hierarchique de classification", Non publie.

[5] S. Soudoplatoff, "Classification de phonemes par estimation de densite", Publication du Centre scientifique IBM de Paris, Aout 1984.

UN SYSTEME DE COMPREHENSION DE LA PAROLE CONTINUE SUR MICROPROCESSEUR.

A. Veloz Guerrero, J. J. Mariani.

LIMSI, CNRS. B. P. 30, 91406 ORSAY, CEDEX.

ABSTRACT

We describe a speech understanding system being implemented on a microprocessor. The system is designed to accept continuous speech from one speaker and to work within the context of a limited task situation and small vocabularies. The system utilizes phonetic recognition at the phonetic level and an optimal one-pass dynamic programming algorithm at the lexical and syntactic levels. The system has an interactive program for the definition of grammars for a given specific task language and a program of orthographic-phonetic translation that takes into account the phonological variations of words.

1. INTRODUCTION.

La communication homme-machine sur support vocal est un domaine de recherche qui a déjà donné des résultats sur le plan pratique. En effet, actuellement il existe sur le marché, des systèmes de reconnaissance de la parole par mots isolés et par mots enchainés, implantés sur une carte à base de microprocesseur.

En général tous ces systèmes se fondent sur les méthodes de reconnaissance de type global. Les paramètres du signal utilisés pour la reconnaissance sont issus d'une analyse spectrale par un banc de filtres, d'une analyse LPC ou d'une analyse cepstrale. L'algorithme de reconnaissance est basé sur une méthode de programmation dynamique [4].

Le système que nous allons décrire a ses fondements dans les travaux de recherche sur les systèmes de compréhension de la parole continue menés au LIMSI par J.J. MARIANI.

Le système utilise dans un premier temps, la méthode de reconnaissance analytique ou phonétique, faisant une transformation du signal de parole du domaine acoustique (sonagramme numérique) au domaine phonétique (treillis phonétique constitué de symboles phonétiques accompagnés de leurs notes de ressemblance) [1]. Ensuite dans un deuxième temps, le système utilise la méthode de reconnaissance globale appliqué au treillis phonétique au moyen d'un algorithme de reconnaissance de mots enchainés à une seule passe [2],[3].

2. DESCRIPTION DU SYSTEME.

La figure 1 montre le diagramme du système de reconnaissance.

2.1 Niveau acoustique.

Le signal de parole est analysé par un banc de 8 filtres numériques recursifs couvrant la bande de 100 à 5000Hz. Les fréquences centrales des canaux sont réparties suivant l'échelle Bark. Le banc de filtres est scruté toutes les 10 millisecondes pour constituer le sonagramme numérique. La sortie de chaque canal est codée selon une fonction logarithmique réalisée par approximation par segments. Après cette compression logarithmique, la segmentation du signal est réalisée en calculant la courbe de stabilité spectrale qui donne la variation relative entre deux spectres distants d'un retard T.

$$VAR(t,T/2) = \sum_{v} |S(t,v) - S(t-T,v)|$$

Où T est le retard ayant comme valeur 40 millisecondes et S(t,v) est la valeur du spectre à l'instant t dans le canal v. Les segments sont determinés par les extrêma de cette courbe. Les maxima correspondent aux frontières des segments et les minima correspondent à une partie très stable de la réalisation du phonème.

Les taux d'erreur de segmentation phonétique obtenus sont de l'ordre de 15%

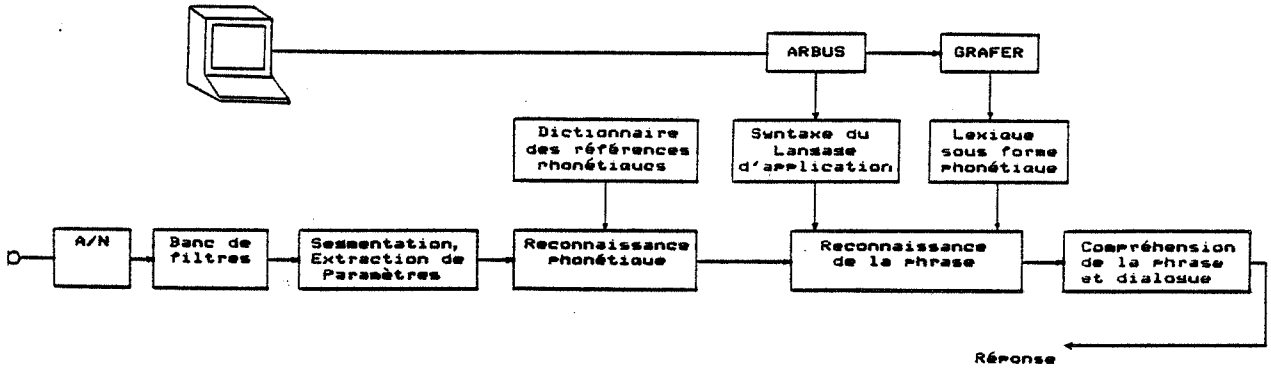


Figure 1. Diagramme du système.

à 20 % de segmentation excédentaire (deux segments succesifs correspondent à un même phonème), et de l'ordre de 5 % de segmentation déficitaire (une segmentation n'apparaît pas et donc un même segment correspond à deux phonèmes).

Après avoir trouvé un minimum de la courbe de stabilité spectrale, on calcule les pentes entre les canaux du spectre correspondant à ce minimum, et on les retient comme des traits pertinents pour la reconnaissance phonétique.

2.2 Niveau Phonétique.

La constitution du dictionnaire de références phonétiques se fait à partir de phrases sans signification telles que "APATAKA" [1]. La première référence préexistante est le spectre de silence "$" pour lequel tous les canaux correspondent au bruit de fond supposé identique pendant l'enoncé de la phrase.

La phrase est segmentée et on obtient le treillis phonétique par rapport à la seule référence de silence "$", ce qui permet d'avoir une différence très grande entre les zones de silence des plosives /p/, /t/ et /k/ (où le silence obtient une bonne note de reconnaissance) et les zones de forte énergie correspondant aux voyelles /a/. Les spectres représentants des phonèmes sont parametrés selon les traits pertinents mentionnés au dessus, ils sont étiquetés et ensuite stockés dans le dictionnaire de références phonétiques pour son utilisation ultérieure pendant la reconnaissance phonétique.

2.3 Reconnaissance phonétique.

La reconnaissance phonétique s'effectue en comparant le spectre S correspondant à la partie la plus stable du phonème, avec tous les spectres références Sri en utilisant la distance de Minkovski du premier ordre sur les pentes:

$$D(S,Sri)=\sum_{v}\left|S(v+1)-S(v)-(Sri(v+1)-Sri(v))\right|$$

Cette distance est calculée pour chacune des références phonétiques Sri.

Les références ainsi reconnues sont ensuite classées selon leur distance et on obtient le treillis phonétique correspondant. Le treillis phonétique est constitué de tous les phonèmes classés, accompagnés de leurs notes de reconnaissance. Le taux de reconnaissance phonétique se situe à environ 45 % de phonèmes correctement reconnus en première position, et à environ 80 % de phonèmes dans les cinq premières positions du treillis phonétique. La figure 2 montre un exemple de treillis phonétique donné par le système.

Toutes les opérations de soustraction du bruit de fond, compression logarithmique, segmentation et obtention du treillis phonétique se réalisent au fur et à mesure de la prononciation de la phrase. Pour la reconnaissance de la phrase prononcée, les notes du treillis phonétique sont stockées dans un "buffer" circulaire. Ainsi, le contenu du "buffer" est examiné et traité par le module de reconnaissance de la phrase dès qu'il y a des telles notes dans le "buffer". Le système de reconnaissance a été implanté sur une carte prototype, en utilisant un microprocesseur 16 bits INTEL 80186.

Les phrases reconnues accompagnées de leurs notes de reconnaissance seront proposées au module de dialogue qui devra assurer l'interface avec l'application particulière donnée.

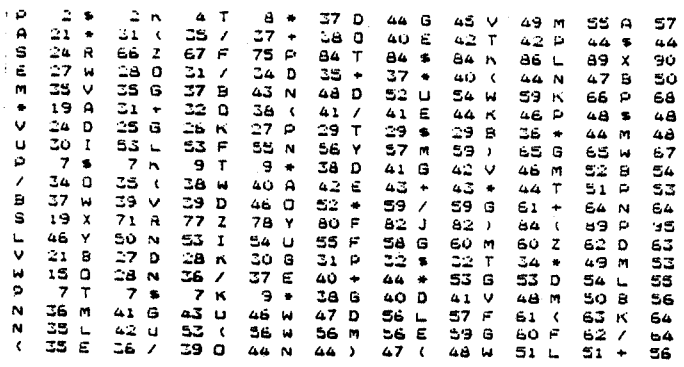


Figure 2. Exemple de Treillis phonétique. Phrase prononcée: "Passez-moi Dupont s'il vous plaît.

2.4 Niveaux Lexical et Syntaxique. Reconnaissance de la phrase prononcée.

Il existe plusieurs stratégies pour reconnaître la phrase prononcée à partir du treillis phonétique, par exemple: la stratégie par îlots de confiance, la stratégie descendante de prédiction-vérification avec conservation des meilleures solutions en parallèle et analyse de gauche à droite sans retour arrière, etc. [1],[7].

Pour notre part, étant donné que le langage d'application est représenté par une grammaire régulière (c'est-à-dire par un automate fini dont chaque transition entre états, porte les symboles des mots), la résolution du problème de la reconnaissance de la phrase prononcée consiste à trouver un chemin optimal entre l'état initial et un état final de l'automate fini. Le chemin optimal est déterminé par un algorithme qui, en effet, explore tous les chemins possibles en parallèle d'une manière très efficace [2],[3]. Cet algorithme, à la différence de l'algorithme à deux niveaux de Sakoe [5], est un algorithme à une seule passe. Il permet le traitement et la sortie des premiers mots de la phrase prononcée avant que celle-ci ne soit terminée.

Le dictionnaire de références lexicales V={R1,R2,...,Rn} est constitué des traductions phonétiques comprennant les variantes phonologiques et la liaison possible après les mots [8]. La séquence à reconnaître T est immergée dans le treillis phonétique issu de l'étape de reconnaissance phonétique. On peut comparer la séquence à reconnaître T à des références composites (appelées super-références) constituées par concaténation de références Rr. Le problème consiste à évaluer K, le nombre de mots de la séquence T, et les indices m1,m2,...mK (1<=mi<=n) qui minimisent la dissemblance entre T et les super-références D(T,Rm1 ⊕ Rm2 ⊕, ..., ⊕ RmK) (le signe ⊕ représente l'opération de concaténation).

L'équation locale de programmation dynamique non symétrique employée est de la forme (en suivant la notation de [9]):

$$gr(i,j)=\min\{gr(i-1,j-1)+d(i,j),\ gr((i-2,j-1)+dr(i-1,j)+dr(i,j),\ gr(i-1,j-2)+(dr(i,j-1)+dr(i,j))/2\} \quad (1)$$

et l'équation de passage de la fin d'une référence au début d'une autre est:

$$gr(i,1) = \min_{r\in P(r)}\{gr(i-1,Jr)+dr(i,1)\} \quad (2)$$

où dr(i,j) et gr(i,j) sont les notes de reconnaissance de phonèmes candidats (présents dans le treillis phonétique) et les distances cummulées pour le i-ème phonème de la référence Rr, à l'instant j. P(r) est l'ensemble de mots qui peuvent précéder le mot r, d'après la syntaxe. Jr est la longueur en caractères phonétiques de la référence r. Les distances gr(i,j) sont évaluées sur tout le domaine des points (r,i,j) qui est parcouru colonne par colonne, en employant des techniques de "pruning" pour réduire le nombre de calculs. On obtient finalement:

$$D(T,Rm1\oplus Rm2\oplus,\ldots,\oplus RmK)=\min_{r\in F}\{gr(I,Jr)/I\}$$

où F est l'ensemble d'états terminaux de l'automate fini et I est la longueur de la séquence à reconnaître.

Pour obtenir les valeurs de K,m1,m2,...,mK, on évalue br(i,j) qui est la position d'une frontière des mots dans T si le point (r,i,j) appartient au chemin optimal, de la manière suivante:

$$br(i,j) = \text{Tracedemin}\begin{Bmatrix} br(i-1,j-2) \\ br(i-1,j-1) \\ br(i-2,j-1) \end{Bmatrix} \quad (3)$$

$$br(i,1)=i-1 \quad (4)$$

où la fonction Tracedemin retourne l'argument correspondant au choix effectué par la fonction de minimisation (1) au même point (r,i,j).

3. DEFINITION DU LANGAGE D'APPLICATION.

Le système comporte un programme intéractif d'aide à la définition de grammaires pour un langage d'application donné (ARBUS) [10]. Ce programme est associé à un programme de phonétisation automatique du lexique (GRAFER) qui donne un ensemble de traductions phonétiques possibles en fonction des variantes phonologiques des mots. L'utilisateeur peut donc, créer une syntaxe, la modifier, remplacer des éléments terminaux, sauvegarder la langage construit, etc. La représentation de la syntaxe sous forme d'automate fini, ainsi que le vocabulaire phonétisé sont fournis à l'algorithme de reconnaissance de la phrase pour son utilisation ultérieure durant le processus de reconnaissance.

