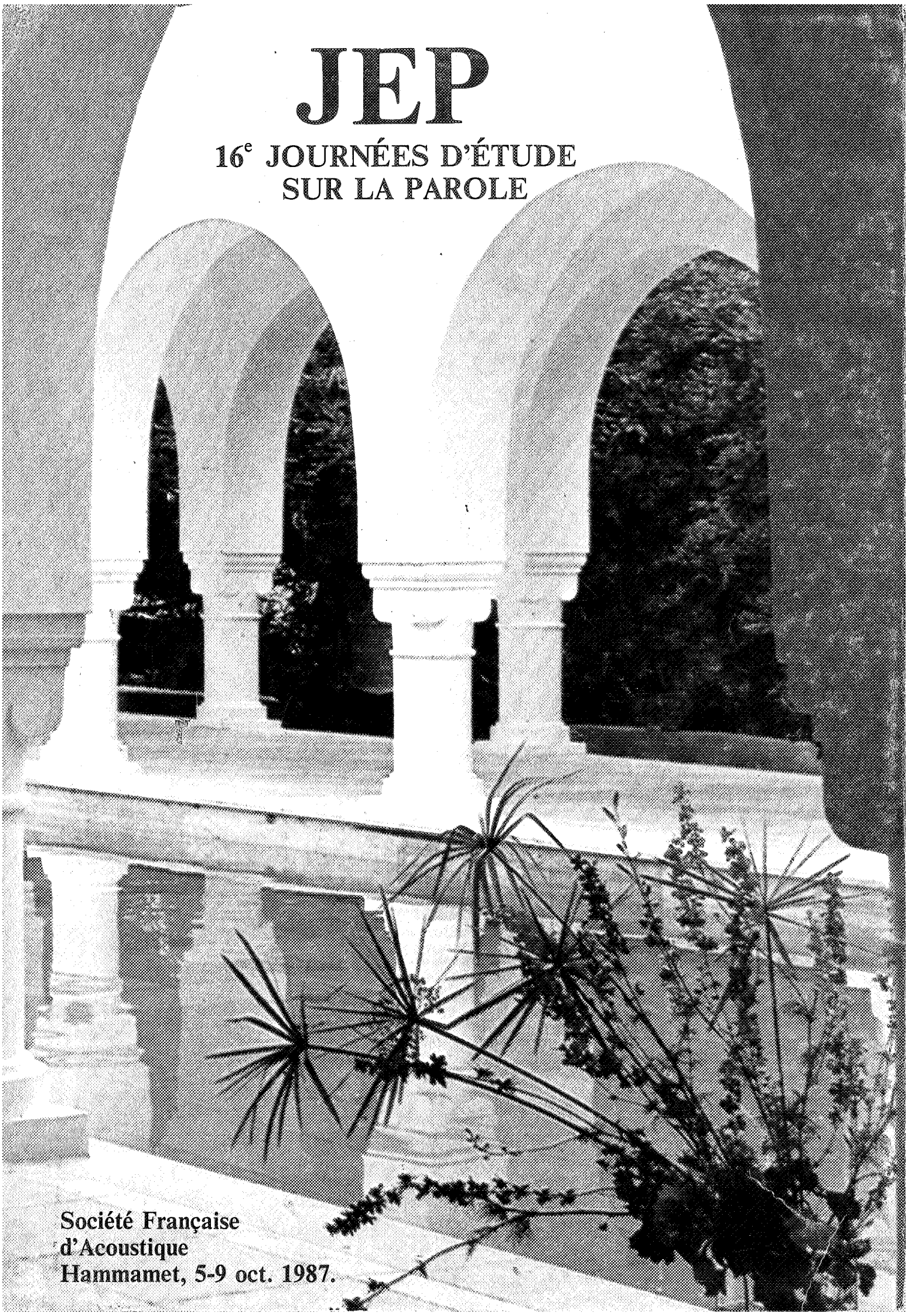
JEP

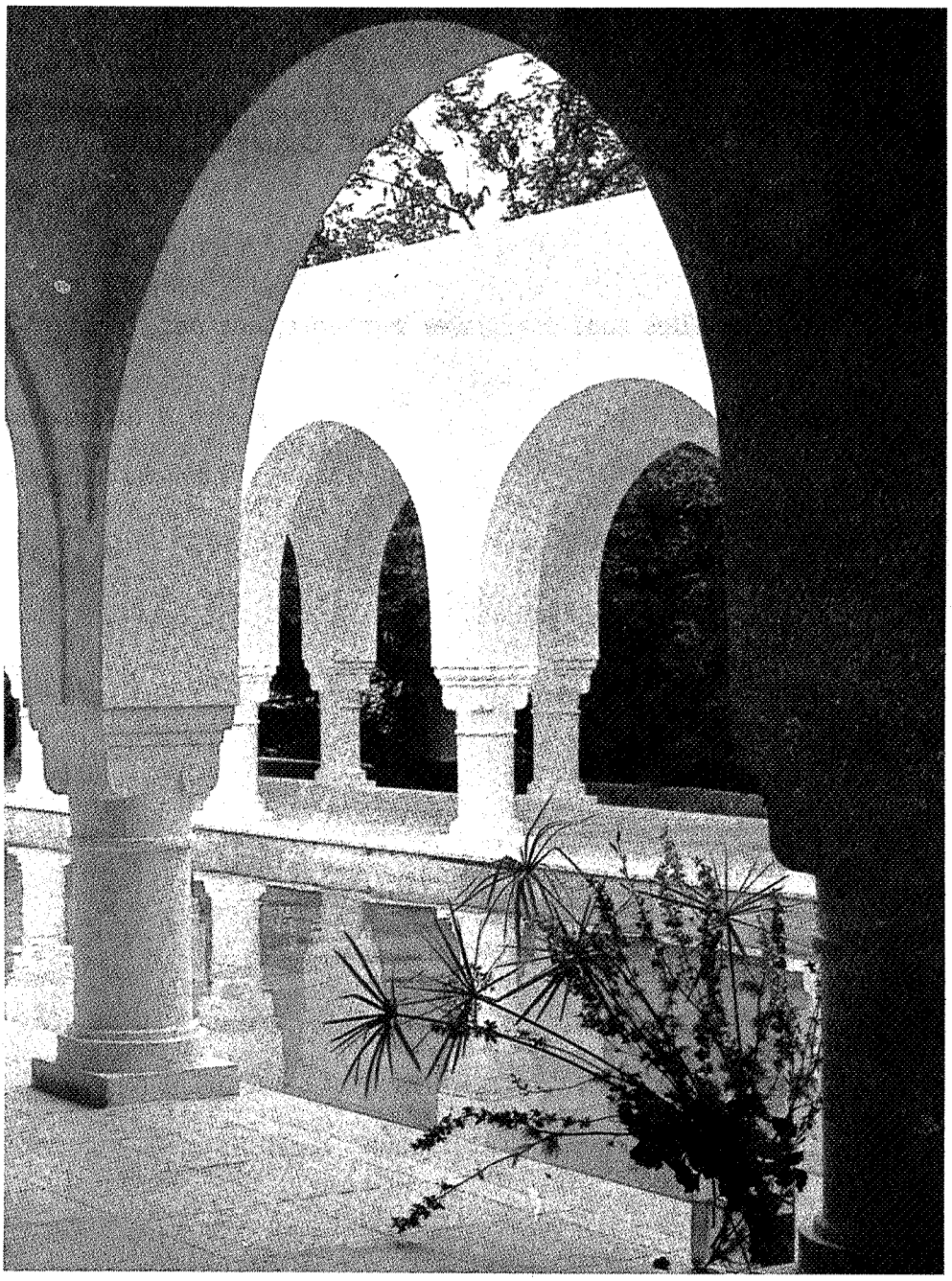
16e JOURNÉES D'ÉTUDE SUR LA PAROLE



Société Française
d'Acoustique
Hammamet, 5-9 oct. 1987.

JEP

16e JOURNÉES D'ÉTUDE SUR LA PAROLE



Société Française
d'Acoustique
Hammamet, 5-9 oct. 1987.

Société Française d'Acoustique SFA
Groupe Communication Parlée GCP

16e JEP

Les 16e Journées d'Etude sur la Parole du groupe Communication Parlée de la Société Française d'Acoustique ont lieu au Centre Culturel d'Hammamet (Tunisie) du 5 au 9 octobre 1987.

Elles sont organisées conjointement par

L'Ecole Nationale d'Ingénieurs de Tunis
ENIT (Tunis)

L'Institut Régional des Sciences Informatiques et des Télécommunications
IRSIT (Tunis)

L'Institut Bourguiba des Langues Vivantes
IBLV (Tunis)

Le Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
LIMSI-CNRS (Orsay)

Comité d'organisation

Pr N. ELLOUZE	**Dr J.S. LIENARD**
Pr S. GHAZALI	**Dr M. ESKENAZI**
Dr B. ZOUABI	**Mme M. CHASTAGNER**
	Mme F. NEEL

Secrétariat SFA : CNET Lannion A, B.P. 40, 22301 LANNION Cedex

Secrétariat GCP : ENST dépt SYC, 46 rue Barrault, 75634 PARIS Cedex 13

Communications aux 16 èmes JEP

I. ANALYSE - SYNTHESE - TRANSMISSION

- transmission -

- synthèse -

II. DECODAGE ACOUSTICO-PHONETIQUE / IA

- IA -

- décodage -

III. SYSTEMES VOCALIQUES

- général -

- F'2 -

IV. PHYSIOLOGIE

- production -

- implants cochléaires -

V. RECONNAISSANCE ET DIALOGUE

- reconnaissance -

- dialogue -

VI. SEGMENTATION ET ETIQUETAGE

VII. PROSODIE / DUREE

- durée -

- prosodie général -

Journée Franco-Arabe

JOURNEE EVALUATION ET BASES DE DONNEES

- évaluation -

- bases de données -

ANALYSE - SYNTHESE
TRANSMISSION

FONCTIONS D'ONDE FORMATIQUE : EXTRACTION DES PARAMETRES ET SYNTHESE VOCALE

*C. d'Alessandro *^, X. Rodet **^*

* LIMSI - CNRS : BP 30 F-91406 ORSAY Cedex
^ LAFORIA Univesité Paris 6 : 4, Place Jussieu F-75005 PARIS
** IRCAM : 31, rue Saint-Merri F-75004 PARIS

ABSTRACT

For many years Formant-Wave-Function synthesis has successfully been used in Musical research. We present results in speech synthesis using Formant-Wave-Function to implement a parallel formant synthesiser. Then, a new method for representing the speech signal as a set of Formant-Wave-Functions is described. The main steps of the analysis/synthesis process are (for each window), LPC modelisation, Formant tracking, filter bank definition using formant parameters, filtering in formant regions with overlap-add short-time Fourier analysis/synthesis, detection of elementary waveforms and modelisation with Formant-Wave-Functions. This method autorizes both spectral and temporal precision, with manipulation of perceptually pertinent parameters.

INTRODUCTION

La synthèse par Fonctions d'Ondes Formantiques (ou FOF) est utilisée avec succès depuis plusieurs années, en particulier pour la recherche musicale (synthèse de voix chantée, de timbres instrumentaux ou imaginaires)(1). Il s'agit de reconstruire - dans le domaine temporel - le signal voulu à l'aide de fonctions élémentaires, les FOF, ayant des caractéristiques spectrales intéressantes.