4. CONCLUSIONS ET PERSPECTIVES.

Le programmme de reconnaissance de la phrase à été implanté sur una carte prototype basée sur le microprocesseur 16 bits INTEL 80186. L'architecture du programme est modulaire. Les modules sont écrits pour la plupart en PLM/86, sauf ceux qui requièrent une vitesse d'exécution plus rapide qui sont écrits en assembleur.

Le langage d'application sur lequel on a testé le système est un langage de standard téléphonique. La taille du vocabulaire de ce langage est de 40 mots. Le facteur de branchement statique est de 10 mots. Les conditions du test ont été les suivantes: salle d'ordinateur, microphone de casque, un locuteur. Le corpus de test de ce langage est constitué de 20 phrases [1]. Les résultats sont comparables à ceux obtenus auparavant sur le système simulé. Par exemple, les taux moyens d'erreur obtenus dans les deux tests sont lees suivants:

sur la phrase	15 %
sur le sens	5 %
sur les mots	5,8 %

Quant au traitement de vocabulaires plus larges, langages plus complexes et pour accélerer le processus de reconnaissance on envisage l'utilisation du coprocesseur de programmation dynamique PREVERT dévéloppé au L.I.M.S.I.

5. REFERENCES.

[1] J.J. Mariani, "Esope: Un système de compréhension de la parole continue", Thèse d'état, Paris VI, Juillet 1982.

[2] J.S. Bridle, M.D. Brown et R.M. Chamberlain, "An algorithm for connected word recognition", Proc. IEEE ICASSP-82, p. 899-902, Mai 1982.

[2] J.S. Bridle,M.D. Brown et R.M. Chamberlain, "Continuous connected word recognition using whole word templates", JSRU Research report No. 1013, Mars 1983.

[4] H.Sakoe et S. Chiba, "Dynamic Programming optimization for spoken word recognition", IEEE Trans. on Acoust. Speech and Signal Processing, Vol. ASSP-26, p. 43-49, Fevrier 1978.

[5] H. Sakoe, "Two level DP-matching. A dynamic programming based pattern matching algorithm for connected word recogniton", IEEE Trans. on Acoust. Speech and Signal Processing, Vol. ASSP-27, p. 588-595, decembre 1979

[6] J.J. Mariani "Contribution à la reconnaissance de la parole continue utilisant la notion de spectre différentiel", Thèse de docteur-ingénieur, Orsay 1977.

[7] W.A. Lea et J.E. Shoup, "Contributions of the ARPA-SUR project" dans "Trends in speech Recognition", Prentice-hall, 1980.

[8] F. Neel, M. Eskenazi, J.J. Mariani Etiquetage phonétique automatique du signal de parole continue à partir de sa transcription orthographique", FASE/DAGA, Gottingen, Septembre 1982.

[9] J. L. Gauvain, "Les techniques de reconnaissance globale", Cours GALF, Janvier 1985.

[10] D. Memmi et J.J. Mariani, "A tool for developing applications gramars", Coling, Prague, 1982.

INFERENCE DE REGLES D'ADAPTATION AU LOCUTEUR DANS UN SYSTEME DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE CONTINUE

J. GUIZOL, H. MELONI, J. GISPERT

Groupe Intelligence Artificielle - Faculté des Sciences de Luminy
CASE 901 - 13288 MARSEILLE Cedex 9

ABSTRACT : This paper describes the main mecanisms of the speaker adaptation phase in an Automatic Continuous Speech Recognition System.
The adaptation is derived from a set of sentences that the speaker must pronounce. For each of them, the analysis of the signal which constitutes its realization produces a pseudo-phonetics events string. This one, and the associated phonemic string are both automatically syllabized. Then the system does an association between phonemes and segments by use of phonetics controls general enough to be valid for every speaker.
From this association we derive, for every encountered phonem, a set of specific contextual rules of its realizations. These rules are then processed by statistical means and by generalization.

I - INTRODUCTION

Le Système de Reconnaissance Automatique de la Parole Continue que nous avons élaboré comportait, dans la description qui en est faite dans (5) et (6), un ensemble de règles d'interprétation d'événements pseudo-phonétiques, propre à un locuteur qui était construit de façon non automatique.
Certaines de ces règles exprimaient la stabilité apparaissant dans les événements fournis par la phase d'analyse pour diverses réalisations d'un phonème dans différents contextes.
D'autres au contraire spécifiaient des caractéristiques contextuelles induites par certaines productions.
D'autres enfin donnaient pour chacun des paramètres calculés à partir du signal son niveau de pertinence et sa valeur "idéale" pour le phonème considéré.

Bien évidemment, ces règles devaient pour la plupart d'entre elles être changées ou réajustées "à la main" pour chaque locuteur.

Nous présentons ici les divers aspects du module d'adaptation au locuteur que nous avons inclus dans notre système afin d'automatiser la génération de ces règles.

II - DONNEES DU MODULE D'APPRENTISSAGE

Notre système de R.A.P.C. comporte un niveau d'identification de mots au moyen d'une structure en treillis dont les noeuds initiaux sont constitués d'événements pseudo-phonétiques issus de l'analyse (7). Le module d'apprentissage s'intégrant à ce système devait donc fournir des règles utilisables ultérieurement en particulier par ce niveau et mettant donc en jeu des objets de nature semblables. Ce sont donc des unités associées à des portions stables et transitoires du signal de parole correspondant à la réalisation d'un ensemble de phrases prononcées par le locuteur et contenant tous les contextes "CV" possibles qui constituent la première forme des données de ce module.

La description du processus permettant de les obtenir ayant déjà été détaillée par ailleurs(5) (7), nous nous bornerons à rappeler que ces unités se présentent sous la forme d'une liste comportant un identificateur de macro-classe et vingt sept variables représentées sur des échelles dont l'amplitude suffit à rendre compte des variations significatives à coder.

En association avec la suite des segments composant chaque phrase du corpus d'apprentissage, est fournie la chaîne phonémique effectivement réalisée.
Nous avons choisi d'opérer la mise en correspondance phonème(s)/segment(s) syllabe par syllabe pour des raisons de performance. En effet, le programme écrit en PROLOG réalisant ce cadrage est d'autant plus déterministe qu'il opère sur un nombre d'unités plus faible. Par ailleurs, le découpage en syllabes de la chaîne phonémique d'une part et de la suite des segments d'autre part s'effectue sans difficulté.

II.1 - Découpage de la chaîne phonémique

Le découpage syllabique a été réalisé simplement en distinguant cinq degrés associés par ordre croissant aux occlusives, fricatives, nasales et liquides, semi-voyelles, voyelles (1). Dans ces conditions, les coupes sont trouvées de gauche à droite avant chaque minimum.

Ce mode de découpage est adapté au mouvement ouvrant de la syllabation en Français et rend compte impli-

citement de la caractéristique d'anticipation vocalique et de non anticipation consonnantique.

II.2 - Découpage de la suite de segments

Afin de pouvoir faire correspondre phonèmes et segments syllabe par syllabe, le découpage de la suite de segments devra suivre la même règle.
Il se fait principalement par détection des noyaux consonnantiques et en disposant la coupe avant chaque segment faisant apparaître un minimum relatif d'énergie. D'autres paramètres tels que la durée relative permettent d'affiner la méthode. On utilise de plus les macro-classes étiquetant les segments afin de pouvoir disposer comme pour les phonèmes de niveaux comparables à des degrés d'aperture.
Un traitement particulier est opéré pour reconnaître les pauses et rendre les segments correspondants inaccessibles par la suite.

II.3 - Cadrage phonèmes - segments

Le regroupement en syllabes étant fait, des règles s'appuyant sur des contraintes suffisamment générales pour être valides pour tout locuteur effectuent la mise en correspondance.
A ce niveau, la démarche ne consiste pas à chercher un indice de ressemblance entre un segment et une représentation "idéale" du phonème que l'on désire lui faire correspondre (3). On vérifie simplement que des indices très larges induits par les phonèmes de la syllabe considérée sont contenus dans les segments que l'on cherche à lui associer.

Exemple : Cherchant à cadrer la chaîne/psi/ avec une suite de segments, nous affecterons d'abord à /i/ le segment déterminé comme étant le plus vocalique ainsi que les segments éventuels qui suivent. Parmi les segments restant en début de liste, /p/ est associé au(x) segment(s) ayant un caractère occlusif sourd (macro-classe "OC", Fondamental nul, énergie inférieure à un seuil), quant à /s/ on lui fait correspondre le ou les segments à caractère constrictif sourd (Densité - Passages par zéro supérieure à un seuil, Fondamental nul, macro-classe "CC").

Il est à noter que dans la liste des segments associés à un phonème, le premier sera le plus caractéristique. Autrement dit les segments transitoires éventuels l'introduisant auront été rattachés au phonème précédent.

Cette mise en correspondance de l'ensemble S des segments produits par les réalisations des phrases du corpus avec l'ensemble P des phonèmes qui les constituent donne donc naissance à un ensemble R(P,S) de règles sur lesquelles vont être appliquées des techniques d'apprentissage.

III - METHODES CHOISIES

La nécessité d'adaptation d'un système est apparue il y a une vingtaine d'années; mais le véritable point de départ de l'étude de ces techniques se situe en 1970 avec la thèse de Winston. Plusieurs formes d'apprentissage ont été imaginées : par mémorisation, par analogie, par exemples, dirigé par les données, par un modèle ...

Dans l'application qui nous intéresse et d'après le paragraphe précédent l'ensemble R(P,S) produit comporte des règles hors contexte(dans le cas où l'on a pu faire correspondre un phonème à un segment) et d'autres contextuelles (lorsqu'il y a eu insertion ou fusion/suppression de segments).

Ne disposant d'aucun modèle à satisfaire à priori, nous avons choisi des méthodes d'apprentissage à partir d'exemples, conduites par les données.
Ces techniques peuvent être utilisées dans plusieurs buts : classification, prédiction ou encore compression de données.

Le traitement que nous effectuons sur les règles de R(P,S) poursuit en fait ces trois buts. Nous opérons en fait :

1°) Une généralisation statistique ayant pour but de déterminer pour chaque phonème dans un espace défini par un sous-ensemble des paramètres des événements, un pavé centré sur le représentant idéal moyen.
2°) En ajoutant à l'ensemble R(P,S) des règles indiquant pour chaque phonème ses caractéristiques acoustiques et en utilisant une technique d'analyse discriminante, on peut représenter un phonème par une série de contraintes (inéquations linéaires) évaluées de façon hiérarchisée selon le système de traits non minimal du Français (4).
3°) Une généralisation sur les règles précisant la suppression ou l'insertion de segments.

IV - INFERENCE DE REGLES HORS CONTEXTE

Nous avons vu que l'ensemble des règles R(P,S) avait été construit de façon à ce que le premier segment de la liste associée à un phonème soit le plus représentatif de cette unité.
On définit alors l'ensemble

$$R'(P,S) = \left\{ /p/ \rightarrow s \text{ tq } /p/ \rightarrow s.x \in R(P,S) \text{ et } s \neq \emptyset \right\}$$

Sur chaque sous ensemble de R'(P,S) dont les éléments ont même partie gauche, on calcule pour chaque paramètre l'écart-type normalisé afin de déterminer ceux ayant une faible dispersion. Pour chaque paramètre retenu un voisinage centré sur la valeur moyenne et d'un rayon évalué à partir de l'écart-type déterminera l'ensemble des valeurs acceptables de ce paramètre pour le phonème considéré.
On obtiendra donc des règles de la forme

$$/p/ \rightarrow seg((p_1.r_1).0.0.(p_4.r_4).(p_5.r_5).0.\ldots.)$$

où P_i détermine la valeur centrale du paramètre i, r_i le rayon du voisinage, 0 désignant les paramètres non significatifs.

V - INFERENCES DE REGLES DISCRIMINANTES

En rajoutant à l'ensemble R'(P,S) des règles précisant les caractéristiques acoustiques de chaque phonème, nous pouvons extraire deux sous-ensembles se rapportant à des réalisations représentatives d'un trait particulier et de son opposé.

Un prédicat évaluable permet alors en passant sous environnement FORTRAN d'appliquer aux deux classes ainsi déterminées un algorithme d'analyse discriminante (2) ayant pour effet de calculer la forme linéaire associée aux paramètres les plus pertinents pour séparer des événements sur la base de traits acoustiques. Cette opération étant effectuée pour tous les traits, le système disposera donc ensuite pour chacun d'eux d'une inéquation correspondant à une contrainte d'application de règle.

Nous avons donc ainsi défini un ensemble de contraintes qui constitueront lors de la reconnaissance les éléments d'un filtre phonétique.

VI - INFERENCE DE REGLES CONTEXTUELLES

Ainsi que nous l'avons dit précédemment, le choix de découpage en syllabes que nous avons fait s'appuie sur la caractéristique du "mode croissant" de la syllabation française. En conséquence les règles contextuelles que nous considérons seront généralement de la forme :

/<consonne>//<voyelle>/ → suite de segments;

Toutefois, il peut advenir qu'un segment regroupe plusieurs consonnes. D'autre part, la méthode de coupe choisie rend possible l'apparition d'une consonne en fin de groupe syllabique si celle commençant le groupe suivant à un degré d'aperture moindre. Dans ces conditions, il peut y avoir fusion entre voyelle et consonne. Par exemple dans la phrase : "Jean pense l'avoir pris", /vwar/ constitue une syllabe dont une réalisation peut faire apparaître une fusion de /v/ et /w/ et une de /a/ et /r/.
En résumé, on aura donc des règles dont la partie gauche sera de la forme /<consonne>//<voyelle>/, /<consonne>//<consonne>/, /<voyelle>//<consonne>/. Les inférences produites vont être réalisées par généralisation sur cet ensemble de règles.

La méthode proposée par Mitchell (8) (9) est séduisante. Elle consiste à considérer trois ensembles de règles :

1°) L'ensemble H qui contient toutes les "règles-exemples" caractérisant un même phénomène.
2°) L'ensemble S initialisé par un élément positif de H et qui est généralisé par les méthodes habituelles (suppression de contraintes, changement de constante en variable...) à chaque nouvelle règle positive de H considérée.
3°) L'ensemble G initialisé par une règle représentant le fait le plus général que l'on désire désigner, que l'on spécifiera par ajout de contraintes successives à chaque fois qu'une règle négative de H est choisie.

Lorsque G devient identique à S, la généralisation optimum est obtenue. Dans l'application qui nous intéresse ici, les règles que nous allons produire ne sont pas des représentations uniques d'un phénomène. Il faut en fait les considérer comme éventuellement applicables dans certains cas de figure et non pas systématiques. D'autre part, on ne dispose que de règles-exemples positives. Le choix et la nature de règles négatives que l'on pourrait imaginer seraient d'une pertinence douteuse.

Nous avons donc préféré la technique proposée par Vere (10) qui procède à un apprentissage seulement sur des règles positives en recherchant la généralisation minimum vérifiant l'ensemble de ces règles.

Les étapes suivies pour procéder à la généralisation sont les suivantes :

1°) Les règles sont regroupées suivant le type de leur membre gauche (Consonne-Voyelle ou Consonne-Consonne ou Voyelle-Consonne).
2°) Les groupes ainsi créés sont traités successivement après avoir développé le membre gauche des règles qui les composent afin de caractériser les phonèmes y apparaissant par leurs traits acoustiques.
3°) Les règles de chaque groupe sont ensuite reclassées en fonction des traits.
4°) On considère alors la plus importante de ces classes afin d'opérer une caractérisation du membre droit des règles la constituant grâce à une évaluation des contraintesfournies par les règles discriminantes.
5°) En otant une à une les contraintes installées dans le membre gauche, on vérifie si éventuellement des règles-exemples d'autres classes ayant une caractérisation similaire du membre droit peuvent être regroupées à la classe de départ et prélevées de leur classe primitive.

Lorsque aucune règle d'aucune autre classe ne peut être prise en compte, la généralisation obtenue constitue une règle qui est ajoutée au système.

Cette suite d'opérations est réalisée pour chaque classe et pour chaque groupe.
Exemple :
Soient les règles du groupe CV :

CV(tt.uu) → seg(OC,t1)seg(CC,t'1) ; (1)
CV(kk.uu) → seg(OC,t2)seg(VC,t'2) ; (2)
CV(dd.ii) → seg(CO,t3)seg(VC,t'3) ; (3)
CV(dd.ii) → seg(OC,t4)seg(VC,t'4) ; (4)
CV(dd.yy) → seg(OC,t5)seg(CC,t'5) ; (5)
CV(pp.yy) → seg(OC,t6)seg(CC,t'6) ; (6)
CV(tt.au) → seg(OC,t7)seg(CO,t'7) ; (7)
CV(tt.ii) → seg(CO,t8)seg(CC,t'8) ; (8)

où ∀i, ti et t'i sont les termes contenant les valeurs des 27 paramètres que nous ne détaillerons pas ici.

Le développement des membres gauches donne:

(1) CV((non vocalique.interrompu.écarté.aigu.sourd).
(oral.écarté.aigu.fermé.bémolisé))
(2) CV((non vocalique.interrompu.compact.Ø.sourd).
(oral.écarté.aigu.fermé.bémolisé))

(3) CV((non vocalique.interrompu.écarté.aigu.sonore).
(oral.écarté.aigu.fermé.diesé))
(4) CV((non vocalique.interrompu.écarté.aigu.sonore).
(oral.écarté.aigu.fermé.diesé))
(5) CV((non vocalique.interrompu.écarté.aigu.sonore).
(oral.écarté.aigu.fermé.diesé))
(6) CV((non vocalique.interrompu.écarté.aigu.sourd).
(oral.écarté.aigu.fermé.diesé))
(7) CV((non vocalique.interrompu.écarté.aigu.sourd).
(oral.écarté.grave.Ø.Ø))
(8) CV((non vocalique.interrompu.écarté.aigu.sourd).
(oral.écarté.aigu.fermé.diesé))

Remarque : Nous avons choisi un traitement spécial pour les semi-consonnes qui sont traitées en tant que voyelles (variantes de /ii/,/uu/,/ou/) lorsqu'on les considère en contexte avec une consonne qui les précède, en tant que consonnes ou dans un contexte où une voyelle les suit.

Dans l'exemple présent, la classe la plus importante obtenue est constituée des règles (3), (4) et (5) et la caractérisation du segment inséré dans le membre droit montre grâce aux termes t'3, t'4 et t'5 qu'il est de nature constrictive. En supprimant la contrainte induite par le trait sourd/sonore de la consonne et en supposant que le terme t'8 permet de confirmer la nature constrictive du segment supplémentaire, la règle (8) est "subsumée" par la règle ainsi généralisée. De même en otant le trait diésé/bémolisé de la voyelle, puis le trait grave/aigu de la consonne, les règles (1) puis (6) se trouvent ainsi prises en compte par la généralisation si les termes t'1 et t'6 traduisent un caractère constrictif. Quant aux règles (2) et (7), leurs termes respectifs t'2 et t'7 ne répondant pas à ce critère, elles seront traitées ultérieurement. La règle inférée est donc la suivante :

CV((non vocalique.interrompu.écarté).
(oral.écarté.aigu.ferme)) ⟶
segment-occlusif(x) segment-constrictif(y);

ou encore :

CV((tt.dd.pp.bb).(ii.uu.yy.uy)) ⟶ segment-occlusif(x) segment-constrictif(y); {tt.dd.pp.bb} étant l'ensemble des consonnes caractérisées par les traits non vocalique, interrompu et écarté, {ii.uu.yy.uy} étant celui des voyelles ou semi-consonnes traitées en tant que voyelles et dont les traits communs sont oral, écarté, aigu, fermé.

CONCLUSION

Les règles inférées par ces méthodes constituent une aide appréciable dans notre système de reconnaissance notamment en ce qui concerne la caractérisation des unités pseudo-phonétiques et le filtrage des accès lexicaux dans la phase remontante. D'autre part, les règles d'interprétation des événements pseudo-phonétique inférées automatiquement permettent une adaptation au locuteur plus précise et moins fastidieuse qu'auparavant.

Toutefois, un problème demeure dans le choix et la taille du corpus d'apprentissage. En effet, la prononciation d'une vingtaine de phrases par un locuteur peut donner lieu à la génération de règles correspondant à des résidus de classes qui ont un caractère exceptionnel. D'autre part, les contraintes limitées à des inéquations linéaires devront être affinées en fonction du contexte et combinées entre elles au moyen d'expressions booléennes.

BIBLIOGRAPHIE

(1) Delattre P., "Principes de phonétique française", (1951).
(2) Guizol J., Méloni H., "Apport de l'analyse discriminante au problème de l'adaptation au locuteur". 12è J.E.P. Montréal (1981).
(3) LeCorre C., Vives R., "Un programme de cadrage pour l'adaptation au locuteur en Reconnaissance Automatique de la Parole". 2è Congrès AFCET-IRIA, Toulouse (1979).
(4) Malmberg B. "Phonétique française" (1972).
(5) Méloni H., "Etude et réalisation d'un système de reconnaissance automatique de la parole continue". Thèse d'Etat (1982).
(6) Méloni H., Guizol J., "A Speech Recognition System". Congrès ICASSP 82. Vol. 3 -pp. 1625-28.
(7) Méloni H., Guizol J., "Identification d'événements pseudo-phonétiques pour la reconnaissance automatique de la parole" 11è I.C.A., Toulouse (1983).
(8) Mitchell T.M., "An Analysis of generalization as a search problem" IJCAI 6, pp. 577-582 (1979).
(9) Mitchell T.M., "Version-Spaces : A candidate elimination approach to rule learning" - IJCAI 5, pp. 305-310 (1977).
(10) Vere S.A., "Inductive learning of relational productions". (Pattern-directed inference systems) NewYork : Academic Press.

****
* DIVERS *
****

ANALYSE DE STRUCTURES D'ENONCES ORAUX

D. Luzzati et F. Néel

LIMSI-CNRS
B.P. 30 - 91406 ORSAY Cedex

ABSTRACT :

The structure of oral utterances in man-machine communication is not easily handled by the use of traditional tools such as ATN ... Moreover, all the words pronounced are not significant. The purpose is to isolate the "information core" in the message. Two types of analyses are proposed here, based on the detection of markers and redondancies very commonly used in fluent speech.

INTRODUCTION

On ne peut à l'aide de schémas théoriques (1) rendre complètement compte des énoncés oraux réellement produits en situation de communication homme-machine. On observe des phénomènes (redondances, reprises, hésitations, structures asyntaxiques ou discursives, marqueurs spécifiques ...) propres à mettre en défaut les analyseurs syntaxiques traditionnellement utilisés dans la plupart des systèmes de compréhension (ATN ...).

Des analyses différentes sont à envisager. Celles que nous proposons ici supposent que le décodage acoustico-phonétique a été capable de restituer sans erreur la totalité du message réellement prononcé, ce qui,en fait, est loin d'être le cas. Ces analyses fonctionnent sur des transcriptions et présentent des problèmes de formalisation informatique si l'on veut tenir compte d'indéterminisme de la reconnaissance vocale.

NOYAUX INFORMATIFS

Nous avons étudié plus particulièrement des énoncés produits à partir d'un scénario de sondages téléphoniques dans lesquels les locuteurs croient être interrogés par une machine dont la compréhension n'est pas limitée :

mach : si vous achetiez une automobile, laquelle choisiriez-vous ?
loc : hé évidemment la moins chère
mach : pourquoi ?
loc : hé je viens de vous dire hein c'est un véhicule euh utilitaire euh/ euh/ j'en vois pas d'autres d'autres fin pour faire autre chose quoi/ hein on est dans un lieu assez iso enfin isolé non mais enfin disons sans transports en commun bon ben/ pour nous une voiture c'est c'est utilitaire c'est tout (2).

Dans de tels énoncés, nous considérons que la totalité des propos tenus n'est pas intéressante au même degré. Une partie d'entre eux peut être ignorée par la machine sans que la "compréhension" en soit perturbée. Avant de chercher à comprendre, il s'agit en somme de sélectionner afin que la machine retienne d'elle-même ce que nous appelons les noyaux informatifs (NI). A partir d'un même énoncé-test, voici deux types d'analyses qui peuvent y contribuer :

a) Le jeu des marqueurs

Quelques termes ont un rôle structurel et sont presque totalement dépourvus de sens propre. Qu'on les appelle des "marqueurs de structuration de la conversation" (3), des "mots du discours" (4) ou des "appuis du discours" (5), il s'agit toujours d'un même ensemble d'outils, c'est-à-dire, dans notre extrait, "hein", "quoi" ou "ben". La liste de ces marqueurs n'est jamais close (tout idiotisme notamment est susceptible d'en devenir un) et leur jeu n'a de valeur que dans un contexte et en fonction de l'échange.

Dans l'exemple ici en cause, il s'agit d'une réponse à une question explicative (pourquoi ?). Dans ce type de réponse, en l'absence d'un "parce que", le NI se trouve en général en tête d'un énoncé. La longueur de l'énoncé et la présence d'un "ben" conclusif, renforcé par "bon", doit nous inciter à retenir également le syntagme final dans les groupes susceptibles de recéler le NI.

b) Le jeu des redondances

En 1979, C. Blanche-Benveniste (6) proposait une structuration de l'oral en grilles, qui mettait en évidence le caractère redondant de cet oral. Une telle structuration peut également permettre de contribuer à isoler les NI.