Plusieurs auteurs se sont par ailleurs attachés à la décomposition du signal de parole en fonctions d'ondes : citons en particulier J.S. Liénard avec l'analyse impulsionelle (2) ainsi que J. Morlet, A. Grossman (et al.)(3) qui ont fourni un cadre mathématique rigoureux pour l'analyse d'un signal sur une base de fonctions élémentaires ou "ondelettes".

Après un rappel de la définition d'une FOF et un exposé des résultats ainsi obtenus en synthèse de parole, nous présentons une méthode d'analyse permettant la représentation du signal vocal par un ensemble de FOF.

DEFINITION DES FOF

Une FOF est une fonction temporelle dont le spectre possède un maximum (qui sera - par analogie avec les maxima de la fonction de transfert du conduit vocal - dénommé ici "formant"). Les FOF peuvent être calculées à l'aide d'une formule mathématique explicite ou bien être extraites de parole réelle (4). La figure 1 présente un exemple de FOF communément utilisée : la réponse d'un résonateur du second ordre à une impulsion (approximativement en forme d'arche). On peut ainsi définir différentes FOF qui possèdent comme expression analytique le produit d'une sinusoïde et d'une fonction d'enveloppe temporelle (exponentielle décroissante, fenêtre d'analyse spectrale (5)...) dont la forme induit les propriétés d'enveloppe spectrale.

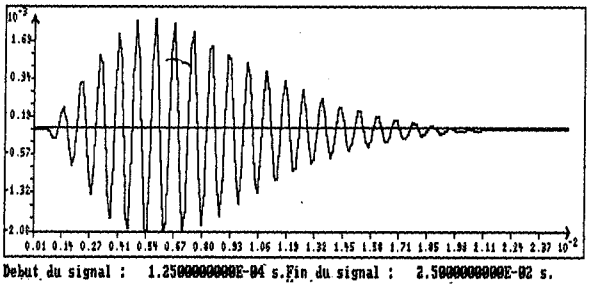


Figure 1a : exemple de FOF : le signal temporel.

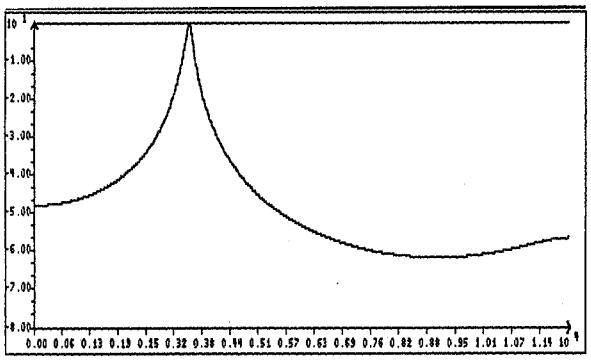


Figure 1b : exemple de FOF : le spectre d'amplitude logarithmique.

L'utilisation de FOF présente plusieurs avantages:

- les paramètres utilisés sont perceptivement pertinents (ce qui explique le succès de la méthode en synthèse musicale).
- on peut obtenir une grande précision temporelle et simultanément une grande définition spectrale.
- la méthode se prête bien à une implantation sur de petites machines (un synthétiseur FOF temps réel a été développé sur microprocesseur TMS32010) (6).

SYNTHESE

La production de diverses sortes de signaux (de parole ou issus d'instruments musicaux par exemple) peut être représentée sous la forme d'une fonction d'excitation et d'un filtre. Ainsi les FOF ont été d'abord utilisées pour simuler un synthétiseur à formant en parallèle excité par un train d'impulsions quasi-périodiques (synthèse de la partie vocalique de la parole, de voix chantée...) (7).

Pour peu que les paramètres (paramètres formantiques, énergie, pitch) fournis au synthétiseur soient correctement estimés, on obtient une qualité de synthèse comparable à un bon codage par prédiction linéaire classique, utilisant un nombre important de coefficients (une trentaine de pôles), en conservant toutefois plein accès aux paramètres acoustiques explicites du signal synthétisé. Dans l'exemple suivant nous avons utilisé le système de détection de formants développé à l'I.R.C.A.M. par P. Depalle (8) et

transcrit par C. Laura (joint à une détection de pitch).

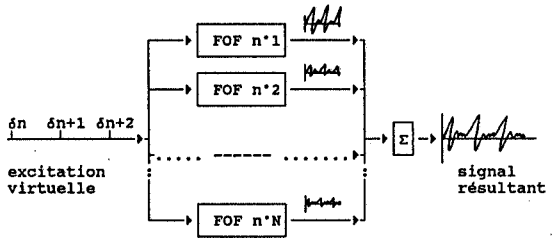


Figure 2 : structure d'un synthétiseur FOF parallèle.

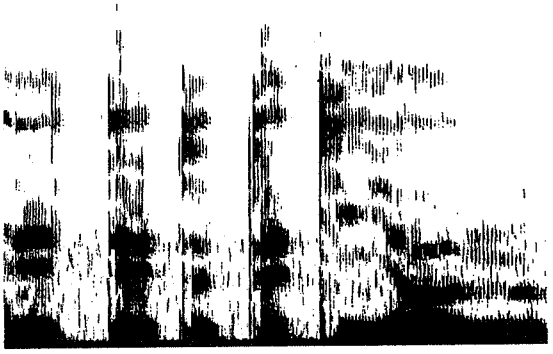


Figure 3 : sonagramme de la phrase "les têtes tatillonnes" parole naturelle.

Figure 4 : sonagramme de la phrase "les têtes tatillonnes" synthèse par FOF.

Il est également possible de simuler la sortie d'un filtre excité par un signal aléatoire (et non plus par un train d'impulsions quasi-périodiques), en générant des FOF de façon aléatoire (de l'ordre d'une FOF par milliseconde en moyenne) : on peut ainsi synthétiser des fricatives, par exemple.

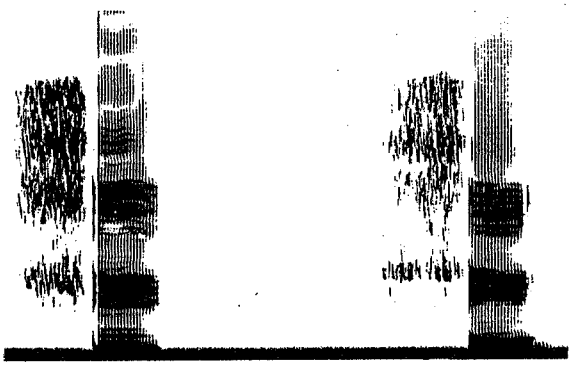


Figure 5 : sonagramme de la syllabe /fy/ naturelle puis synthétique (estimation manuelle des paramètres à partir du son naturel).

Pour obtenir une synthèse de bonne qualité il semble bien sûr nécessaire de dépasser - quant à l'excitation - l'opposition voisé / non voisé, tant pour les segments de parole qui relèvent directement d'un modèle de production "mixte" (un bruit fricatif se superposant aux vibrations quasi-périodiques des cordes vocales) comme les consonnes fricatives voisées, que pour les segments de parole qui paraissent "bien voisés" : en effet d'une part tous les degrés de mélange (entre signaux quasi-périodiques et signaux aléatoires) coexistent dans la parole naturelle - de la voix chuchotée à la voix théâtrale -, d'autre part, jusque dans les voyelles chez certains locuteurs on constate plusieurs excitations du conduit vocal au sein du même cycle vocalique (par exemple à l'ouverture et à la fermeture des cordes vocales). Le succès du codage par prédiction linéaire multi-impulsionnelle (9) est à cet égard révélateur.

ANALYSE-SYNTHESE

Nous ne cherchons plus ici à modéliser le signal comme étant issu du filtrage d'une certaine source d'excitation (démarche developpée à partir - comme nous le verrons - d'un noyau d'analyse commun par X. Rodet, P. Depalle et G. Poirot (10)), mais (comme le propose J.S. Liénard(11)), à le représenter par un ensemble de fonctions d'ondes élémentaires. Ces fonctions d'ondes, formant une concentration d'énergie dans une région spectro-temporelle, peuvent néanmoins être modélisées comme la réponse d'un certain filtre à une certaine excitation (filtre et excitation en quelque sorte "locaux" du point de vue spectro-temporel). L'analyse parcourt les étapes suivantes :

1. Modélisation du signal par prédiction linéaire. Nous utilisons un filtre d'analyse en treillis (12), (cette méthode d'analyse se trouve également à la base des travaux de X. Rodet, P. Depalle et G. poirot), pour obtenir une estimation (non synchrone au pitch) de l'envellope spectrale. La fenêtre d'analyse est choisie assez longue pour englober une période de voisement.

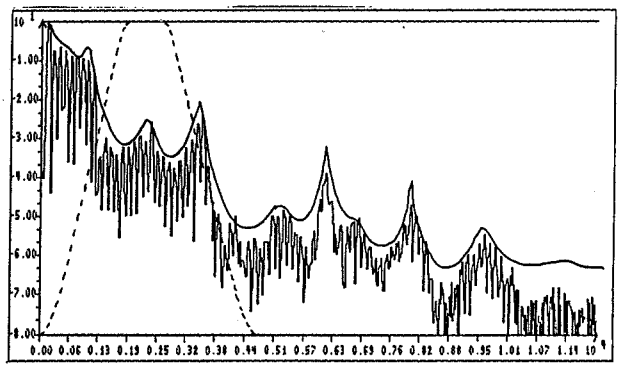


Figure 6 : spectre d'amplitude logarithmique d'un /a/ superposé à son enveloppe spectrale modélisé par prédiction linéaire.On à également représenté le gain du filtre dans la troisième région "formantique".

2. Détection des maxima spectraux, sur chaque fenêtre d'analyse (par suivi de la courbe spectrale d'après des programmes de M.J. Caraty). On acquiert ainsi la fréquence centrale, l'amplitude et la largeur de bande de chaque formant (au sens défini plus haut).

3. Définition pour chaque fenêtre d'un jeu de filtres passe-bande, dont le gain est centré sur les fréquences centrales des formants, plats dans une bande définie grâce aux largeurs de bande des formants, et doucement amorti jusqu'à zéro en arche de sinusoïde. La méthode de filtrage que nous avons choisi autorise en effet la définition de filtres de gains quelconques. La somme des gains de ce jeu de filtres est bien sûr partout égale à l'unité.

4. Filtrage dans chaque région spectrale définie par le jeu de filtres précédent, et dans la région temporelle définie par l'instant d'analyse (avec bien sur une marge temporelle suffisante autour de cet instant pour la détection qui va suivre). Le filtrage est effectué par analyse-synthèse de Fourier à court terme (13), ce qui préserve les caractéristiques de phase des signaux.

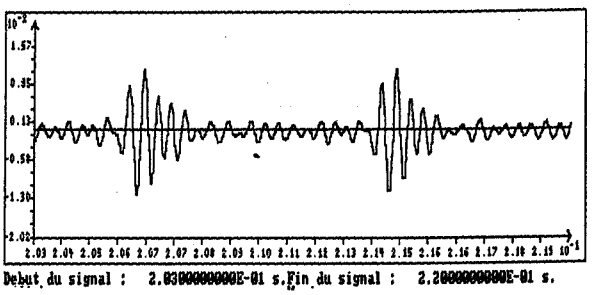


Figure 8 : signal issu du filtre précédent.

5. Détection des FOF dans chaque région spectro-temporelle, par corrélation, ou bien par filtrage inverse ce qui donne un signal local d'excitation.

6. Jusqu'ici le système est rigoureusement additif, c'est-à-dire que l'on reconstitue le signal original par sommation des diverses FOF "naturelles" détectées. Nous pouvons aussi constituer un signal synthétique grâce à une modélisation (telle que celles présentées plus haut) des FOF par une famille de fonctions mathématiques simples, ce qui autorise à nouveau la manipulation des caractéristiques acoustiques du signal. Ce point est en cours de développement au moment d'écrire cet article.

CONCLUSIONS

Notre système de décomposition d'un signal de parole en fonctions d'ondes élémentaires utilise la connaissance a priori des régions de maximum spectral (nous exploitons ainsi cette particularité du signal de parole, qui est de posséder un spectre "à formants").

Un tel système permet de garder automatiquement une grande précision fréquentielle (due à la qualité de l'analyse par prédiction linéaire) mais aussi de ne pas délaisser les évolutions temporelles rapides, par le filtrage et la détection qui suit (même en cas de mauvaise estimation spectrale pour, par exemple, une plosive, l'aspect temporel du phénomène reste accessible).

Notre méthode offre ainsi à la fois un mode d'analyse-synthèse automatique (codage) et un mode de synthèse à partir de paramètres acoustiques (formants, pitch...).

REFERENCES

(1) Rodet,X.,Potard,Y.,Barrière,J.B.,1980 : The Chant project : from the synthesis of the singing voice to synthesis in general. Computer Music Journal vol.8 No3.

(2) Liènard,J.S.,1985 : Analyse à très court terme de la parole : un outil et quelques directions de recherche. 14ièmes JEP PARIS.

(3) Goupillaud,P.,Grossmann,A.,Morlet,J., 1985 : Cycle-octave and related transforms in seismic signal analysis. Geoexploration 23.

(4) Rodet,X.,Delatre,J.L.,Ražzam,M. 1979 : Construction du signal vocal dans le domaine temporel. 10ièmes JEP GRENOBLE.

(5) Harris,F. 1978 : On the use of windows for harmonic analysis with the discrete Fourier transform. Proceedings of the IEEE vol.66 No1.

(6) Déchelle,F.,d'Alessandro,C. 1984 : Rapports de DEA TAI. Université Paris 6.

(7) Rodet,X. 1980 : Time domain Formant-Wave-Function synthesis. Spoken language generation and understanding J.C. Simon ed, D. Reidel publishing Compagny, Dordrecht Holland.

(8) Rodet,X.,Depalle,P. 1985 : Synthesis by rule : LPC diphones and calculation of formant trajectories. IEEE ICASSP.

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

(10) Rodet,X.,Depalle,P.,Poirot,G. 1987 : Analyse et synthèse de voix parlée et chantée par modélisation de l'enveloppe spectrale et de l'excitation. 16ièmes JEP HAMMAMET.

(11) Liènard,J.S. 1987 : Speech analysis and reconstruction using short-time, elementary waveforms. IEEE ICASSP.

(12) Makhoul,J.,Lynn,K. 1981 : Adaptative lattice analysis of speech. IEEE ICASSP.

(13) Crochiere,R.E. 1979 : A weighted overlap-add method of short-time Fourier analysis/synthesis. IEEE ICASSP.

SUPPRESSION DES OSCILLATIONS NON LINEAIRES DANS LES SYSTEMES APC PAR TRAITEMENT INTER-TRAME

G. FENG (1) & P. COMBESCURE (2)

(1) Institut de Phonétique de Grenoble
Institut de la Communication Parlée
Université III
38400 St-Martin-d'Hères

(2) C.N.E.T. Lannion A
Département TSS/CMC
Route de Trégastel
22301 Lannion

ABSTRACT

In this paper, we present several techniques that we used to cope with nonlinear oscillations in APC systems. The main idea is to reduce discontinuities between two successive frames, since they produce strong impulses in the system, which can provoke oscillations. Three approaches will be presented. They concern determination of prediction coefficients, initialization of the lattice filter/synthesizer, and diminution of the prediction error discontinuity between the frames. These techniques have been proved efficient in eliminating the oscillations and in improving the coder's performance.

INTRODUCTION

Les codeurs APC (codage prédictif adaptatif) sont souvent utilisés pour la transmission du signal de parole à moyen et à bas débit (20 à 10 kbit/s). Ces codeurs peuvent offrir une bonne qualité pour le signal codé tout en n'introduisant pas de grande complexité. Une structure commune des systèmes APC consiste à mettre le quantificateur dans les boucles de prédiction, aplatissant ainsi le spectre du bruit de codage et augmentant le rapport signal sur bruit /1,2,3/.

Cependant, cette insertion du quantificateur dans les boucles de prédiction constitue un système non linéaire bouclé dont la stabilité n'est pas toujours assurée, même si le prédicteur linéaire est lui-même stable. En fait, l'instabilité d'un codeur APC est une source importante de la dégradation de qualité. Krasner et al. ont les premiers étudié ce problème de la stabilité des systèmes APC en fonction du gain du prédicteur /4/.

Dans notre étude précédente /5/, ayant d'abord relié l'instabilité des systèmes APC à l'existence d'oscillations non linéaires, nous avons étudié particulièrement le mécanisme de leur formation. A l'aide de la méthode du premier harmonique /6,7/, nous avons obtenu des conditions d'oscillations en fonction des coefficients de prédiction à court terme et, en particulier en fonction du degré d'inclinaison du spectre du prédicteur. Nous avons également proposé une méthode de modification des coefficients afin de stabiliser le système. Ces approches se sont montrées efficaces pour éliminer les principaux modes d'oscillation non linéaire qui sont de type sinusoïdal, souvent de forte amplitude et de longue durée ; leur présence est relativement indépendante du signal d'entrée.

Comme pour tous les systèmes non linéaires, la stabilité dépend, non seulement du système, mais aussi du signal d'entrée ; les oscillations non linéaires dans un système APC peuvent se manifester de différentes façons. A part les oscillations dites "stationnaires", on observe parfois des oscillations irrégulières, de forte amplitude mais de courte durée, ou des oscillations ayant une forme d'onde complexe, etc.

Il est encore difficile d'expliquer l'origine exacte du mécanisme de ces oscillations, et de trouver des conditions analytiques pour prévoir leur existence. Par ailleurs, la structure complexe d'un système réel (l'utilisation d'un quantificateur à plusieurs bits, d'une boucle de prédiction à long terme, etc.) augmente la difficulté de l'analyse.

Nous avons pourtant constaté que la présence de ces oscillations est plus ou moins liée aux discontinuïtés des paramètres entre les trames successives. Comme on utilise en général des trames de longueur fixe dans les codeurs à complexité limitée, ces discontinuïtés peuvent parfois être très grandes dans les zones où le signal n'est pas stationnaire. Ces discontinuïtés peuvent provoquer des signaux impulsionnels dans le système lors du changement de trames, et ceci influence sa stabilité puisque celle-ci dépend du signal d'entrée.

Ainsi, dans la mesure où l'on utilise des trames de longueur fixe, il semble nécessaire d'introduire un traitement inter-trame pour réduire les discontinuïtés et éliminer les oscillations non linéaires. Nous allons présenter dans cette communication trois procédures que nous avons mises à l'épreuve.

DETERMINATION DES COEFFICIENTS DE PREDICTION

Typiquement un système APC se réduit à un ensemble codeur/décodeur (Fig. 1).

Dans la partie codeur, la boucle APC est constituée par le quantificateur et deux prédicteurs. Le prédicteur à court terme a une fonction de transfert :

$$P(z) = -\sum_{i=1}^{p} a_i \cdot z^{-i}, \qquad (1)$$

et le prédicteur à long terme peut se présenter par :

$$P_M(z) = \beta_1 z^{-M+1} + \beta_2 z^{-M} + \beta_3 z^{-M-1}, \qquad (2)$$

ici, M représente un délai relativement long qui correspond, en général, à la valeur de la période fondamentale.

Pour chaque trame d'analyse (16 ms), on détermine d'abord les coefficients de prédiction a_i, β_i et le pas de quantification à partir du signal de parole. Ensuite, la boucle entre en fonction et génère le signal résiduel quantifié pour la transmission.

Le fonctionnement de la boucle dépend ainsi totalement de ces paramètres. Un "mauvais" jeu de coefficients peut être à l'origine d'oscillations /5/.

Compte tenu des influences entre les trames successives dues aux mémoires des prédicteurs, la détermination des coefficients de prédiction doit respecter les deux exigences suivantes :

- rendre compte de l'influence de la trame précédente,
- s'adapter au fonctionnement de la boucle.

Nous avons ainsi déterminé les coefficients de prédiction de la façon suivante.

Pour les coefficients de prédiction à court terme de la trame /0,N-1/, on utilise les données sur l'intervalle /-p,N-1/ pour minimiser les erreurs de prédiction sur l'intervalle /0,N-1/. On applique donc le critère des moindres carrés qui correspond dans ce cas au calcul des coefficients LPC par la méthode de covariance /8/.

Comme les coefficients calculés avec ce critère ne permettent pas d'obtenir un filtre toujours stable, on utilise en pratique la structure en treillis pour estimer les coefficients PARCOR Ki, ce qui conduit toujours à un filtre stable. Dans le calcul, l'intervalle des données est /-p,N-1/ et celui de la sommation effectuée pour le calcul des Ki est /0,N-1/.

Notons que l'utilisation d'une fenêtre n'est pas nécessaire dans ce cas, puisque l'intervalle des données est suffisamment long par rapport à l'intervalle de minimisation des erreurs ; il ne faudrait d'ailleurs pas en utiliser puisque les erreurs de prédiction doivent correspondre au signal non pondéré.

Dans la boucle APC, l'entrée du prédicteur à court terme provenant du signal différencié u' du décodeur local (Fig. 1), il contient le bruit de quantification. Les coefficients de prédiction optimaux devraient être calculés à partir du signal u'. Or, on ne connaît pas u' lors de la détermination des coefficients. Calculés à partir du signal u_0 qui ne contient pas de bruit de quantification, les coefficients sont légèrement sub-optimaux. On retient ces coefficients parceque l'on a guère le choix ; on estime d'autre part que la différence entre u et u' est relativement petite.

En fait, le signal u' de la trame précédente est disponible ; il est donc logique de l'utiliser pour l'intervalle /-p,-1/. Les coefficients calculés de cette façon rendent compte au mieux des données de la trame précédente et de la situation réelle de la boucle. Ceci est particulièrement utile dans le cas où la différence entre u et u' de la trame précédente n'est pas négligeable. En pratique cette méthode s'est révélée efficace pour diminuer la discontinuité entre deux trames successives.

Les mêmes principes s'appliquent au calcul des coefficients de prédiction à long terme. Une fois que la valeur M est déterminée, on utilise les données sur l'intervalle /-M-1,N-1/ pour minimiser les erreurs de prédiction sur l'intervalle /0,N-1/. Cette minimisation conduit à un ensemble d'équations des moindres carrés classiques, et leur résolution se réduit à une inversion de matrice.