On peut relever trois types de redondances :
- les redondances conjointes (-----) : hésitation
- les redondances itératives (———) : reformulation, énumération
- les redondances disjointes (=====) : mot clé

hé je viens de vous dire hein c'est un véhicule euh utilitaire euh/ euh/ j'en vois pas

d'autres d'autres fin pour faire autre chose quoi/ hein on est dans un lieu assez iso enfin isolé non mais enfin disons sans transports en commun bon ben/ pour nous une voiture c'est c'est utilitaire c'est tout (2).

On constate, dans cet énoncé, que le jeu cumulé des marqueurs et des redondances met en évidence le mot clé "utilitaire". C'est bien là le coeur d'un NI qui, pour notre locuteur, répondrait à la question : "Pourquoi, si vous avhetiez une automobile, choisiriez-vous la moins chère ?".

LIMITES D'APPLICATION

Ces deux types d'analyse exigent évidemment du décodage acoustico-phonétique une précision capable de détecter soit ces marqueurs (souvent monosyllabiques) soit des répétitions de mots inachevés (dont la première syllabe ne doit pas être assimilée à un mot complet de la langue).

Ces techniques de traitement du discours et les formalisations qui en découlent sont en outre limitées par leur marge d'erreur et par leur sphère d'application.

a) Marge d'erreur

mach : si vous achetiez une automobile, laquelle choisiriez-vous ?
loc : alors là j'y ai pas pensé
mach : pourquoi ?
loc : ah ben pasque euh/ on n'a plus vingt ans et que/ pas question d'en racheter une autre/ alors on garde celle-là. (2)

Un énoncé tel que celui-ci est difficile à traiter dans la mesure où il est fondé sur une incompréhension : le locuteur ne répond pas à la question réellement posée par la machine. Le jeu des marqueurs peut cependant conduire sur une fausse piste : on sera enclin à chercher le NI après "pasque" et après "alors" et le jeu inexistant des redondances ne permet pas d'y remédier.

En fait, le problème des marges d'erreur est vaste. Il s'agit de poser des bornes en précisant si le NI les précède ou les suit, la marge d'erreur est faible. S'il s'agit de relever des mots-clés, cette marge est beaucoup plus forte.

b) Sphère d'application

Tout changement de sujet ou de type d'échange peut rendre caducs des résultats obtenus dans des circonstances différentes. Souvent, lorque le vocabulaire change, les marqueurs se modifient. Le jeu des redondances s'atténue dès lors que l'échange est plus contraint,c'est-à-dire lorsque le locuteur est plus motivé pour obtenir un renseignement.

L'exemple suivant illustre un type de corpus différent de celui dont ont été tirés les précédents exemples :

loc : Allo oui bonjour je suis actuellement étudiante à Paris euh/ ex Vincennes St Denis je prépare un DEUG de langue littérature et culture et je serais intéressée pour poursuivre mes études à Paris III/ euh pour préparer une licence d'espagnol/ et étant donc en première année j'aimerais l'année prochaine euh/ poursuivre mes études ici à Paris III et j'aimerais des informations pour euh le transfert de mon dossier savoir si je suis acceptée/ à cette faculté. (7)

Les marqueurs sont moins nombreux et ils sont différents : "donc", "et". Mais ils sont cependant présents et permettraient d'opérer une première segmentation du flot continu de parole.

Le jeu des redondances est plus sommaire et la présence d'une seule redondance itérative interdit ici de conclure.

CONCLUSION

Le jeu des marqueurs et celui des redondances sont des techniques de traitement du discours qui aboutissent à quelques résultats. Il faut toutefois rester prudent :

- Ces types d'analyse ne tiennent pas compte des difficultés de reconnaissance orale des monosyllabes.
- Il faut développer plusieurs techniques d'analyse pour parvenir à une quelconque fiabilité.
- Chaque corpus ou presque nécessite une analyse particulière : les marqueurs sont dépendants de la tâche et du locuteur.
- Dans les oraux dits finalisés (type renseignements SNCF), la fréquence des marqueurs est moindre mais ils sont présents de même que les hésitations, reformulations ou auto-corrections. Le problème tient alors tout autant à la structure du NI qu'à son repérage.

REFERENCES BIBLIOGRAPHIQUES

(1) J.P. Declès, Actes du séminaire GRECO CP "Dialogue Homme-Machine à composante orale", Nancy, 11-12 octobre 1984.

(2) D. Luzzati, ORSO n° 1, note scientifique LIMSI, septembre 1984.

(3) E. Cülich et A. Auchlin, Cahiers de Linguistique Française, Genève, 1981.

(4) O. Ducrot et Al., Les mots du discours, Edition de Minuit, Paris, 1980.

(5) D. Luzzati, "Ben, appui du discours", Le Français Moderne, juillet 1983.

(6) GARS n° 2, Aix-en-Provence, 1979. Repris par B. Borel, GARS n° 3, 1981.

(7) D. Luzzati, ORSO n° 3 (corpus CIO Paris III, oral finalisé) à paraître, note scientifique LIMSI, 1985.

POUR UNE BASE DE DONNEES DE TRANSCRIPTIONS PHONETIQUES: MOTIVATIONS ET PREMIERE EXPLOITATION STATISTIQUE

J.P. Tubach(*) et L.J. Boë(**)

(*) ENST, Département SYC (CNRS, UA 820), 46 Rue Barrault, 75013 Paris.
(**) Institut de Phonétique (I.C.P.), Grenoble.

A very large (over 300 000 phones) corpus of phonetic texts has been built from different sources. It incorporates different levels of language, and a common codification of the IPA has been used for all texts. We have gathered statistics on frequencies of phones and sequences of 2, 3 and 4 phones, including a previously unpublished list of most frequent quadriphones. We plan to conduct other research using this corpus. We suggest to other researchers to share such phonetic data in order to build a large data base.

I INTRODUCTION

Nous présentons ici la constitution et quelques résultats de l'exploitation d'un vaste ensemble de transcriptions phonétiques, que nous avons constitué aux centres de calcul de l'ENST et de l'Institut de Phonétique. Un important volume de statistiques obtenues sur ces textes se trouve dans TUBACH et BOE, 1985a. Le texte complet des corpus est publié par ailleurs, pour référence (TUBACH et BOE, 1985b)

II Motivations et historique

Le traitement informatique de corpus de transcriptions phonétiques pour en extraire des statistiques diverses ont commencé, pour le français, il y a de plus de quinze ans. On peut citer BONNOT, 1968, TUBACH, 1969, LIENARD, 1972, HATON, 1974. Après une interruption, ce type de travaux connait actuellement de nouveaux développements.

Les motivations des travaux passés et en cours sont variées; il s'agit toujours de rendre compte de la structure de la langue au niveau phonotactique, mais les utilisations possibles de ces informations sont diverses:
- applications à la reconnaissance automatique de la parole, que nous évoquerons de façon plus détaillée au paragraphe suivant.
- applications à l'apprentissage de la lecture, à l'enseignement du français (relations entre le code orthographique et le code phonétique
- étude de la constitution des mots : longueur, occurences, groupes consonantiques ...
- étude des syllabes : règles de syllabation, contenu des syllabes
- étude des traits phonétiques à partir de statistiques
- génération de corpus de test prenant en compte les fréquences des sons et suites de sons
- étude des liaisons
- étude du e muet
-etc...

En ce qui concerne la reconnaissance automatique de la parole, les premières tentatives pour utiliser des statistiques sur les suites de sons sont déjà anciennes: après que les pionniers de l'époque héroïque aient perdu leurs illusions sur la possibilité de réaliser rapidement une "machine à écrire phonétique" (OLSON et BELAR, 1957), le premier travail fut celui de FRY et DENES, à University College, à Londres: la prise en compte des probabilités de transition entre "phonèmes" leur permettait d'améliorer de façon significative les résultats obtenus par un reconnaisseur de "phonèmes" rudimentaire (DENES, 1959. DENES, 1963).

Dans son système de reconnaissance pour le japonais, DOSHITA, 1965, utilisait les fréquences d'occurences des suites de trois phones (qu'il appelait trigrammes) pour tenir compte de la structure linguistique du message.

Les années 70 ont vu une tendance à utiliser des traitements syntaxiques et sémantiques pour compléter le décodage acoustique phonétique. Mais plusieurs équipes étudient à nouveau l'introduction de statistiques sur les suites de phones, par exemple pour mieux cheminer dans un treillis phonétique, en tenant compte des séquences qui ne se rencontrent pas dans la langue.

III LES CORPUS UTILISES

Tous les textes composant ces corpus sont des transcriptions phonétiques faites par des phonéticiens expérimentés, en écoutant l'enregistrement des conférences ou conversations. Cette méthode est couteuse, mais fournit une information plus exacte et plus précise sur la langue parlée que les textes phonétiques obtenus en appliquant un programme de traduction orthographique - phonétique à des textes écrits.

Nous appelons "phone" (on pourrait aussi dire "son") l'unité de base des transcriptions. Elle est une réalisation d'un phonème, unité linguistique plus abstraite. Un certain nombre de diacritiques précise éventuellement cette réalisation (par exemple assourdissement, allongement, accentuation etc.) Lorsque nous employons le mot diphone, nous voulons seulement désigner une suite de deux phones. On notera que ce n'est en général pas dans ce sens qu'est employé le mot diphone par les chercheurs en reconnaissance et en synthèse de parole. Nous pensons que dans le contexte de ce rapport, aucune confusion n'est possible. De la même façon, nous appelons triphones et quadriphones les suites de trois et quatre phones, respectivement.

-- Corpus "G"

Ce corpus, de 86360 phones a été constitué à l'Institut de Phonétique de Grenoble, en 1967 et 68, sous la direction de R. GSELL. Il contient 3 heures de conférences de France - Culture . Il est donc représentatif d'un langage parlé cultivé et assez formel. La transcription est fine et comporte de nombreux diacritiques. Ce corpus avait déjà été utilisé par l'un des auteurs, lors de sa création (TUBACH, 1969) pour une étude statistique, et plus récemment (TUBACH et al., 1981) pour une étude fondée sur la théorie de l'information.

- Corpus "M"

Ce corpus de 201281 phones à été constitué à l'Université de Californie (Santa Barbara) par A. MALECOT (MALECOT, 1972). Nous avons pu en disposer, avec son autorisation, grace à l'obligeance de G. CHOLLET. MALECOT a enregistré cinquante conversations informelles d'une demi-heure, sur des sujets très divers, avec des membres de "l'intelligentia" parisienne. De ces enregistrments, il a conservé les informations se rapportant à ce qui était prononcé par ses interlocuteurs. Il s'agit d'un niveau de langue plus familier, mais pour des locuteurs cultivés. La transcription est moins fine (par exemple les deux a ne sont pas distingués).

- Corpus "J"

Ce corpus de 17111 phones a été constitué recemment à l'Institut de Phonétique de Grenoble par Joelle VAN EIBERGEN (VAN EIBERGEN, 1985). Il comporte les transcriptions de 16 brèves conversations, pour 16 locuteurs, d'origines linguistiques et socio-professionnelles variées (4 enfants de moins de 20 ans, 8 adultes de 20 à 60 ans, 4 adultes de 60 à 80 ans). Ces conversations sont en situation informelle, entre familiers, sur un pied d'égalité. Le niveau de langue représenté est un français très spontané La transcription est très proche de celle utilisée pour le corpus "G".

L'ensemble des trois corpus représente en tout 304752 phones.

CODIFICATION

Etant données leurs origines différentes, ces corpus n'étaient pas codés de la même façon. Nous avons défini une représentation du sous-ensemble de l'API utilisé pour le français, destinée à être imprimable par une imprimante d'ordinateur standard, et à être facilement lisible par un phonéticien. Nous avons recueilli les suggestions d'un certain nombre de phonéticiens et d'utilisateurs potentiels de ces corpus.

Nous avons tenu compte des nombreuses listes déjà proposées (HATON, 1976. PERENNOU, 1984). Nous avons considéré que les anciens codes à un caractère par phone, souvent en majuscules seulement, et qui datent de l'époque des perforatrices, sont par trop "cabalistiques", et ne sont lisibles qu'après un fastidieux et stérile entrainement. Par ailleurs, un code comportant deux caractères par phone serait bien mal utilisé. Nous avons donc adopté ce code à 1 ou 2 caractères par phone. Il est sans ambiguité, et le sous - programme capable de fournir le code du phone suivant la position courante dans le texte est assez simple.

Cette codification est assez proche de celle utilisée à l'INRS de Montréal (LENNIG et BRASSARD, 1984). les différences visent à accroitre la lisibilité (en contre partie, elle ne permet de transcrire que le français). Nous proposons de l'utiliser pour d'autres corpus qui pourraient venir augmenter la base de données phonétique.