Comme pour les coefficients à court terme, on utilise ici le signal re-synthétisé s' du décodeur local pour l'intervalle /-M-1,-1/ (trame précédente), tandis que le signal s se situe dans l'intervalle /0,N-1/ (trame courante). Compte tenu de la valeur relativement grande de M (M > 20 en général), une partie importante du signal s' de la trame précédente participe au calcul des coefficients. Ceci permet d'obtenir une meilleure évaluation.

INITIALISATION DES MEMOIRES DU PREDICTEUR ET DU SYNTETISEUR EN TREILLIS

Dans la boucle APC, la sortie du prédicteur à court terme $\tilde{u}$ peut s'écrire :

$$\tilde{u}_k = -\sum_{i=1}^{p} a_i \, u'_{k-i} \quad , \qquad (3)$$

ici, a_i sont les coefficients de prédiction et u'_k représente le signal différencié venu du décodeur local (cf. Fig. 1).

Avec la détermination des coefficients de prédiction que nous avons adoptée, quand on commence une nouvelle trame (à l'instant k=0), ce sont les p derniers échantillons u'_{-i} (i=1,...,p) de la trame précédente que le prédicteur doit utiliser.

Supposons maintenant qu'on réalise le prédicteur par une structure directe (filtre tranversal), dans laquelle il y a une série de registres (un "FiFo") contenant les données u'_{k-i} (i=1,...,p). Avec l'arrivée d'une nouvelle trame, ces registres contiennent exactement les p derniers échantillons de la trame précédente. Au chaque changement de trame, il n'est donc pas nécessaire de modifier les contenues des mémoires du prédicteur.

Cependant ceci n'est plus le cas avec un prédicteur en treillis.

Dans un prédicteur en treillis, le calcul de $\tilde{u}_k$ à l'instant k s'effectue en deux étapes. On effectue d'abord un calcul de filtre sur u'_{k-1} :

$$b^0_{k-1} = x^0_{k-1} = u'_{k-1}$$

$$\begin{cases} b^j_{k-1} = b^{j-1}_{k-2} - K_j \cdot x^{j-1}_{k-1} \\ x^j_{k-1} = x^{j-1}_{k-1} - K_j \cdot b^{j-1}_{k-2} \end{cases} \text{pour } j=1,\ldots,p \quad (4)$$

Ceci a uniquement pour but de déterminer l'état des mémoires :

$$b^{j-1}_{k-1} \quad , \; j = 1,\ldots,p \qquad (5)$$

Ensuite, on calcule $\tilde{u}_k$:

$$\tilde{u}_k = \sum_{j=1}^{p} K_j \cdot b^{j-1}_{k-1} \, . \qquad (6)$$

Lors de l'arrivée d'une nouvelle trame, les coefficients Ki changent. Une question se pose donc : est-il possible de continuer à calculer $\tilde{u}_k$ en changeant uniquement les Ki, mais sans modifier les mémoires b_{k-1} ?

Pour répondre cette question, nous allons représenter les b_{k-1} sous forme suivante (obtenue par un développement de (4)):

$$b_{k-1}^{i-1} = \sum_{i=1}^{i} f_j^i(K) \cdot u'_{k-i}, \quad j = 1,\ldots,p \quad (7)$$

ici $f_j^i(K)$ représente une fonction des coefficients Ki, par exemple, $f_2^1 = 1$, $f_2^2 = -K_1$, etc...

On voit clairement qu'à l'instant d'arrivée d'une nouvelle trame, les mémoires b_{k-1} contiennent une combinaison des p échantillons u'_i (i=1,...,p) (ce qu'il faut) et les coefficients de prédiction de la trame précédente (ce qu'il ne faut pas).

Avec le principe de détermination des coefficients que nous avons adopté, on a besoin, à cet instant, des mémoires $-b_{k-1}$ qui contiennent la combinaison de u'_i (i=1,...,p) et les coefficients de la trame courante. Pour réaliser ceci, on doit réinitialiser les mémoires du prédicteur lorsqu'on introduit une nouvelle trame. Il s'agit donc de recalculer le filtrage à partir de u'_{k-p} jusqu'à u'_{k-1} par (4); tout en utilisant les coefficients de la trame courante.

Cette réinitialisation des mémoires s'applique également au synthétiseur en treillis du décodeur. Mais, le calcul des mémoires doit s'effectuer à l'aide d'un filtre auxiliaire compte tenu d'une légère différence entre le synthétiseur et le prédicteur.

A PROPOS DE LA DISCONTINUITE DANS LES ERREURS DE PREDICTION

Selon le principe de la détermination des coefficients l'erreur de prédiction est minimisée sur l'intervalle /0,N-1/. Cependant, la continuïté des erreurs de prédiction à la frontière de deux trames successives n'est pas toujours assurée.

Si le système ne contient qu'un seul prédicteur, la discontinuïté des erreurs de prédictions ne pose pas de problème particulier. En effet, les erreurs de prédiction de part et d'autre de la frontière s'adaptent respectivement aux pas de quantification, qui sont déjà déterminés en fonction des écarts-types des erreurs de prédiction.

Pour un système qui contient deux prédicteurs, l'erreur de prédiction du premier prédicteur est le signal d'entrée du deuxième prédicteur. Une forte discontinuïté de la première erreur risque d'entraîner une forte impulsion dans la deuxième. Comme les fortes impulsions ne sont pas correctement transmises par le quantificateur, elles risquent d'introduire des "clics" dans le signal codé, ou même des oscillations ce qui est plus grave. Ces effets sont manifestes dans les zones de non-stationnarité.

Il est pourtant difficile de résoudre complètement ce problème parcequ'il n'est pas possible de prendre en compte les trames précédente et suivante lors du calcul des coefficients.

Nous avons trouvé une méthode simple pour contourner cette difficulté. Comme la discontinuïté se produit surtout à l'arrivée de chaque trame, les fortes impulsions apparaîssent souvent dans le premier échantillon de l'erreur de prédiction. Il est donc logique de quantifier plus précisémment (sur plus de bit) cet échantillons. Les simulations nous ont montré qu'un ajout d'un seul bit pour cet échantillon suffit à améliorer le résultat. Dans le cas d'un quantificateur à un bit, une quantification à un niveau de 5σ pour toutes les impulsions dont l'amplitude dépasse 4σ donne des résultats assez satisfaisants.

CONCLUSION

Dans cette communication, nous avons proposé trois améliorations visant à réduire les discontinuïtés entre deux trames successives. Notre expérience a montré qu'une détermination des coefficients de prédiction prenant mieux en compte l'influence de la trame précédente permet de diminuer les discontinuïtés. La réinitialisation des mémoires du prédicteur et du synthétiseur en treillis assure une bonne cohérence entre le fonctionnement de la boucle et la détermination des coefficients. Enfin, l'introduction d'un bit supplémentaire atténue la discontinuïté des erreurs de prédiction. Tous ces procédés permettent de réduire les oscillations du sytème et d'augmenter les performance des codeurs.

Nous pensons que les procédures présentées ici sont utilisable pour réduire les discontinuïtés de tout système traitant des trames de longueur fixe.

REFERENCES

/1/ ATAL B.S. & SCHROEDER M.R., Adaptive predictive coding of speech signals. - Bell Syst. Tech. J., vol.49, pp.1973-1986, 1970.

/2/ MAKHOUL J. & BEROUTI M., Adaptive noise spectral shaping and entropy coding in predictive coding of speech. - IEEE Trans. on ASSP, ASSP-27, pp.63-73, 1979.

/3/ ATAL B.S., Predictive coding of speech at low bit rates. - IEEE Trans. on Comm., COM-30, pp.600-614, 1982.

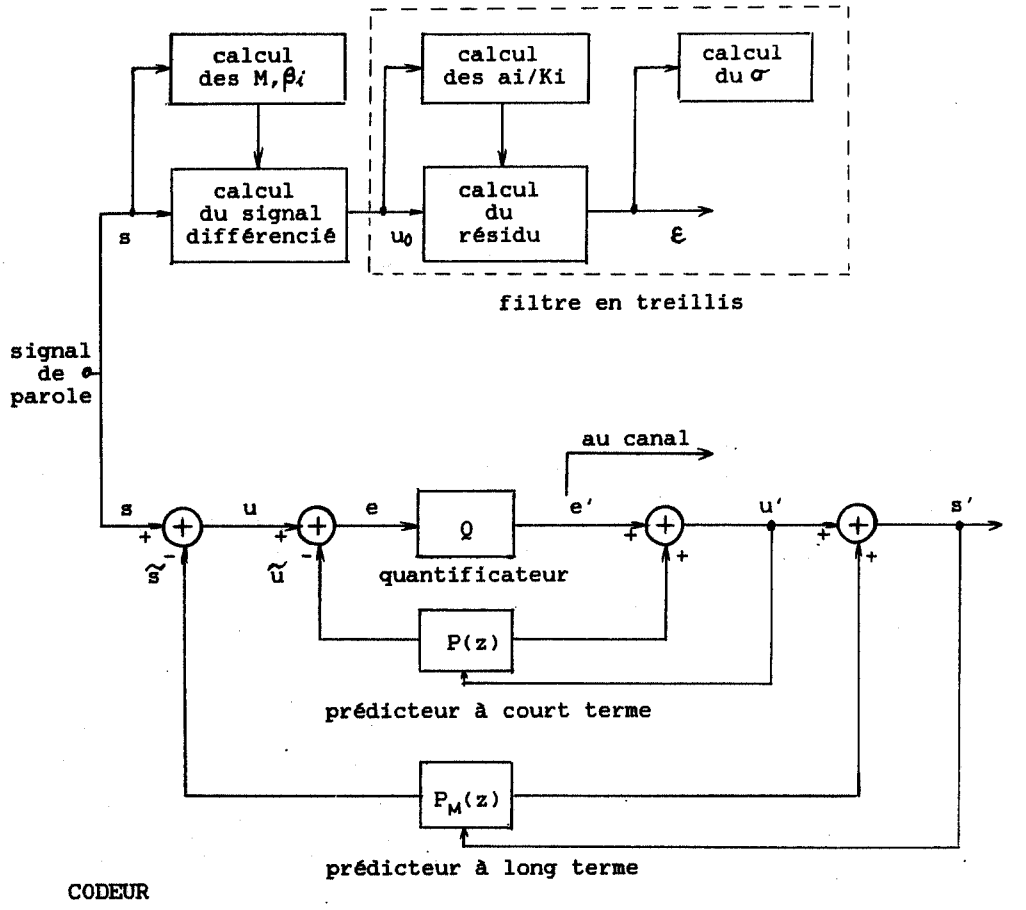
/4/ KRASNER M., BEROUTI M. & MAKHOUL J., Stability analysis of APC systems. - Proc. IEEE ICASSP, pp.627-630, 1981.

/5/ FENG G. & COMBESCURE P., Nonlinear oscillations in APC systems and their prevention. - Int. Conf. on Comm. Tech., Nanjing, China, (to appear), 1987.

/6/ MILLER R.K., MICHEL A.N. & KRENZ G.S., On the stability of limit cycles in nonliear feedback systems : analysis using describing functions. - IEEE Trans. on Cir. and Syst., CAS-30, pp.684-696, 1983.

/7/ MEES A.I. & BERGEN A.R., Describing functions revisited. - IEEE Trans. on Auto. Control, AC-20, pp.473-478, 1975.

/8/ MARKEL J.D. & GRAY A.H., Linear prediction of speech. - Springer-Verlag, Berlin, Heidelberg, New York, 1976.



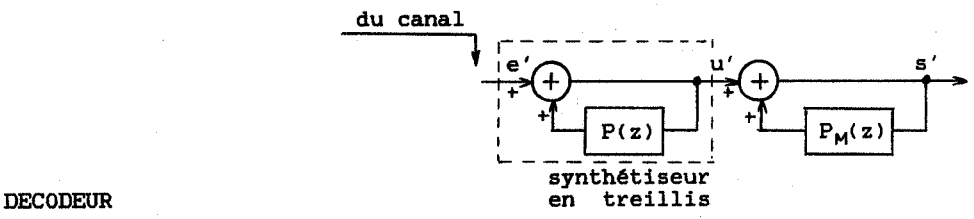


Figure 1. Structure d'un système APC.

- s - signal de parole.
- u_0 - signal différencié, c'est-à-dire, débarassé de la redondance à long terme (calcul des paramètres).
- ε - résidu LPC.
- u - signal différencié (boucle APC).
- e - signal résiduel.
- e' - signal résiduel quantifié.
- u' - signal différencié re-synthétisé (décodeur local/lointain).
- s' - signal re-synthétisé.
- $\tilde{u}$ - prédiction du signal différencié (sortie du prédicteur à court terme).
- $\tilde{s}$ - prédiction du signal de parole (sortie du prédicteur à long terme).

DE LA PRECISION D'UNE METHODE DE VARIATION DU DEBIT DE PAROLE

R. ESPESSER

Institut de Phonétique, UA 261 CNRS,
29, av. Robert-Schuman, 13621 Aix-en-Provence Cedex

ABSTRACT

A technique for modifying speech rate has recently been proposed (3), which is much simpler than the usual phase vocoder method. The implementation is first described, which minimises the temporal deviations inherent to the method. In order to estimate the temporal displacements (particularly for time variable speech rate modifications), the RMS energy is calculated on the modified signal and the RMS energy of the original signal is then resampled. The absolute difference, the correlation coefficient between the two series then obtained, and their graphic display are used as cues to measure temporal and energetical precision.

INTRODUCTION

Il est courant, lors d'études sur la prosodie, de chercher à modifier le débit de parole en préservant les autres paramètres (Fo, spectre...). Deux outils classiques réalisent cette modification : le vocoder à prédiction linéaire et le vocodeur à phase (1, 2). Le premier a quelques difficultés avec des Fo un peu élevés (voix de femmes) et donne une synthèse de qualité variable; le second donne une synthèse de bonne qualité, mais est très coûteux en (moyens de) calcul. Plus récemment une méthode temporelle a été proposée (3), peu exigeante en calcul, tout en donnant une synthèse de haute qualité : elle est l'objet de la présente étude.

Outre la qualité de synthèse, les études prosodiques ont des exigences de précision; on entend ici précision temporelle : la précision des trois techniques mentionnées quant à la conservation des caractéristiques spectrales et autres est déjà assurée. La plupart des études sur ces techniques de variation du débit (ou, plus généralement, de variation de l'échelle temporelle, VET) prennent comme exemple-test des modifications importantes : coefficient de 0.5, 2, 3 etc., portant sur des séquences de plus de 1 s. Tout autre est la situation phonétique typique : dans un mot et phrase porteurs, variation d'une ou deux voyelles, soit pour un cas déjà favorable, une zone de 100 ms, portée de 60 à 180 ms par pas de 20 ms, ceci aux fins d'expérience de perception. Le respect de cette finesse de modification est donc crucial pour qu'une technique de VET soit applicable en phonétique. C'est ce que nous aborderons dans la troisième partie, précédée d'un rappel de la méthode, puis d'une description de l'implantation - et de ses quelques particularités - qui en ont été faites.

I. - RAPPEL DE LA METHODE "SOLA" (Synchronised overlapping-and-add)

Le signal d'entrée est lu par blocs de d ms, toutes les pl ms (1/pl : fréquence de lecture) et fenêtré; chaque fenêtre est additionnée à la queue de la séquence déjà synthétisée à une fréquence d'écriture 1/pe (pe : pas d'écriture) différente, par une méthode "overlapping-and-add" (OLA); k, le facteur de VET, est défini par :

$$k = \frac{pe}{pl}$$

Pour $k > 1$, le débit est ralenti, pour $k < 1$, accéléré. Le point où s'applique la procédure OLA est appelé point de soudure (PS) (voir Fig. 1).

Un PS défini seulement par k (PS "brut" : PSB) ne respecterait évidemment pas la cohérence de phase entre deux fenêtres successives originales. La méthode SOLA consiste à rechercher un PS optimal (PSO) autour du PSB; cet optimum est le maximum de la fonction d'intercorrélation (dans le domaine temporel) des deux séquences à souder; les fenêtres sont ainsi dites synchrones. On remarque que l'on n'est plus maître du coefficient de VET, le PSO étant en général différent du PSB : on a (voir Fig. 2) :

$$PSB - 1 < PSO < PSB + 1$$

avec PSB : valeur temporelle du PSB
PSO : valeur temporelle du PSO
2 1 : plage de calcul de la fonction d'intercorrélation.

Cette méthode permet de traiter des VET variant dans le temps (et non pas restant constant sur toute la séquence à traiter).

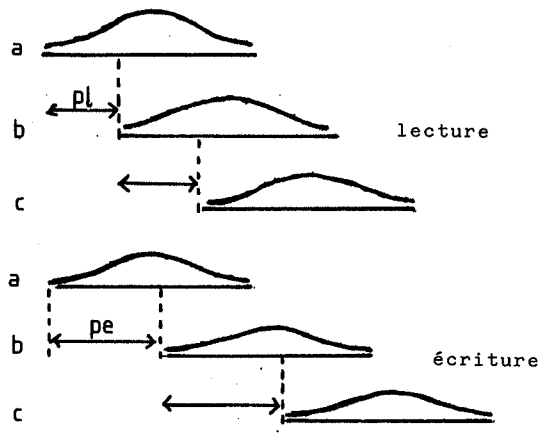


Figure 1. - Procédure OLA.

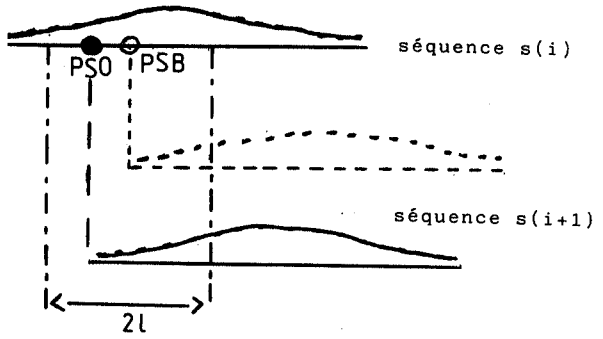


Figure 2. - Procédure SOLA.

II. - IMPLANTATION, PARTICULARITES

- Pour faciliter le traitement de VET en f(t), nous avons pris pe constant, fixé à 12 ms, (pl varie donc en pe/k). La fenêtre d vaut 32 ms; on utilise une fenêtre de Hamming.
- On a vu que le principe de la méthode induisait une erreur temporelle e, écart entre le PSO et le PSB, avec

$$|e| < l$$

l : plage de calcul de l'intercorrélation; l vaut 5.5 ms, soit une demi-période à 90 Hz; en effet, e correspond généralement à un alignement sur une période des deux séquences.

Pour des Fo élevés (>200 à 300 Hz), la fonction d'intercorrélation a plusieurs sommets sur une plage de 11 ms. Au lieu de prendre le maximum de cette fonction, on prend le sommet (> 2/3 du maximum, pour s'assurer de sa "qualité") le plus proche du PSB; e est ainsi minimisé.

A chaque soudure, e est calculé, pour correction sur la soudure suivante : soit PSB(i) la soudure brute des séquences s(i) et s(i+1), g la fonction de calcul de PSO(i) :

$$PSO(i) := g(s(i), s(i+1), PSB(i))$$
$$e(i) := PSB(i) - PSO(i)$$

pour les séquences s(i+1), s(i+2), il vient :

$$PSB(i+1) := PSB(i+1) + e(i)$$
$$PSO(i+1) := g(s(i+1), s(i+2), PSB(i+1))$$
$$e(i+1) := PSB(i+1) - PSO(i+1)$$

(:= symbolise l'assignation des langages classiques de programmation).

- Les valeurs retenues de pe, d et l paraissent être un triplet satisfaisant (optimal ?) pour les sons de parole. Elles autorisent une dynamique de VET de 0.4 à 4 environ, tout en respectant certaines contraintes (par ex., pour k = 0.4, pl = 30 ms, l'algorithme de soudure rapproche ces deux séquences, quasi contiguës à l'origine, de 18 ± 5.5 ms; compte tenu de la stationnarité du signal de parole, k = 0.4 paraît être une limite).
- Les VET en f(t) sont réalisés simplement en rafraîchissant à une certaine fréquence le coefficient k; l'évolution de k est calculée préalablement (à partir de points cibles par exemple) et constitue un fichier pilote; le temps de référence de ce fichier pilote est celui de la séquence d'origine. k n'est pas tenu de varier par paliers, mais peut varier "continuement".

La synthèse obtenue est excellente, en particulier la "qualité du naturel" est totalement préservée, pour des locuteurs masculins ou féminins.

III. - PRECISION TEMPORELLE

On a cherché à voir si la procédure SOLA n'entraînait pas des erreurs temporelles excessives, particulièrement pour des VET en f(t).

La méthode directe, mais manuelle, de mesure est vite apparue fastidieuse et imprécise : les zones de transition entre phonèmes sont elles-mêmes dilatées (ou contractées), et la précision de la segmentation manuelle illusoire.

On a donc tout d'abord vérifié rapidement (sur quelques mots et quelques locuteurs) la précision temporelle et énergétique pour des VET constants dans le temps. L'énergie RMS moyenne est conservée à mieux de 1 %, l'erreur temporelle maximale est de 3 %, ceci pour k variant de 0.4 à 3.5 par pas de 0.1. Sur ces bases, on a appliqué la méthode indirecte suivante : l'énergie RMS est calculée sur le signal d'origine, puis rééchantillonnée (par interpolation linéaire) selon le même fichier pilote que celui utilisé pour le VET; cette série temporelle est alors comparable terme à terme avec l'énergie RMS du signal modifié (la fenêtre de calcul est de 8 ms, le pas de calcul et celui du fichier pilote de 2 ms). Le coefficient de corrélation et l'écart absolu moyen servent d'indices de mesure de l'écart entre les deux séries. Le décalage temporel maximal est relevé sur leur superposition graphique.

Le corpus d'évaluation comprend deux phrases, numérisées à 10 kHz sur 16 bits (/la pipe de Jean s'est cassée en tombant de la poche de ta gabardine /; /la fille de Charles Sablon a voulu un petit chien en guise de cadeau/), prononcées par 3 femmes et 3 hommes. Le k(t) du VET vaut x (x ≠ 1) sur 77 ms toutes les 400 ms, et 1 le reste du temps, ceci jusqu'à la fin de la phrase (soit environ de 6 à 8 plages de variation par phrase). Cinq valeurs de x ont été testées : 0.4, 0.85, 1.65, 3.25, et 1 à titre de contrôle. La Fig. 3 donne l'écart absolu (moyenne sur 2 fois 6 phrases) par valeur de x, et la Fig. 4 la corrélation entre les deux séries d'énergie . L'énergie RMS moyenne du signal original varie de 60 à 65 dB.