API	symbole	séparateurs	
u	u	blanc	de mots
i	i	/	de groupe rythmique
e	e	#	de phrase
ɛ	E	-	de liaison
a	a	[	debut de transcriptrion
ɑ	A	]	fin de transcription
o	o	x	partie non transcrite
ɔ	O		
y	y	diacritiques	
ø	EU		
œ	OE	.	assourdissement
ə	&	^	sonorisation
ɑ̃	a~	:	allongement
ɔ̃	o~	'	accent
œ̃	u~	"	accent renforcé
ɛ̃	e~		
j	j		
w	w		
ɥ	Y		
p	p		
t	t		
k	k		
b	b		
d	d		
g	g		
v	v		
z	z		
ʒ	3		

f	f	!
s	s	!
ʃ	S	!
m	m	!
n	n	!
ɲ	n~	!
l	l	!
R	R	!

voici un exemple de texte codé (début du corpus "G")

```
[x / pRezi'da~ / d& la kOmisjo~ nasjO'nal /
d amena3'ma~ / dy tERi'twa:R / pRezi'da~ /
d& la ko~pa'n~i / d amena3'ma~ / dy baRo d& x /
m&sjEU x / depy'te / a~sje~ mi'nistR. /
pRezi'da~ / dy ko~sEj nasjO'nal /
de-z ekOnOmi Re3jO'nal / x diREk'tOE:R /
d& l Eks'pREs / e o'tOE:R / dy
defi ameRi'ke~ # swasa~tmiljo~ d& fRa~'sE /
```

IV EXPLOITATION

Le but de ce travail étant de faire des statistiques sur les suites de phones, nous avons créé des versions des corpus ne comprenant que les informations nécessaires:
- phones: tous conservés. [&] est remplacé par sa réalisation [OE], d'où 35 phones différents (en fait 34 seulement dans le corpus "M" qui ne distingue pas les deux a).
- diacritiques : tous supprimés.
- séparateurs: seuls sont conservés # (séparateur de phrases) et x (partie non transcrite).

Les groupes de plusieurs phones ne traverseront bien entendu pas les frontières # et x.

V QUELQUES RESULTATS

On peut examiner le rapportdu nombre de poly-phones différents rencontrés à leur nombre maximal possible: On obtient, évidemment 100% à l'ordre 1 (phones); 87% à l'ordre 2: la plupart des diphones possibles sont rencontrés; à l'ordre 3:environ 30% (12019 sur 34^3 = 39304); ceci constitue un résultat très intéressant et riche d'applications potentielles en assistance à la reconnaissance de la parole. On peut considérer qu'il n'est pas du à une taille trop faible du corpus, car il est stabilisé bien avant la longueur traitée ici.

Par contre, à l'ordre 4, le corpus n'est pas assez long pour une étude des quadriphones autres que les plus fréquents. En effet, sur les 50451 quadriphones différents rencontrés, les quatre cinquièmes n'interviennent que de 1 à 5 fois. Le fait de ne rencontrer que 3.78% des quadriphones possibles ne correspond pas à une information stable, et croîtrait avec la taille du corpus.

Il est intéressant pour les poly-phones fréquents, d'examiner les rapports du type pij/(pi*pj) (expression pour les diphones): on peut ainsi distinguer ceux qui le sont parce que leurs composants sont par eux-mêmes fréquents et ceux qui le sont par effet de distribution spécifique. Par exemple, [st] (1.36) et [ak] (1.29) s'opposent à [Yi] (16.92) et [jo~] (12.26).

Pour les triphones, [sjo~] (33), [bje~] (81), [a~fe~] (51) appartiennent à la seconde catégorie,et [tRE] (2.4), [ala] (2.7), [ete] (3.2) à la première. (On remarque, pour les triphones fréquents, qu'on peut souvent repèrer quels mots, parties de mots ou suites de mots fréquents les fournissent).

Pour les quadriphones fréquents, c'est le second type qui prédomine: [30EsY] (137), [OEkRw] (82) etc. Rares sont les valeurs inférieures à 3 ([paRl], [paRt], [paRi], [paRE] ...) (Il est ici encore plus facile que pour les triphones de voir de quels mots fréquents proviennent ces quadriphones (alors, parce que, il y a, de la, je crois, je suis, etc...).

quadriphones, liste triée

format: rang phi phj phk phl nijkl pijkl% pijkl/(pijk*pl) pijkl%cumule

nombre quadriphones: 259903

1	a	l	O	R	567	0.22	13.88	0.22
2	p	a	R	s	472	0.18	4.76	0.40
3	s	OE	k	OE	447	0.17	10.89	0.57
4	a	s	j	o~	446	0.17	48.96	0.74
5	i	l	i	a	436	0.17	7.19	0.91
6	d	OE	l	a	418	0.16	7.68	1.07
7	R	s	OE	k	382	0.15	20.78	1.22
8	a	R	s	OE	374	0.14	12.74	1.36
9	a	v	E	k	346	0.13	18.22	1.50
10	f	R	a~	s	338	0.13	15.93	1.63
11	e	p	a	R	334	0.13	5.44	1.75
12	k	R	w	a	273	0.11	13.23	1.86
13	OE	s	Y	i	272	0.10	19.18	1.96
14	v	w	a	R	272	0.10	6.69	2.07
15	s	e	p	a	265	0.10	8.92	2.17
16	p	a	R	l	264	0.10	2.83	2.27
17	l	OE	m	a~	258	0.10	17.91	2.37
18	OE	k	R	w	231	0.09	82.00	2.46
19	e	b	j	e~	229	0.09	83.52	2.55
20	u	z	a	v	223	0.09	35.32	2.63
21	3	OE	k	R	220	0.08	12.00	2.72
22	b	o	k	u	217	0.08	52.35	2.80
23	p	a	R	t	215	0.08	2.54	2.89
24	k	E	l	k	214	0.08	15.12	2.97
25	a	p	a	R	213	0.08	6.84	3.05
26	s	E	R	t	204	0.08	11.54	3.13
27	t	R	u	v	203	0.08	42.25	3.21
28	a	v	w	a	197	0.08	13.16	3.28
29	t	R	a	v	188	0.07	19.92	3.35
30	a	R	t	i	186	0.07	11.77	3.43
31	E	t	R	OE	185	0.07	8.50	3.50
32	v	u	z	a	185	0.07	7.94	3.57
33	3	OE	s	Y	185	0.07	136.93	3.64
34	E	l	k	OE	184	0.07	12.68	3.71
35	e	p	Y	i	182	0.07	19.18	3.78
36	p	u	R	l	182	0.07	4.34	3.85
37	e	t	R	e	179	0.07	6.96	3.92
38	s	OE	k	i	179	0.07	4.43	3.99
39	m	e	R	i	176	0.07	17.05	4.06

40	a	m	e	R	175	0.07	7.50	4.12
41	l	i	t	e	175	0.07	10.00	4.19
42	p	a	R	i	175	0.07	1.96	4.26
43	e	R	i	k	174	0.07	18.45	4.32
44	m	w	a	3	169	0.07	22.17	4.39
45	v	u	l	e	169	0.07	9.03	4.45
46	k	OE	3	OE	169	0.07	10.96	4.52
47	i	l	i	j	166	0.06	12.37	4.58
48	a~	m	E	m	165	0.06	26.36	4.65
49	p	E	R	s	165	0.06	9.33	4.71
50	OE	m	a~	d	162	0.06	5.12	4.77

VI CONCLUSION

Nous menons également sur ces corpus des travaux de syllabation automatique, d'étude de rendement des oppositions, de construction de dictionnaire phonétique.

Bien que cet ensemble de textes ait déjà un volume respectable (plus de 300 000 phones), on a vu qu'il est souhaitable, pour certaines études, de l'agrandir encore. C'est pourquoi nous voulons lancer ici un appel aux chercheurs qui disposeraient de données de même nature pour les mettre en commun, et constituer ainsi une importante base de données de textes phonétiques.

BIBLIOGRAPHIE

J.P. TUBACH, L.J. BOE, 1985a: Un corpus de transcriptions phonétiques (300000 phones): constitution et exploitation statistique. Document ENST, Paris, Avril 1985. 85-D001

J.P. TUBACH, L.J. BOE, 1985b: Un corpus de transcriptions phonétiques (300 000 phones): texte intégral. Travaux de recherche de 'Institut de Phonétique de Grenoble (Institut de la Communication Parlée). Avril 1985.

H.F. OLSON, H. BELAR, 1957: Phonetic typewriter. IRE Transactions on audio, July August 1957.

P.B. DENES, 1959: The design and operations of the mechanical speech recognizer at University College, London. Journal of British IRE, 19 Avil 1959.

P.B. DENES, 1963: On the statistics of spoken english. Journal of the acoustical society of America, vol 35, 1963

S. DOSHITA,1965: Studies on the analysis and recognition of japanese speech sounds. Thèse, Université de Kyoto, Japon, 1965

J.S. LIENARD, 1972: Analyse, synthèse et reconnaissance automatique de la parole, Thèse d'Etat, Université Paris VI, 1972.

J.P. HATON, 1974: Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole. Thèse d'Etat, Université Nancy I, Janvier 1974.

J.A. BONNOT, 1968: Information, redondance et répartition des lettres et des phonèmes en français. Séminaire de culture générale linguistique du Pr. G. BIBEAU, document LING 744, Université de Montréal, Mars 1968,

J.P. TUBACH, 1969: Etude des contraintes statistiques des groupements phonématiques. Congrès l'Informatique au service de l'Homme, Grenoble 1969.

J.P. TUBACH et al, 1981: Utilisation de la théorie de l'information pour une étude quantitative de l'ambiguïté en lecture labiale. 12èmes Journées d'Etudes sur la Parole du GALF, Montréal, mai 1981. A. MALECOT, 1972: New procedures for descriptive phonetics. in Papers in linguistics and phonetics to the memory of Pierre DELATTRE. Mouton, The Hague and Paris, 1972.

J. VAN EIBERGEN, 1985: Etude du e muet dans le français parlé. A paraitre.

J.P. HATON, 1976: Enquête sur les codes phonétiques et orthographiques utilisés par les équipes de langue française. Bulletin du groupe reconnaissance automatique de la parole de l'AFCET vol 5 no 2, Dec. 76.

G. PERENNOU, 1984: Base de données lexicales du français parlé. in Rapport scientifique du GRECO Communication parlée, CNRS. Edité par le CRIN, Université de Nancy I, Juin 1984.

M. LENNIG, J.P. BRASSARD, 1984: A machine-readable phonetic alphabet for French and English. Speech Communication, 1984

FREQUENCES ELEVEES ET PHENOMENES TRANSITOIRES DANS LA RECEPTION D UN MESSAGE VOCAL INATTENDU .

Jacques JANDOT

Département I.S.A.V.
Université Paul Valéry . 34000 MONTPELLIER

A number of experiments on the information-bearing features of recorded British speech , first with French listeners of different age-groups particularly English-learning pupils ,and teachers, then with native speakers in similar listening conditions , have led to the conclusion that higher frequencies , especially between 6000 and 12500 Hz , are of decisive importance for correct perception of highly unexpected words or sentences.

Le but de l'expérimentation , qui a fait l'objet d'une Thèse d'état (U.R.L. de l'université Paris VII) , était de déterminer un seuil qualitatif minimal pour les appareils audio utilisés dans l'apprentissage de l'anglais par des français.

Pendant plusieurs années consécutives, des phrases ou textes en langue anglaise furent présentés à des groupes d'auditeurs , surtout des enfants scolarisés en Collège , sous deux formes bien distinctes: d'abord sans altération autre que celle globalement imposée par toute la chaîne électro-acoustique et acoustique de reproduction (appareil, local d'écoute,etc.); ensuite, sous une forme volontairement modifiée par filtrage.

Dans un premier temps , le filtrage fut passif et relativement grossier, surtout au niveau du contrôle. Il se borna à atténuer la zone dite conversationnelle , de façon à accentuer par contraste perceptif la partie plus grave et plus aiguë du spectre audible.

Les résultats obtenus vinrent confirmer très vite ce que l'intuition "pédagogique" semblait indiquer: la mise en avant de la ligne mélodique , d'une part, et le renforcement artificiel de la partie aiguë du message vocal d'autre part , se traduisaient concrètement par une meilleure intelligibilité des énoncés présentés (par magnétophone).

La différence constatée était particulièrement évidente pour les textes présentés de façon dite "audio-orale" , c'est à dire sans lecture visuelle préalable.

La présentation devint progressivement plus sélective , les énoncés filtrés étant choisis dans un domaine lexical et syntaxique moins connu (voire inconnu) des auditeurs. Les résultats précédents furent confirmés.

L'amélioration notable de l'intelligibilité portait surtout sur la segmentation correcte des énoncés , et la reconnaissance des éléments consonantiques de type plosif ou occlusif (/p/ ; /t/; /k/),notamment en position initiale.

Une expérience ultérieure - fortuite d'ailleurs- apporta une confirmation à ces résultats: lors d'une réunion pédagogique portant justement sur l'étude de textes présentés de façon "audio-orale", un groupe de professeurs d'anglais , comprenant deux personnes d'origine britannique , écouta un texte poétique (relativement difficile , il est vrai) à partir d'un magnétophone de mauvaise qualité qui, à l'évidence , ne transmettait rien au dessus de 3000/6000 Hz , avec une dynamique fort réduite.