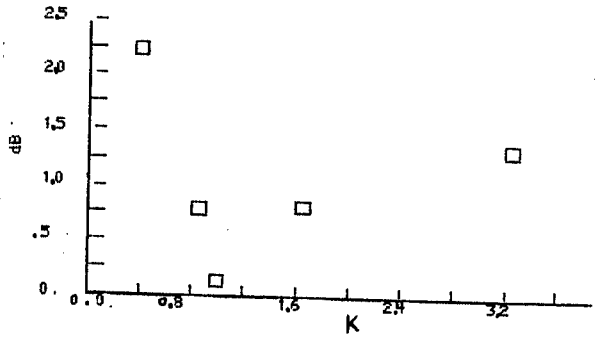


Figure 3. - Ecart absolu moyen pour k = 0.4, 0.85, 1, 1.65, 3.25

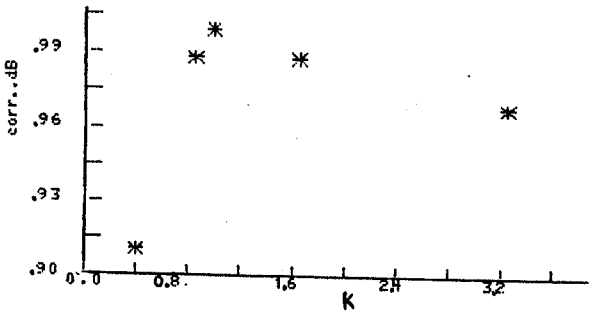


Figure 4. - Corrélation pour k = 0.4, 0.85, 1, 1.65, 3.25

Plus k est différent de 1, plus les deux séries diffèrent. Sur la Fig. 5, on a porté l'écart absolu moyen (sur 6 phrases) en différenciant locutrices (▽) et locuteurs (+) : l'erreur est plus forte chez les hommes, leur période moyenne étant plus importante. L'énergie comme indice de précision a au moins le mérite d'une certaine cohérence. Le décalage temporel maximal, mesuré sur la superposition graphique des deux séries, est d'environ 16 ms; il a été trouvé sur des valeurs de k extrémales (0.4 ou 3.25).

Deux jeux d'essais voisins de k(t) ont donné des résultats très similaires.

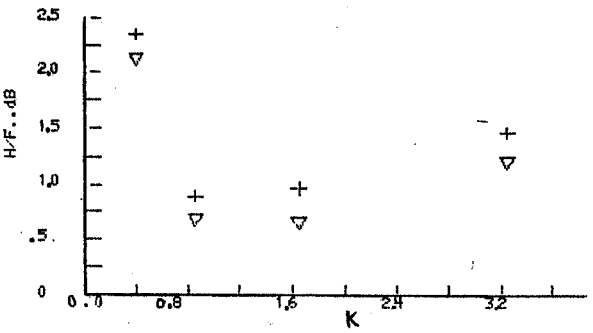


Figure 5. - Ecart absolu moyen; ▽ : femme, + : homme.

CONCLUSION

Les résultats obtenus, satisfaisants, incitent néanmoins à une certaine prudence : la méthode d'évaluation n'est pas infaillible, et la précision atteinte est le minimum demandé pour certains travaux. Mais il paraît difficile de faire mieux, dès lors que des fenêtres de 32 ms, des pas de 12 ms sont en jeu. Une précision accrue relève plutôt de l'édition de signal (couper/coller des périodes). La méthode SOLA, dans l'implantation ici décrite, ne paraît pas elle-même en cause.

REFERENCES

(1) M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis", **IEEE Trans. Acoust. Speech-Signal Processing**, vol. ASSP-29, Nr. 3, June 1981, p. 374-390.

(2) S. Seneff, "A system to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", **IEEE Trans. Acoust. Speech-Signal Processing**, vol. ASSP-30, Nr. 4, August 1982, p. 566-578.

(3) S. Roucos & A. M. Wilgus, "High quality time-scale modification for speech", **Proc. ICASSP**, March 1985, p. 493-496.

VOCODEUR TEMPS REEL POUR LA TRANSMISSION DE PAROLE A FAIBLE DEBIT

M. LEVER, I. LECOMTE, L. LELIEVRE, A. TASSY

Groupe MATRA COMMUNICATION
rue J.P. Timbaud 78390 BOIS D'ARCY

Abstract

The purpose of this paper is to present a real time vocoder at a bit rate lower than 6 Kbps implemented on a single TMS32010 digital signal processor. We describe the speech coding algorithm and the hardware implementation. Then we expose its performances which allow its use in narrowband digital radio communications.

Introduction

Dans le domaine récent de la communication numérique, le besoin de transmettre la parole avec une bonne qualité à moins de 6 Kbits/s devient de plus en plus sensible.

Or il s'agit d'une gamme de codeurs intermédiaires entre les vocodeurs à 2400 bits/s, produisant une qualité "synthétique", et les techniques hybrides (ex: codage de la parole par prédiction linéaire à excitation multi-impulsionnelle ou MPLPC [1]), permettant une très bonne restitution au delà de 9600 bits/s. Ces dernières ont, grâce à l'utilisation généralisée de la quantification vectorielle [2,3], vu leur débit décroître au prix d'une augmentation considérable de leur complexité. Malgré des efforts récents de simplification [4], de telles méthodes semblent, pour l'instant, peu compatibles avec une réalisation matérielle à bon marché.

La qualité de parole reproduite et la complexité ne sont pas les seuls critères de choix d'un algorithme de codage. En transmission, la qualité de phonie ne doit pas être trop affectée par la présence de bruit de fond à l'émission et doit pouvoir être préservée malgré les erreurs dues au canal.

Le vocodeur que nous présentons dans cet article répond à la recherche d'un compromis entre ces différentes contraintes pour la transmission à faible débit. Il utilise une technique MPLPC [1], qui a pu être réalisée en temps réel à l'aide d'un processeur de signal TMS32010.

Après une description de cette technique de codage, des solutions retenues pour palier aux imperfections du canal et de la réalisation matérielle, nous discuterons les choix effectués et les résultats obtenus lors d'essais en transmission sur voie radio à bande étroite pour laquelle la méthode MPLPC s'est avérée d'un intérêt tout particulier.

Méthode de codage de la parole

Le principe du codage par prédiction linéaire à excitation multi-impulsionnelle est de représenter un segment de résidu LPC par quelques échantillons non nuls caractérisés par leurs positions et amplitudes sur le segment. Ces "impulsions" sont utilisées à la synthèse pour exciter le filtre tout pôle (cf figure n° 1).

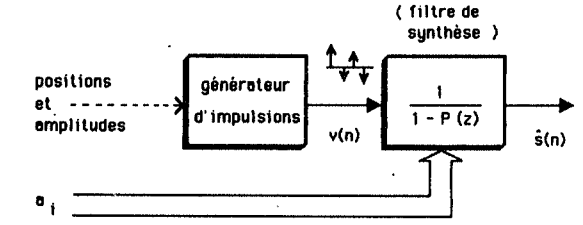


figure n° 1: synthèse MPLPC.

La parole étant divisée en fenêtres d'échantillons, le but de l'analyse MPLPC est donc, pour chaque fenêtre, d'extraire du signal les coefficients du filtre LPC et d'estimer les caractéristiques des impulsions d'excitation. Les coefficients de réflexion (k_i) peuvent être obtenus par l'algorithme de Leroux [5]. Quant aux impulsions d'excitation, elles sont choisies de façon à rendre minimale l'erreur quadratique moyenne entre le signal reconstitué et le signal original. Il est préférable, afin de tenir compte de propriétés de la perception auditive, d'effectuer une pondération du signal d'erreur afin de déplacer son énergie dans les zones formantiques où elle se trouve "masquée" par l'énergie du signal de parole. L'erreur "perceptuelle" s'écrit alors:

$$E(z) = W(z) . [S(z) - \hat{S}(z)] \text{ , où}$$

$$W(z) = \frac{A(z)}{A(z/\partial)} \text{ , } 0 < \partial < 1 \text{ ,}$$

est la fonction de transfert du filtre de pondération, avec $A(z) = 1 - P(z)$, fonction de transfert du filtre de synthèse inverse.

$1/A(z/\partial)$ définit un filtre de synthèse modifié et l'erreur s'écrit encore:

$$E(z) = \frac{1}{A(z/\partial)} . [\, R(z) - V(z) \,] = Y(z) - \hat{Y}(z) ,$$

V(z) étant la séquence d'excitation et R(z) le résidu LPC.
Y(z) et $\hat{Y}(z)$ ne sont autres que le signal original et le signal de synthèse pondérés par le filtre de masquage spectral.

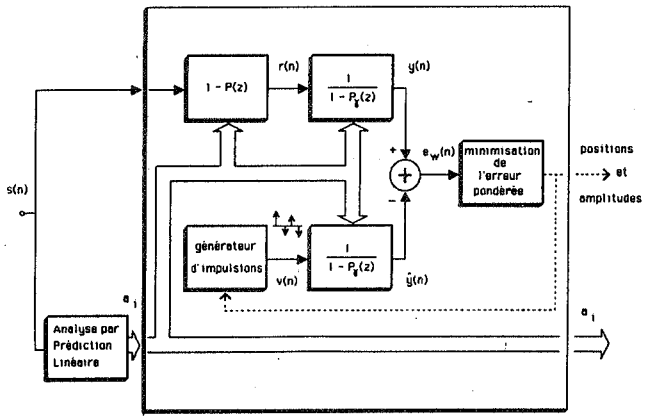


figure n°2 : chaîne d'analyse MPLPC.

Chaque fenêtre d'analyse LPC est divisée en intervalles de N échantillons dans lesquels il s'agit de placer K impulsions minimisant l'énergie de l'erreur pondérée. On note A_k et p_k l'amplitude et la position de l'impulsion k. Si h(n) désigne la réponse impulsionnelle du filtre de synthèse modifié, cette énergie s'écrit:

$$E = \sum_{n=0}^{N-1} [\, y(n) - \sum_{k=1}^{K} A_k . h(n-p_k) \,]^2$$

Si l'on suppose connues les K positions p_k, minimiser **E** par rapport aux A_k conduit au système de K équations suivant:

$$\sum_{i=1}^{K} A_i . Q(p_i , p_k) = c(p_k) \quad , \quad k = 1, ..., K \qquad (1)$$

avec

$$c(i) = \sum_{n=0}^{N-1} y(n) . h(n-i)$$

et

$$Q(i,j) = \sum_{n=0}^{N-1} h(n-i) . h(n-j)$$

Les positions des impulsions sont obtenues de manière itérative [6,7]. On suppose ainsi caractérisées i impulsions sur l'intervalle considéré. L'impulsion i+1 va rendre minimale l'erreur quadratique:

$$E_{i+1} = \sum_{n=0}^{N-1} [y_i(n) - A_{i+1} . h(n-p_{i+1})]^2$$

$$= \sum_{n=0}^{N-1} y_i(n)^2 - A_{i+1} . c_i(p_{i+1})$$

avec

$$y_0(n) = y(n) \ , \quad y_i(n) = y_{i-1}(n) - A_i . h(n-p_i) \ ,$$

(en fait, y_0 est y privé de la contribution de l'intervalle précédent)

et

$$c_0(p) = c(p) \ ,$$
$$c_i(p) = c_{i-1}(p) - A_i . Q(p_i,p) \ , \quad \text{pour } p = 0, ..., N-1 \qquad (2)$$

Algorithme:

I. Pour chaque impulsion i (i=1,...,K),
1. La position p_i de l'impulsion i qui minimise E_{i+1} rend maximale la quantité
$c(p)^2 / Q(p,p)$.
2. Son amplitude est donnée par
$A_i = c_{i-1}(p_i) / Q(p_i,p_i)$
3. Calculer $c_i(p)$ par (2)

II. Calculer les amplitudes des impulsions 1 à K par résolution du système (1).

Une amélioration consiste à remplacer à chaque itération le point 2 par la résolution du système (1) appliqué à i impulsions pour réactualiser ainsi l'ensemble des amplitudes. Le point 3 doit alors être remplacé par:

$$c_i(p) = c(p) - \sum_{k=1}^{i} A_k . Q(p_k,p) \ , \quad \text{pour } p = 0, ..., N-1$$

Les paramètres représentant une fenêtre de signal sont le facteur de gain du filtre LPC, ses coefficients (k_i) et, pour chaque intervalle d'excitation, un ensemble de couples amplitude-position d'impulsions. La plupart de ces paramètres sont quantifiés à l'aide de tables, suivant des lois non linéaires [7]. Seules les valeurs des positions d'impulsions sur un intervalle sont transmises exactement grâce à une méthode combinatoire proposée dans [6]. Le débit binaire de l'information de parole à transmettre est ainsi porté à une valeur inférieure à 6 Kbits/s.

Protection contre les erreurs

Quand on veut utiliser des techniques de codage pour transmettre de la parole, l'information disponible au récepteur risque d'être différente de celle émise par le vocodeur. En effet, dans la majorité des applications pratiques le canal de transmission n'est pas parfait et l'information utile peut toujours être entachée d'erreurs. Dans ce cas, la qualité de la transmission ne dépend pas uniquement des caractéristiques intrinsèques du vocodeur mais aussi du canal de transmission et de la robustesse de la technique de codage. Pour pouvoir réaliser un système utilisable en radiotéléphonie, nous avons analysé avec précision les effets des erreurs sur les différents bits de la trame vocodeur et mis au point une technique de codage qui permette de s'affranchir au mieux des perturbations d'un canal radio.

Une trame correspondant au codage MPLPC d'une fenêtre de parole est constituée de bits véhiculant trois types d'information: les coefficients du filtre de prédiction, le facteur de gain de ce filtre et les impulsions. Des expériences ont montré qu'une erreur sur les bits codant les impulsions avait moins de conséquence qu'une erreur frappant certains bits codant le filtre (cf. Fig. n°3). Ainsi les bits de la trame vocodeur ont été classés suivant la dégradation que provoquait une erreur les affectant et un code correcteur en bloc a été utilisé pour protéger les bits les plus sensibles de la trame.

Dans le cas d'une transmission radiotéléphonique, les perturbations du canal ne peuvent être modélisées par un bruit blanc gaussien. Les ondes radio se propageant suivant des trajets multiples, un phénomène d'évanouissement, appelé Fading de Raleigh [8], vient s'ajouter aux erreurs gaussiennes traditionnelles. Lors de ces baisses subites de champ, la perturbation est tellement importante qu'il est impossible de décoder la trame vocodeur reçue. En fait le fading, qu'il est impossible d'analyser dans le détail dans le cadre de cet article, perturbe tellement la communication que l'on peut perdre plusieurs trames vocodeur consécutives. Des techniques sophistiquées, fondées sur la diversité en fréquence permettent de s'affranchir de ce phénomène. Nous proposons une autre solution, moins performante mais beaucoup moins coûteuse, qui consiste à masquer le fading en tenant compte des caractéristiques du signal de parole. Dans certaines parties comme les voyelles, le signal évolue peu entre deux fenêtres consécutives. Il est alors possible d'atténuer les perturbations éventuelles par un traitement au niveau du récepteur.

Par exemple, en stockant les informations sur l'excitation pendant les trames considérées comme bonnes, on peut effectuer lorsque c'est nécessaire une prédiction du fondamental sur les trames passées [9]. Les coefficients du filtre de prédiction de la fenêtre précedente sont conservés pendant le trou puis interpolés lors du retour d'informations.

Beaucoup de travaux restent à faire dans ce domaine car, si l'on s'intéresse toujours à la répartition des erreurs sur le canal de transmission, peu d'études ont été publiées sur l'utilisation des caractéristiques de l'information à transmettre pour améliorer la qualité de réception.

Réalisation matérielle

Pour effectuer des essais en vraie grandeur, le codage MPLPC a été implanté sur une maquette possédant un processeur de signal TMS32010 [10]. La puissance de calcul de ce processeur et la petite quantité de mémoire interne qu'il offre ont permis de réaliser un vocodeur half-duplex. Il ne semble pas possible de faire un vocodeur full-duplex sans simplification de l'algorithme. De nouveaux processeurs, type TMS320C17 ou C25 permettront certainement d'atteindre cet objectif.

La figure n°4 présente le synoptique de la maquette. Le TMS32010 réalise l'analyse MPLPC et le codage correcteur à l'émission, le décodage et la restitution de parole à la réception. Un monochip programmable effectue la gestion de la ligne et la synchronisation trame.

Afin d'analyser le comportement du vocodeur, ces maquettes ont été connectées à des systèmes de transmission sur voie radio et des communications ont été établies dans les conditions réelles d'utilisation d'un radiotéléphone.

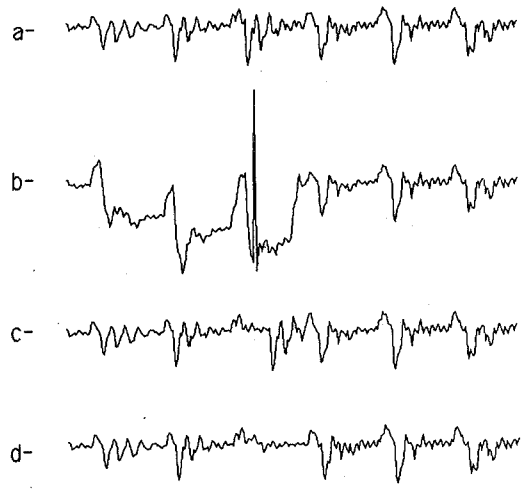


Figure n°3 : Influence des erreurs de transmission sur la trame vocodeur : a-deux fenêtres de signal reconstitué sans erreur ; b-un bit faux sur k_1; c-un bit faux sur les positions d'impulsion ; d-un bit faux sur une amplitude.

Résultats et discussion

Ces essais ont prouvé la faisabilité d'un système radiotéléphonique numérique bas débit et nous ont permis d'appréhender ses avantages et ses inconvénients.

En comparant avec un système analogique, si la qualité intrinsèque de la transmission est moins bonne, le numérique semble beaucoup moins sensible aux perturbations dans la zone de couverture. En limite de transmission, la qualité de la parole se dégrade subitement et la communication devient brutalement inaudible.

Les mesures qui ont été effectuées ont permis de vérifier réellement que le MPLPC conserve une bonne intelligibilité tant que le canal de transmission possède un taux d'erreur gaussienne résiduelle inférieur à 10^{-2}. Ce résultat est conforme à ceux déja publiés pour d'autres algorithmes [11]. Les expériences nous ont permis d'établir que la qualité de parole reste acceptable pour ce même taux d'erreur en présence de fading.

En ce qui concerne le codage de la parole, la méthode multi-impulsionnelle est utilisée pour des débits considérés comme insuffisants pour une bonne reproduction de la parole. Celle-ci est perçue comme "chevrotante" ou "enrouée". Cependant, elle reste intelligible et permet de préserver le timbre du locuteur. De plus, elle est moins sensible aux erreurs de transmission et au bruit ambiant lors de l'analyse qu'une technique bas débit du fait qu'elle ne repose pas sur une décision voisée/non voisée.

Un autre avantage de cette méthode est de pouvoir exploiter la corrélation inter-trame pour reconstituer le signal d'excitation lors des pertes de trame.

La complexité totale est compatible avec une implantation en temps réel contrairement à d'autres algorithmes de meilleure qualité.

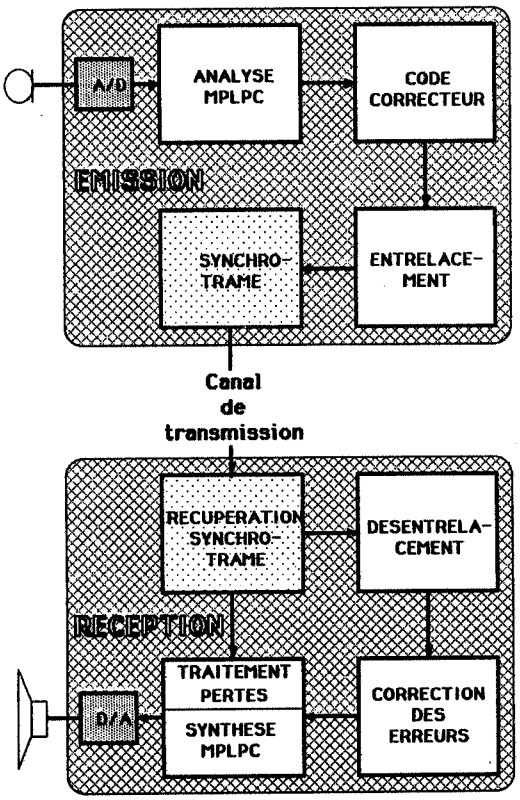


Figure n°4 : Synoptique du vocodeur

Conclusion

Le but de cette expérience était de démontrer la faisabilité d'une liaison numérique sur un équipement radio existant et d'étudier le comportement d'un type d'algorithme de codage de la parole (MPLPC) face aux erreurs de transmission.

Le codage de la parole ainsi que l'adjonction de redondance sont effectués sur un seul TMS32010 fonctionnant en temps réel.

Le système de correction des erreurs ainsi que le traitement des pertes de trame permettent d'obtenir une qualité de parole relativement bonne sur la zone de couverture contrairement aux systèmes analogiques actuels.

Références

[1] B.S. ATAL, J.R. REMDE, "A new model of LPC excitation for producing natural sounding speech at low bit rates", Proc. ICASSP-82, pp 614-617.

[2] J.P. ADOUL, C. LAMBLIN, A. LE GUYADER, "Baseband speech coding at 2400 bits/s using spherical vector quantization", Proc. ICASSP-84, pp 1.12.1-1.12.4, 1984.

[3] M.R. SCHROEDER, B.S. ATAL, "Code-excited linear prediction (CELP): high quality at very low bit rates", Proc. ICASSP-85, pp 25.1.1-25.1.4, 1985.

[4] J.P. ADOUL, P. MABILLEAU, M. DELPRAT, S. MORISSETTE, "Fast CELP coding based on algebraic codes", Proc. ICASSP-87, pp 45.9.1-45.9.4, 1987.