Il fut impossible à l'ensemble des auditeurs présents de reconnaître plus d'un cinquième des énoncés vocaux entendus , alors qu'ils étaient prononcés par un locuteur britannique sans accent régional particulièrement marqué.

L'élimination des aigus (la partie grave étant à peu près correctement reproduite) et le manque de dynamique venaient s'ajouter à l'impossibilité des auditeurs, pourtant habitués à l'écoute de la langue parlée , y compris reproduite au magnétophone, de prévoir l'enchaînement et l'organisation du contenu linguistique du message entendu.

Par la suite , les résultats des expérimentations systématiques , réalisées avec des élèves ou des adultes volontaires , purent être affinés.

D'une part, la précision du filtrage et de son contrôle fut améliorée; dans un premier temps en conservant la technique de filtrage passif , ensuite en faisant appel à des modules de filtrage actif à pente plus raide.

D'autre part, une enceinte acoustique mieux adaptée à la diffusion sonore en salle de classe permit de s'assurer que chaque élève pouvait percevoir (à peu près) le même message vocal. (L'enceinte fut adaptée au local d'écoute.).

L'expérimentation recommença dans ces conditions , et il s'avéra que , dans tous les cas où les auditeurs devaient faire appel , non à des énoncés préalablement entendus et mémorisés , ou leur connaissance des éléments présents dans le message, et l'attente relative qui pouvait en découler, mais plutôt à la reconstitution d'unités signifiantes à partir des simples indices acoustico-phonétiques, les fréquences comprises (approximativement) entre 6000 et 12500 Hz jouaient un rôle décisif dans la reconnnaissance correcte de l'énoncé entendu.

Il fut donc décidé de poursuivre l'expérience avec des sujets britanniques. Dans un premier temps les conditions d'écoute furent similaires, et les résultats très voisins - en tenant compte d'une plus grande facilité pour un natif d'opérer une rapide reconstitution à partir de quelques éléments (indices) perçus.

Par la suite , l'audition fut faite à partir de casques d'écoute. Les personnes qui furent volontaires pour ces expériences (assistantes anglaises) devaient écouter et transcrire des paroles de chansons qu'elles avaient la possibilité de re-diffuser à volonté , ce , pour compenser la difficulté due au masquage relatif introduit par la musique.

Les sons étaient filtrés , comme précédemment. Là encore, l'élimination des fréquences aiguës à partir de 6000 Hz environ (précision relative des filtres...) se traduisait par des erreurs d'interprétation tout à fait grossières ; l'auditeur entendait un terme sans rapport avec le contenu précédent des paroles, parfois même sans aucune justification de type syntaxique , mais dont les sons vocaliques présentaient une suite à peu près similaire.

Un exemple: "tiny" fut entendu comme "twenny"(un américanisme connu pour "twenty") alors que rien dans le texte ne pouvait laisser prévoir l'intrusion d'un nombre ou d'un américanisme...

Ces expériences furent complétées par un relevé systématique d' "incidents" de compréhension pour des auditeurs français entendant des énoncés français dans de mauvaises conditions (écoute d'un film dont la piste son était mauvaise , ou au travers d'une chaîne électro-acoustique peu "fidèle",etc.).

La confirmation de l'importance des aigus, au dessus de 6000 à 8000 Hz (selon les individus),ainsi que de la restitution correcte des attaques des consonnes plosives et occlusives , fut apportée.

L'effet est cependant moindre en français, comme d'ailleurs en américain ; cela est dû au moindre "poids" des éléments consonantiques (pour la compréhension) dans ces langues, par rapport à l'anglais.

En conclusion , il semble bien que les systèmes présentant (comme le téléphone)une forte atténuation des composantes les plus aiguës de la parole, et un "gommage" des transitoires d'attaque , rendent quasiment impossible une reconnaissance de type phonétique , même s'ils permettent, en moyenne, une bonne transmission des messages redondants et/ou fortement prévisibles.

L'expérience fort simple qui consiste à dire un

nom propre relativement mal connu , soit au téléphone , soit dans une sonorisation médiocre, a pour résultat infaillible une interprétation erronée (sonore et orthographique) de la part des auditeurs, ce qui montre à l'évidence le poids de la prévisibilité sur l'intelligibilité réelle.

Par ailleurs , on peut noter systématiquement que les auditeurs interprètent ce qu'ils "entendent" en fonction de leur vécu personnel , jusqu'à des limites quasiment absurdes , c'est à dire des cas où le "nom" entendu ne possède qu'un phonème identique à celui qui a effectivement été prononcé.

Dans ces conditions , les habituelles contraintes économiques et techniques liées au traitement et à la transmission des messages vocaux sont-elles compatibles avec les impératifs qualitatifs d'une transmission-réception correcte de ces messages?...

RECHERCHES POUR L'ENSEIGNEMENT DES PRONONCIATIONS ETRANGERES

THOMAS Raymond

Directeur du Laboratoire de Langues, Faculté des Sciences de Marseille-Luminy, 70 Route Léon-Lachamp, 13288 MARSEILLE Cedex 9

ABSTRACT Thanks to information processing, the vast and complex information available in various scientific and linguistic fields on the ways and mechanisms of human pronunciation, can now be usefully organized. Present corrective methods may be better understood and rated, and new and more efficient ones can be made available to teachers and students. This should also lead to a more powerful description of speech, thus benefitting acoustics and neuro-physiology.

Utilité d'une recherche commune.

L'enseignement des prononciations étrangères ne doit pas ignorer les découvertes récentes. L'utilité de la division des syllabes en consonnes et voyelles a fini par imposer dans la pratique l'idée qu'on pourrait reproduire exactement les sons du langage par des lettres qui représenteraient les mêmes phénomènes acoustiques, du moins dans nos langues européennes.

Cette découverte ancienne doit être complétée au plus tôt pour la pédagogie car ses inexactitudes nuisent beaucoup à l'enseignement des langues étrangères. Or cet enseignement s'étend partout et il est urgent de le rendre moins décevant.

Pardonnez ce qui paraîtra peut-être une digression mais est indispensable, en fait. Deux façons d'apprendre sont nécessairement en conflit. L'une fait appel à la réflexion et, pour notre cas, ramène des incohérences apparentes dans les langues à des applications diverses d'une tendance générale. C'est celle dont nous voulons faire usage dans notre enseignement car, tout en le rendant capable d'utiliser l'autre langue, elle dispose l'esprit de l'étudiant à la critique précise et profonde, à la recherche et à la découverte, tout en lui étant utile pour sa propre langue, d'autres idiomes et toute étude linguistique.

L'autre façon d'enseigner cherche seulement le résultat brut: rendre capable, pour notre cas, de se servir de telle ou telle langue étrangère. A cause des techniques modernes, cette seconde méthode est en passe de déclasser la première. L'amélioration de cette autre façon d'apprendre les langues est évidemment une bonne chose en soi, et il serait odieux de chercher en ceci à diminuer la liberté du choix de l'étudiant.

Mais cette seconde méthode est aussi moins utile, et moins ennoblissante pour les personnes, et ceci fait partie des dégradations qui menacent de plus en plus l'esprit humain. Notre impérieux devoir d'enseignant est donc de rendre la première méthode beaucoup plus efficace pour l'avantage final de l'étudiant, mais aussi, pensons-nous pouvoir l'affirmer, pour celui de vos propres recherches, acousticiens et neuro-physiologues.

On le voit de façon éclatante par l'influence profonde du professeur de français-langue étrangère Pierre Delattre, chez vous, acousticiens, et en particulier aux Etats-Unis. Mais, bien que notre propre collaboration avec les vôtres ne soit guère encore qu'amorcée, nous y avons déjà trouvé quelque utilité pour la description du comportement du larynx dans la parole (1). Nous avons pu montrer comment réconcilier les points de vue opposés de de P. Lieberman (2) et de R. Husson (3) sur ce point fondamental.

Il est donc raisonnable de penser que vos études comparatives des diverses prononciations, ou plus généralement, les recherches que vous jugeriez bon d'entreprendre dans notre domaine, devraient vous apporter des éclaircissements sur des points inexplicables autrement.

Et voilà donc les deux raisons pour lesquelles je viens vous demander votre aide.

Observations préliminaires.

Mais tout d'abord, pour mieux cerner les problèmes que poserait une éventuelle coopération, il sera utile de voir par quels hasards nous avons été amené à rompre avec nos traditions pédagogiques sur un point où leur inefficacité n'est hélas que trop connue.

Ayant eu l'avantage insigne de pouvoir longuement appliquer les leçons de Pierre Delattre pour l'enseignement du français aux anglophones, nous avons pu, en pratique aussi bien qu'en théorie, constater la supériorité de la méthode consistant à faire acquérir à l'étudiant certains comportements généraux, au lieu de se contenter de chercher à corriger les consonnes et les voyelles, la nature et la position des accents, et certains rythmes et intonations.

Mais, pas plus qu'aucun d'entre nous, Delattre ne pouvait se dégager d'un seul coup de toutes les erreurs sévissant dans nos traditions pédagogiques (4), et nous avons dû chercher longuement avant de trouver une méthode équivalente à celle qu'il employait pour le français, pour enseigner, cette fois, l'anglais aux francophones.

On pourrait penser que, puisque nous savions que la méthode de Delattre est efficace, il nous suffisait d'inverser ses procédés pour faire acquérir une prononciation vraiment anglaise aux francophones. Nous l'avons, en effet, tenté d'abord, en insistant successivement sur les points nous paraissant devoir être les plus efficaces. En vain: ce qui était essentiel pour nous, n'avait guère d'importance dans sa méthode, directement ou inversement. En termes plus parlants pour des ingénieurs: nous avons alors appris que si l'on veut employer des analyseurs pour ces deux enseignements, ces machines devront être complètement différentes.

Et ce fut notre second avantage que d'être amené, par des recherches en vue de l'automatisation de l'enseignement, à isoler l'habitude de prononciation qui suffit à transformer aisément une prononciation française en une prononciation anglaise ne différant pas plus des prononciations britanniques ou américaines que celles-ci ne diffèrent entre elles(5). Pour encourager nos étudiants, nous avons donc affirmé qu'ils pourraient acquérir une prononciation authentiquement anglaise, tout en sachant bien que cette authenticité ne peut être, de nos jours, qu'une notion assez vague. Ce n'était pourtant pas un vain jeu de mots, car nos étudiants voyaient bien qu'ils comprenaient rapidement beaucoup mieux les prononciations des anglophones, ceci étant bien plus facile à contrôler que l'amélioration de la production des sons du langage.

Il y avait donc dans notre succès des éléments de chance, bien courtisée, il est vrai! Nous voyions aussi que Delattre aurait finalement pu atteindre à la cohérence et ne pas rester dans l'incertitude au sujet de ses trois "modes". Quant à nous, nous y étions parvenu, mais pour le point essentiel seulement, et la tâche de clarification à opérer parmi les autres causes diverses nous a longtemps paru insurmontable.

Problèmes à résoudre.

Nous demandons d'abord si un progrès semblable est possible pour l'enseignement des autres langues ou, en d'autres termes, nous aimerions savoir s'il y a ainsi, pour chaque langue étrangère, une seule habitude de prononciation qui puisse être comme une clé pour son acquisition correcte par une autre communauté linguistique. Ceci semble bien improbable, et Delattre faisait d'ailleurs remarquer d'avance que, si c'est possible entre nos deux langues (6), c'est parce qu'elle sont les plus opposées, en Europe, pour leur prononciation!

De toutes façons, en plus de cette clé que nous avons trouvée pour les francophones apprenant l'anglais, d'autres habitudes, moins cruciales mais non négligeables, resteront à acquérir. C'est donc à une recherche très complexe qu'il faut nous attendre et notre coopération devra être sérieuse, large et bien organisée.

Découvrir quels sont ces mécanismes ou habitudes de prononciation à acquérir, et surtout, dans quel ordre ils peuvent ou doivent être enseignées, est resté jusqu'ici un problème aussi insurmontable que lourd de conséquence. Et cela bien que les recherches, publications et enseignements n'aient manqué ni en nombre, ni en diversité; Pour s'en convaincre, d'une façon non scientifique mais hélas trop suffisante, il suffit de constater nos déficiences à nous, francophones, pour les prononciations étrangères, une remarque semblable se faisant aussi couramment pour les anglophones.

Mettant donc à profit notre travail antérieur dans la recherche et dans l'enseignement du français et de l'anglais aux membres de l'autre communauté linguistique, nous venons vous demander de nous aider en pensant que nous pourrions malgré tout obtenir rapidement des enseignements beaucoup plus efficaces. Car si les problèmes restent pratiquement tous irrésolus pour l'instant, nous disposons de nos jours, de moyens d'investigation, de comparaison et d'organisation qui faisaient défaut à nos devanciers.

Et nous ne sommes tout de même pas sans quelque orientation. Il serait vain de prétendre que le programme de recherche que nous vous présentons aujourd'hui est le meilleur possible, ou même simplement réalisable tel quel. Notre intention est autre.

D'abord, oser parler ouvertement de cette tare de nos enseignements, tare dont nous n'avons pas à rougir plus que de beaucoup d'autres, puisque tant d'entre nous faisons tout ce que nous pouvons pour y porter remède. Ensuite, proposer une démarche, et ainsi appeler critiques et suggestions qui permettront peu à peu d'établir un plan de travail réellement efficace, s'il s'avère que celui que nous proposons n'est pas bon.

Préliminaires au plan de recherche.

Nous gagnerons du temps en nous orientant d'abord par un exemple aux leçons assez claires.

Il se trouve que l'habitude de prononciation la plus utile aux francophones pour acquérir une prononciation vraiment anglaise, à savoir: la modulation intra-syllabique est, d'un point de vue très souvent adopté, sans grande importance pour les anglophones eux-mêmes. Ceux-ci peuvent, dans certaines parties très étendues et importantes de leurs activités, parler, même très naturellement et correctement, sans moduler leur voix dans le cours des syllabes accentuées. Car cette habitude compte surtout pour manifester l'émotion et a donc une valeur plus esthétique que sémantique.

Mais on ne peut, non plus, négliger l'aspect émotif dans bien des relations humaines. Si donc on peut s'attendre, en théorie, à ce que la modulation intra-syllabique compte assez peu dans les communications scientifiques et techniques, ou en général, professionnelles, et c'est bien ce qui se produit, elle reste très utilisée dans la conversa-

tion courante, sans parler des productions littéraires, théâtrales, etc. Et il reste bien indéniable que si nous, francophones, acquérons cette capacité, alors nous pouvons aisément prononcer comme eux. Pourtant le fait que son utilisation n'est pas universelle en anglais a puissamment contribué à cacher son importance pédagogique, la faisant ranger par beaucoup parmi les faits esthétiques ou de diction.

Et donc, bien qu'elle soit enseignée par certains de nos collègues, et parfois même avec grand soin, si les élèves de ces professeurs font des progrès particulièrement remarquables en anglais, on peut ne pas voir que c'est à cause de cette modulation intra-syllabique, mais trouver d'autres raisons sur lesquelles d'ailleurs les opinions varieront, en général!

De telles incertitudes et confusions se produisent probablement pour la plupart des langues étrangères et il faudra en tenir grand compte dans nos efforts.

Etapes de la recherche.

D'abord, établir une liste, exhaustive si possible, des diverses options dans les mécanismes et habitudes des prononciations humaines, et, parallèlement, des moyens utilisés pour les inculquer, dans le plus grand nombre possible de pédagogies. En effet, les useges d'une communauté linguistique peuvent être très révélateurs pour ce qui reste caché, tout en étant important, dans une ou plusieurs autres.

Partant de là, et des autres connaissances dont nous disposons, chercher à voir la nature et le degré des liaisons entre ces habitudes, non pas seulement des points de vue acoustique et neurophysiologique, mais aussi pour d'autres raisons, peut-être imprévisibles ainsi, mais que les comparaisons entre les diverses langues, ou d'autres procédures encore, pourront rendre certaines.

Il peut être très utile de savoir, par exemple, que les Japonais ont beaucoup de mal à prononcer certaines suites de consonnes, suites pourtant banales chez nous, ou que les locuteurs du lingala (7) sont désemparés devant les syllabes commençant par une voyelle. Une immense moisson de tels faits peut être recueillie si nous sommes assez nombreux De même, les vérifications par applications précises dans des cas variés, pourront alors être conduites rapidement.

Il deviendra donc assez vite possible de comprendre la raison de l'efficacité de certaines méthodes ou "recettes", et d'en créer de meilleures et de mieux adaptées à l'acquis antérieur de chaque étudiant. On saura mieux comment déterminer celui-ci à coup sûr, grâce à ces vues plus larges, et comment y ajouter ce qu'il faudra.

On sait, par exemple, quelles difficultés les francophones peuvent avoir pour placer correctement l'accent sur un mot anglais, selon les indications du dictionnaire peut-être, ou même, en répétant un mot entendu. Il existe pourtant pour cela une méthode simple et infaillible: dire à l'étudiant de marquer une pause après la syllabe à accentuer (8). Mais fort peu connaissent cette méthode que nous avons déduite de l'accentuation en français, accentuation qui, en France, n'est étudiée que par de rares spécialistes. De même, comme autre exemple, une connaissance même réduite du phonétisme chinois, empêchera de tirer des conclusions hâtives de comparaisons, même relativement larges, entre diverses langues européennes.

Il est donc bien raisonnable de penser que, si nous coopérons en assez grand nombre, les faits révélateurs ne nous feront pas défaut, et grâce à l'informatique, on devrait pouvoir en déterminer les relations de façon précise et efficace.

Il n'est plus interdit d'espérer que bientôt tout homme pourra connaître les lois, peut-être fort simples, qui gouvernent toutes les prononciations humaines et parler, correctement et sans difficulté, un nombre incomparablement plus grand de langues qu'il n'est raisonnable d'escompter de nos jours.

Quelques précisions.

Pour ne pas rester dans le vague, nous indiquerons des habitudes ou mécanismes employés pour la prononciation, mais nous ne les répartirons qu'en deux catégories. En effet, si beaucoup d'autres classifications sont possibles et souhaitables, le souvenir des erreurs commises à ce sujet par nos prédécesseurs, y compris par ce Pierre Delattre que nous admirons tant, suffit à nous interdire d'aller plus avant pour l'instant, sans parler des méprises dont nous-même avons déjà pu nous défaire en cherchant à réduire notre ignorance!

Mouvements et positions articulatoires se rapportant individuellement aux diverses consonnes, voyelles et semi-voyelles:
-Direction dominante, variable ou exclusive, adoptée par l'extrémité de la langue: la pointe de la langue est-elle dirigée vers l'orifice de la bouche, vers le palais antérieur ou, plus ou moins, vers l'arrière de la bouche?
-Utilisation de l'arrondissement des lèvres, ou de leur rétraction, pour des consonnes et voyelles, et à quel degré;
-Par rapport à un système qu'il faudra définir, malgré l'arbitraire inévitable au départ, quelles voyelles ou consonnes sont ajoutées ou retranchées dans les diverses langues? à quel degré ces variations sont-elles liées? (ce point appartient aux deux catégories);
- Quelles sortes de h, l, et r sont permises pour chaque langue? (même remarque).

Habitudes plus "mentales" ou plus générales, c'est-à-dire ne se rapportant pas aux consonnes, voyelles ou semi-voyelles prises séparément, mais plus ou moins, à elles toutes, à la syllabe, ou à tout l'ensemble de la prononciation:
-Ouverture ou fermeture habituelle du larynx au cours de la parole, c'est-à-dire: les cordes vocales sont-elles fermées ou ouvertes quand leur ouverture ou fermeture n'est pas indispensable?
-Les laryngales sont-elles utilisées, et à quel degré (déjà dit)? Les consonnes p, t, k, sont-elles ou non, aspirées? Est-ce une marque distinctive?
-La palatalisation est-elle admise? dans quels cas

et à quels degrés?
-La nasalisation se diffuse-t-elle, ou non, sur les sons avoisinants et à quels degrés?
-La diphtongaison des voyelles existe-t-elle? et à quel degré?
-Les voyelles ou les consonnes sont-elles anticipées?
-L'attaque syllabique sans consonne existe-t-elle? Est-elle liée à des laryngales? Auxquelles?
-La syllabation est-elle ouverte? fermée? mixte?
-Quelles successions de sons sont proscrites?
-L'accentuation se fait-elle? par allongement ou par des changements a/de hauteur soit dans le cours de la syllabe, soit par rapport aux syllabes environnantes (voir ci-après) b/ de l'intensité de la voix c/ autrement (voir juste après)?
-L'accent a-t-il une place fixe dans les mots ou dépend-il du sens ou de la positions dans les groupes de syllabes?
-Quels sont les rythmes prédominants en poésie et en prose?
-Nature de l'évolution de la force vocale au cours de la syllabe: varie-t-elle? comment? à quel degré?
-Quels groupes de muscles sont plus souvent tendus ou souples au cours de la parole? (Il ne s'agit ici que d'un point de vue différant des précédents).
-Les changements de hauteur de la voix au cours des syllabes, en dehors de l'accent, existent-ils? ont-ils une valeur sémantique dans la langue?

etc... ce signe contenant peut-être le plus important.

En tout cas il faut marquer, dans l'inégalité des deux catégories proposées, un indice du fait qu'il faut cesser d'apporter, en pratique aussi bien qu'en théorie, une attention si prédominante aux consonnes et voyelles, alors que le véritable élément fondamental du langage semble bien être la syllabe. Il est urgent que nous en tenions compte.

Puis, sans attendre que les premières listes soient complètes, chercher par des comparaisons entre les diverses langues, aussi bien que par l'acoustique et la physiologie, quelles habitudes entraînent, et avec quel degré de nécessité, quelles autres.

On cherchera aussi quelles habitudes étrangères à une langue donnée sont les plus difficiles à acquérir.

Enfin, si l'on a déterminé, comme nous l'avons fait entre le français et l'anglais, les clés de la prononciation de telle langue étrangère pour telle communauté linguistique, on en verra aussi la probabilité et les limites.

Conclusion.

Bien loin de nous effrayer, le grand nombre de ces possibilités doit, au contraire, être le garant d'un succès assez rapide dans nos efforts. En effet cette multiplicité des options explique et justifie nos échecs passés: sans l'informatique, il ne pouvait être question pour nous d'organiser un savoir si étendu: il fallait trop souvent défaire ce que l'on venait péniblement de bâtir. Désormais au contraire, la coordination entre acousticiens, neuro-physiologues et pédagogues, et des classifications plus pratiques, mettront à la disposition de chacun des ressources plus que suffisantes. Car, pédagogiquement, nous n'avons à tenir compte, pour chaque étudiant, que des capacités qui lui manquent pour la langue en question.

La zone réservée au langage ne peut recevoir les informations que contient déjà en général le reste du cerveau qu'à certaines conditions fort mal connues. Dans un avenir non défini, les biologistes sauront organiser ce tranfert. Nous, pédagogues, ne l'avons pas fait assez bien jusqu'ici, mais désormais, grâce à votre aide, cela peut fort bien nous devenir possible, rapidement et à coup sûr.

La menace, non seulement de dispersion et d'incompréhension mutuelle, mais même de dégradation mentale, qui pèserait encore beaucoup plus sur nos descendants, n'est redoutable que dans la mesure où nous fermons les yeux pour ne pas voir ce que nous avons à faire. De nos jours, nous cherchons et surtout nous enseignons, inévitablement, au hasard; or une erreur enseignée rend notre intervention non seulement nulle, mais négative, et souvent lourdement; ne l'oublions jamais.

Il est grotesque à la fois, d'une part, de se plaindre de la dégradation des études et de la vulgarité et de l'ignorance de plus en plus triomphantes, et d'autre part, de continuer comme par le passé, alors que des moyens presque féériques sont à notre disposition. L'enseignement des langues, si durement décrié, peut prendre, grâce à notre coopération, la tête d'un renouveau aussi merveilleux que de plus en plus indispensable à nos enfants. Et, en outre, vos disciplines propres ne devraient pas manquer d'en tirer un profit plus immédiat encore.

NOTES

1/Cf Compte-Rendu Scientifique, CNET Lannion, août 1977 Annexe 1, page 2 (Note)
2/Lieberman P. Intonation, Perception, and Language p. 15
3/Husson R. C-R Acad. Sc. 1955 pp 241, 242-244
4/On trouvera l'état, encore généralement accepté, de la question, en Principes de phonétique française Middlebury College, Vt, 2ème éd. 1951 par P. Delattre.
5/Cf notre C-R Lannion ci-dessus.
6/Sans être aussi affirmatif que pour l'anglais, nous ne croyons pas nous tromper en disant que, pour les anglophones, l'essentiel est d'acquérir le rythme égal du français, à une cadence assez rapide.
7/Lingala: une des langues parlées au Congo et au Zaïre.
8/Mais il ne faut pas oublier que les dictionnaires de langue anglaise, en général, ignorent superbement les découvertes récentes sur la syllabation...

BIBLIOGRAPHIE

En pratique, elle reste à créer, car à peu près tout est à revoir. Ce sera une des tâches indispensables de la coopération envisagée. On peut se référer, non sélectivement pour l'instant, hélas, aux immenses publications en acoustique de la parole, neuro-physiologie et surtout linguistique, phonétique et pédagogie.

ANALYSE DES PERFORMANCES DU SYSTEME DE RECONNAISSANCE SYRIL 2*

B. Flocon, P. Lockwood, L. Sauter

LABORATOIRES de MARCOUSSIS
CR-C.G.E. - Route de Nozay - 91460 Marcoussis

ABSTRACT

SYRIL 2 is a speaker independent isolated word recognizer based on cepstral analysis, and dynamic time warping algorithm. Clustering techniques applied on a large data base are used during the building of the reference set. In its original version, SYRIL 2 is working with 5 representatives for each word of the vocabulary. 10 parameters are calculated every 16 milliseconds and stored on a 16-bit word. Different tests have been performed in order to see the influence of the reduction of the number of parameters or the number of bit for storing the data. Comparison between different ways for the calculation of the cepstra coefficients have been made. Finally a simple building of the reference set by choosing 5 speakers among 40 has been made : results have been compared with those obtained with the use of a clustering procedure on these 40 representatives.

INTRODUCTION

Les systèmes de reconnaissance de mots isolés multilocuteur donnent à l'heure actuelle des performances suffisamment bonnes pour que la réalisation de maquettes ou prototypes puisse être envisagée. Un certain nombre de contraintes liées à la conception des systèmes de reconnaissance rendent ces maquettes chères. Disposant d'un système de reconnaissance de mots multilocuteur simulé sur ordinateur (SYRIL 2), nous avons pensé qu'il serait intéressant de chercher à réduire le "coût" de certains éléments de notre système. Dans un premier temps, les éléments que nous avons cherché à simplifier concernent, d'une part, la partie "analyse du signal", et d'autre part, le principe utilisé pour la création de l'ensemble de références.

Les résultats présentés montrent que certaines simplifications n'entrainent pas de dégradation et ne remettent donc pas en cause les performances du système. D'autres, par contre, peuvent être jugées inacceptables car elles entrainent trop de différences avec le système original SYRIL 2.

*Etude partiellement financée par un contrat ESPRIT.

LE SYSTEME SYRIL 2

Ce système a été développé à partir de SYRIL [1]. Il s'agit d'un système de reconnaissance de mots isolés multilocuteur. Le vocabulaire disponible sur SYRIL 2 comporte 130 mots sans hiérarchisation du vocabulaire ; ces 130 mots correspondent aux commandes d'une machine de traitement de textes ; ils n'ont pas été particulièrement choisis et comportent certains groupes de mots voisins ("placer", "classer" ; "mot", "non" ...). Afin de resituer SYRIL, nous donnons brièvement une description des modules qui le composent :

1. Le module d'extraction d'extraction des paramètres de SYRIL et SYRIL 2 est effectué sur un array processor qui réalise en temps réel le calcul de 10 paramètres toutes les 16 millisecondes, après filtrage passe-bas à 4000 Hz, lissage par une fenêtre de Hamming de largeur 32 millisecondes et codage sur 12 bits. Les 10 paramètres utilisés dans SYRIL 2 sont 9 coefficients du cepstre (MFCC) calculés à l'aide de l'échelle Mel [2] et un paramètre représentant la puissance du signal (nous utilisons les MFCC à partir de MFCC(1) ; il n'y a donc pas directement de terme représentant l'énergie dans les 9 MFCC que nous utilisons).

2. L'ensemble de référence de SYRIL 2 comporte 5 références pour chaque mot du vocabulaire. Ces références ont été obtenues après traitement de 40 élocutions de chaque mot du vocabulaire provenant de 40 locuteurs différents (20 hommes et 20 femmes). Un algorithme de classification [1] utilisant un seuil adaptatif est utilisé pour déterminer 10 classes à l'aide des 40 éléments ; seules les 5 classes les plus représentatives sont conservées. Un algorithme de calcul de moyenne au sein d'une classe est appliqué afin d'obtenir un représentant unique pour chaque classe.

3. L'algorithme de reconnaissance de SYRIL 2 est basé sur la programmation dynamique [3] ; la règle du plus proche voisin est appliquée afin de décider quel est le mot reconnu. Dans la version de SYRIL 2 utilisée dans les expériences qui vont être décrites, aucun seuil de rejet ou de proximité n'est utilisé.

Les données utilisées dans SYRIL 2 sont stockées sur 16 bits.

MODIFICATIONS PAR RAPPORT AU SYSTEME INITIAL

Nous avons réalisé un certain nombre de modifications afin de tester les possibilités de simplifications qui pourraient être apportées au système SYRIL, sans entraîner une dégradation trop importante des performances. Ces modifications ont été effectuées à trois niveaux :

1. Modification sur le calcul des paramètres d'analyse

. calcul des paramètres cepstraux à l'aide de la prédiction linéaire (utilisation des LPCC : Linear Prediction Cepstrum Coefficients).

. utilisation d'un banc de filtres simulés pour calculer les coefficients cepstraux, au lieu d'une transformée de Fourier suivie d'un filtrage. Des bancs de 19 et 16 filtres ont été simulés ; les sorties de ces filtres ont été moyennées deux à deux, de manière à effectuer un lissage comparable à celui obtenu lors de l'utilisation de la transformée de Fourier.

2. Réduction de la quantité d'information représentant l'ensemble des données : deux orientations ont été testées :

. calcul de 5 paramètres pour représenter le signal toutes les 16 millisecondes (4 MFCC et un paramètre de puissance)

. compression des données sur 8 bits au lieu de 16 (après recentrage de la puissance) ; les essais de compression ont été effectués sur 5 et 10 paramètres.

3. Utilisation de 5 locuteurs pris au hasard pour constituer l'ensemble de références : au lieu d'enregistrer 40 personnes et d'effectuer les traitements de classification et de moyennage décrits plus haut, nous avons pris aléatoirement 5 locuteurs et nous avons utilisé leurs élocutions pour constituer l'ensemble de références, sans effectuer de traitement particulier sur les données correspondantes.

VOCABULAIRES DE TEST

Les 3 séries de tests décrits ci-dessus ont été réalisées sur différents types de vocabulaire. Dans tous les cas, l'ensemble de références est constitué de 5 références pour les 130 mots du vocabulaire de traitement de texte.

1. Premier vocabulaire de test : il s'agit d'utiliser 57 élocutions des 130 mots en tant que base de test. Cela représente une base de données de 130 x 57 = 7410 mots. Les performances obtenues correspondent à des performances d'un système fonctionnant pour une centaine de mots ; la difficulté de ce vocabulaire peut être considérée comme moyenne.

2. Second vocabulaire de test : le test de reconnaissance est effectué en comparant les 10 chiffres (0, 1, ... 9) aux 130 mots précédemment cités. Là encore 57 locuteurs ont été testés. La base de test comporte donc 10 x 57 = 570 mots. Le fait de comparer ces 10 mots aux 130 ne présume pas des performances réelles d'un système qui reconnaîtrait uniquement les 10 chiffres ; mais ce test permet de se placer dans des conditions où le nombre d'erreurs est suffisamment important pour que l'on puisse mesurer l'impact d'une modification au niveau de la détérioration des performances du système.

3. Troisième vocabulaire de test : un sous-vocabulaire de 18 mots dont la liste est donnée dans la figure 1 a été extrait des 130 mots. Ce sous-vocabulaire comporte un certain nombre de mots présentant plus de confusions que la plupart des autres mots. La figure 1 présente les mots avec lesquels sont confondus le plus fréquemment ces 18 mots. Ce 3è test consiste également à comparer ces 18 mots aux 130. La base de test comporte 18 x 57 locuteurs, soit 1026 mots. La liste de ces mots est donnée dans la figure 1.

La composition des vocabulaires de tests est rappelée dans la figure 5.

MODE	
MATHEMATIQUE	AUTOMATIQUE
MULTIPLIER	
MOT	NON
MODIFIER	
NUMERO	ZERO
NOUVEAU	
NON	MOT
OPERER	
OPTION	
OUI	
POSITION	
POURCENTAGE	
PAGE	
PRECEDENT	
PARTAGER	
PETIT	
PLACER	CLASSER

Figure 1 : Liste des mots du 3ème vocabulaire de test avec, en face, les confusions les plus fréquentes.

RESULTATS

Nous présenterons les résultats en différents tableaux. Dans chaque tableau, nous présenterons en première ligne le système de références (SYRIL 2) pour les 3 vocabulaires de tests définis plus haut. Les résultats sont donnés en nombres d'erreurs avec, entre parenthèses, l'augmentation du taux d'erreurs en pourcentage par rapport au système de référence (1ère ligne).

Une analyse de chacun des tests est faite après la présentation de chaque table de résultats.

- **Influence du type d'analyse**

	Vocab. 1	Vocab. 2	Vocab. 3
MFCC 256 pts	175	37	18
LPCC 256 pts	344 (+97%)	51 (+38%)	19
19 filt.	213 (+22%)	31 (-16%)	27 (+50%)
16 filt.	230 (+31%)	31 (-16%)	31 (+72%)

Figure 2 : résultats suivant le type d'analyse

Les résultats ci-dessus montrent que l'utilisation d'un banc de 19 filtres ou de 16 filtres pour calculer les MFCC au lieu d'une transformée de Fourier dégrade légèrement les performances. Par contre, le calcul des coefficients cepstraux par prédiction linéaire fait pratiquement doubler le nombre d'erreurs. Ceci semble confirmer la supériorité des MFCC [4] par rapport à d'autres types de paramètres.

- **Influence du nombre de bits et du nombre de paramètres utilisés pour la représentation des données**

énergie +	Vocab. 1	Vocab. 2	Vocab. 3
9 MFCC (16 bits)	175	37	18
9 MFCC (8 bits)	183 (+5%)	35 (-5%)	19 (+5%)
4 MFCC (16 bits)	235 (+26%)	49 (+32%)	27 (+50%)
4 MFCC (8 bits)	574 (+228%)	131 (+254%)	55 (+205 %)

Figure 3 : résultats suivant le nombre de bits et/ou paramètres

On voit clairement apparaître, sur les résultats ci-dessus, le fait que l'utilisation de 8 bits suffit largement pour la représentation des 10 paramètres utilisés dans le système de reconnaissance. Le fait de n'utiliser que 5 paramètres, au lieu de 10, dégrade légèrement les performances ; par contre, le cumul des deux opérations (réduction du nombre de paramètres et du nombre de bits) apporte une dégradation très importante.

- **Influence de la méthode utilisée pour choisir les représentants**

Pour réaliser ces tests, nous avons effectué successivement 4 tirages aléatoires de 5 locuteurs parmi les 40 dont nous disposions. Les tests ont été effectués pour les 3 vocabulaires définis précédemment.

	Vocab. 1	Vocab. 2	Vocab. 3
5 classes	175	37	18
tirage 1	307 (+75%)	71 (+92%)	34 (+89%)
tirage 2	351 (+100%)	67 (+81%)	34 (+89%)
tirage 3	341 (+95%)	44 (+19%)	34 (+89%)
tirage 4	354 (+100%)	68 (+84%)	27 (+50%)

Figure 4 : résultats suivant la création des références

Le fait d'utiliser des références calculées par moyennage dans 5 classes obtenues après classification automatique appliquée à un ensemble de 40 éléments permet d'obtenir 2 fois moins d'erreurs que lorsque l'on prend aléatoirement 5 locuteurs pour créer l'ensemble de références. Rappelons que 175 erreurs pour le vocabulaire 1 correspondent à un taux d'erreurs de 2,3 %.

vocabulaire numéro	nombre mots	nombre locuteurs	nombre mots test
1	130	57	7410
2	10	57	570
3	18	57	1026

Figure 5 : vocabulaires de test

CONCLUSION

Comme cela a été dit plus haut, il ne s'agit pas ici de donner des performances du système SYRIL 2 dans son mode de fonctionnement normal. Le but est de comparer l'influence de certaines simplifications ou réductions de la complexité de SYRIL 2 sur les performances du système. Il en ressort un certain nombre d'informations intéressantes à différents niveaux : coût de la création de l'ensemble de références (multiplication du nombre d'erreurs par 2 si l'on n'utilise pas la classification), coût du stockage des données de l'ensemble de références (peu de changements en diminuant par 2 la quantité à stocker, mais forte dégradation quand on divise par 4), coût du calcul des paramètres (diminution des performances si on préfère un banc de 19 filtres à une transformée de Fourier), coût de la prédiction linéaire (diminution des performances par rapport aux résultats obtenus par la FFT et le filtrage selon l'échelle Mel).

Tous ces tests montrent bien que les systèmes SYRIL et SYRIL 2 doivent une part non négligeable de leurs performances à certains éléments qui les rendent "chers". Bien évidemment, dans le cas d'une application pratique, on cherchera à réduire les coûts ; mais un compromis coût/performance devra être choisi ; nous pensons que les résultats ci-dessus pourront jouer un rôle dans ce choix.

REFERENCES

[1] B. FLOCON, N. BRIANT
"SYRIL : Système temps réel de reconnaissance de mots isolés indépendant du locuteur", 4ème congrès AFCET RFIA, Paris 1984.

[2] S.B. DAVIS, P. MERMELSTEIN, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences", : IEEE-ASSP 28 n° 4, Août 1980.

[3] H. SAKOE, S. CHIBA, "Dynamic programming optimization for spoken word recognition", IEEE - ASSP 26 n° 1, Février 1978.

[4] C. GAGNOULET, M. COUVRAT, "SERAPHINE : a connected word speech recognition system", ICASSP - vol 2, p. 887, Paris, 1982.

****
* LISTE DES AUTEURS DE COMMUNICATIONS *
****