[5] J. LEROUX, C. GUEGUEN, "A fixed point computation of partial correlation coefficients", IEEE Trans. on ASSP, juin 1977.

[6] M. BEROUTI, H. GARTEN, P. KABAL, P. MERMELSTEIN, "Efficient computation and encoding of the multipulse excitation for LPC", Proc. ICASSP-84, pp 10.1.1-10.1.4, 1984

[7] M. LEVER, "Etude d'un vocodeur à prédiction linéaire et excitation multi-impulsionnelle. Réalisation du décodeur en temps réel sur TMS32010", Thèse de 3ème cycle, Univ. de Rennes I, décembre 1985.

[8] K. HIRADE, "Mobile-radio communications" chapitre 10 de " Advanced digital communications", K. Feher, Prentice-Hall, N. J., 1987.

[9] D.J. GOODMAN, G.B. LOCKHART, O.J. WASEM, W.C. WONG, "Waveform substitution techniques for recovering missing speech segments in packet voice communications", IEEE Trans. on ASSP, pp 1440-1448, decembre 1986.

[10] A. TASSY, L. LELIEVRE, Brevet n°85 06866, mai 1985.

[11] A. LE GUYADER, P. COMBESCURE, C. LAMBLIN, M. MOULY, J.F. ZURCHER,"A robust 16Kbit/s vector adaptive predictive coder for mobile communications", Proc. ICASSP-86.

REDUCTION DU BRUIT PAR SOMMATION SYNCHRONE APPLICATION A UN CODEUR DE TYPE R.E.L.P.

ZERUBIA J., MAYORAN T., MATHIEU P., MENEZ J.

LASSY - Université de Nice - UA CNRS N° 814 - GRECO "SARTA" -
- 41, Bd Napoleon III - 06041 NICE Cedex - France

ABSTRACT

This paper describes a method to enhance corrupted speech. The procedure which is based on pitch-synchronous averaging uses the fact that voiced sounds are quasi-periodic whereas the noise is completely random. The pitch-detection is done by means of the maximum of the normalised correlation. An appropriate weighting avoids the voiced/unvoiced detection. An exponential forgetting factor enables us to work with stationary data. Herein, we report on tests performed with synthetic waveforms and real speech signals. Then, we give results of listening tests. At last, we propose to use this method which requires a moderate computational work as a pre-filtering technique before using a R.E.L.P coder.

INTRODUCTION

La sommation synchrone est une technique de moyennage [2],[6] qui s'applique à des signaux présentant un caractère périodique, perturbés par un bruit additif aléatoire sans aucune corrélation avec le signal. Le procédé consiste à aligner temporellement N réalisations du signal bruité et à en prendre la moyenne. Dans le cas idéal, la sommation synchrone de N signaux a pour effet d'augmenter le rapport signal sur bruit d'un facteur racine carrée de N.

Dans la parole, les sons voisés présentent une périodicité qui permet d'appliquer la méthode décrite ci-dessus. Cependant, on ne peut réaliser une sommation synchrone que sur des intervalles de temps où le signal est supposé stationnaire (typiquement 15 à 25 ms). Il faut donc limiter le nombre de termes. Dans la méthode que nous proposons, ce sera le rôle d'un facteur d'oubli exponentiel. L'alignement des réalisations sommées sera quant à lui obtenu en cherchant le maximum de l'autocorrélation normalisée.

FORMULATION MATHEMATIQUE DU PROBLEME

Soit X_n le signal utile perturbé par un bruit blanc, b_n, supposé gaussien, de moyenne nulle et sans corrélation avec le signal. Nous disposons exclusivement de l'observation bruitée :

$$Y_n = X_n + b_n \qquad (1)$$

Le signal estimé $\hat{X}_n$ est obtenu par l'expression suivante [8] :

$$\hat{X}_n = \frac{\alpha\, Y_{n-m_i} + \sum_{i=1}^{\infty} \alpha^i\, r_{ij}\, b_{ij}\, Y_{n-m_i}}{\alpha + \sum_{i=1}^{\infty} \alpha^i\, r_{ij}} \qquad (2)$$

où :

- $b_{i,j}$ désigne la valeur de "u" estimée à la jième fenêtre précédente dans le ième terme de la sommation
- α est un facteur d'oubli, $\alpha \in [0,1]$
- r est le maximum de l'autocorrélation normalisée [1].
- b est un facteur d'amplitude [1]
- $m_i = \sum_{l=1}^{i} M_{l,j}$ est un retard de i pseudo-périodes constitué par l'accumulation des valeurs estimées, M_{ij}, de la période du fondamental.

L'indice j détermine sur quelle fenêtre sont estimés les paramètres $M_{i,j}$, $r_{i,j}$, $b_{i,j}$. Si LFEN désigne la longueur d'une fenêtre de calcul, on a :

* $j = ENT\left[\frac{i \,.\, M_{k,i-1}}{LFEN}\right] + 1$ pour $i > 1$

* $j = 1$ pour $i = 1$

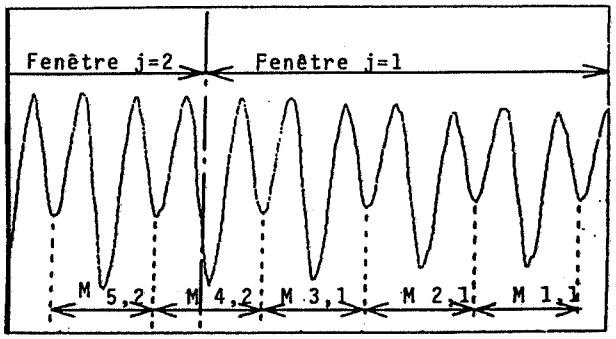


Figure 1: Figure montrant comment s'effectue la réactualisation de la période fondamentale.

Dans l'exemple de la figure précédente:

$$\sum_{i=1}^{\infty} \alpha_{i,j}\, r_{i,j}\, b_{i,j}\, Y_{n-m} = \alpha_{1,1}\, r_{1,1}\, b_{1,1}\, Y_{n-M_{1,1}} + \alpha_{2,1}\, r_{2,1}\, b_{2,1}\, Y_{n-M_{1,1}-M_{2,1}} + \ldots$$

Sur une fenêtre (chaque fenêtre correspond à 20 ms de parole), les paramètres M, r et b ne sont estimés qu'une seule fois. Ils gardent une valeur constante pour j fixé. Ainsi

$$M_{1,1} = M_{2,1} = M_{3,1} \quad \text{et} \quad M_{4,2} = M_{5,2}$$

mais $M_{1,1} \neq M_{4,2}$

Revenons à l'expression de X_n.

Dans cette formule, il faut signaler que le facteur r permet d'affecter à chaque terme de la sommation un poids qui tient compte de la qualité de l'estimation de la période du fondamental. Cet artifice permet, d'éviter la détection voisé/non voisé. En effet, pour des sons voisés r est proche de l'unité tandis que pour des sons non voisés r tend vers zéro.

RESULTATS EXPERIMENTAUX

1°/ Tests sur signal synthétique [8], [9]

Le signal utilisé dans ces tests a été généré par l'intermédiaire d'un modèle AR d'ordre 8 excité par un train d'impulsions périodiques. Nous avons travaillé sur une fenêtre rectangulaire de 480 points. Trente réalisations d'un bruit blanc gaussien obtenues avec un générateur pseudo-aléatoire nous ont permis de faire des tests statistiques. Les figures (2) et (3) représentent la DSP et les pôles du signal non-bruité tandis que les figures (4) et (5) rendent compte des résultats des tests statistiques. On constate sur ces courbes un lissage du troisième formant. On peut montrer [5] que la diminution du bruit apporté par la sommation synchrone est nécessairement associée à une atténuation des hautes fréquences (effet de filtrage passe-bas).
Enfin, la figure (6) donne la distance cepstrale [4] obtenue pour différents rapports S/B.

2°/ Tests sur signaux réels

a/ Signal reconstitué

Les figures (7) à (9) présentent un fichier de signal vocal dans ses versions non bruitée, bruitée à 20 dB, et reconstituée après sommation synchrone. L'observation de ces graphiques permet d'apprécier la réduction du bruit apportée par cette technique de moyennage.

b/ Tests d'écoute

Ils ont été réalisés à partir de voix d'hommes et de femmes échantillonnées à 8 kHz. Pour pouvoir juger de l'influence du rapport S/B sur l'efficacité de la méthode, nous avons travaillé avec des signaux ayant un rapport S/B variant de 5 à 30 dB.
D'une façon générale, nous avons constaté que la sommation synchrone apporte une réduction de bruit nettement perceptible, mais que celle-ci est associée à un glissement vers les basses fréquences et à une légère perte d'intelligibilité sur les voix de femmes.
C'est pourquoi le domaine d'application privilégié de cette méthode se situe dans une tranche de rapports S/B allant de 10 à 20 dB. En effet, pour des signaux très faiblement bruités, le choix entre la diminution du bruit ou une légère perte d'intelligibilité est très subjectif.
De même, pour des signaux très fortement bruités (S/B < 5 dB) l'amélioration est moins sensible. Par contre, aux alentours de 15 dB, après traitement, l'écoute est ressentie comme étant plus agréable par la plupart des auditeurs.
Nous proposons dans le paragraphe suivant d'appliquer la sommation synchrone en prétraitement à un codeur de type RELP. Cet appareil introduit, en effet, de par son principe même, un effet de blanchiment, qui, si on lui ajoute celui de la sommation synchrone, peut permettre lors de la transmission d'un signal bruité d'éliminer une grande partie du bruit. Cette étude a été réalisée dans le cadre d'une application industrielle [7].

APPLICATION A UN CODEUR DE TYPE RELP

Le principe d'un tel codeur est décrit dans la première partie de [3]. Succinctement, la bande de base est obtenue par filtrage passe-bas du signal résiduel tandis que la bande haute est obtenue par différence entre le signal résiduel et la bande de base.
Cependant, seule l'énergie de la bande haute est transmise au récepteur. Par conséquent, lors de la reconstitution du signal une grande partie des hautes fréquences du bruit auront disparu. En faisant précéder le RELP d'un prétraitement par la sommation synchrone on va accroître l'effet de blanchiment déjà inhérent au procédé de codage.

Résultats des tests d'écoute

La diminution du bruit due au prétraitement est très sensible particulièrement sur des signaux assez fortement bruités (typiquement 10 dB). Dans le cas des voix de femme, elle est liée à une légère perte d'intelligibilité.

Pour des voix d'hommes, aucune dégradation n'apparaît.

CONCLUSION

Si l'on travaille avec un rapport S/B inférieur ou égal à 15 dB, la sommation synchrone utilisée seule ou en prétraitement a un résultat toujours positif. Par contre, pour un rapport S/B s'étendant de ∞ à 20 dB environ, il est difficile de choisir entre une légère perte d'intelligibilité sur les voix de femmes ou une augmentation du bruit.

BIBLIOGRAPHIE

[1] **Atal, Schroeder** :"Adaptive predictive coding of speech signals" - Bell Techn. Journal - Oct 70.

[2] **Craelius, Restivo, Assadi, El-Sherif** "Criteria for optimal averaging of cardiac signals" - IEEE Biomedical Eng. vol 33 n°10 - Oct. 86.

[3] **Galand, Arnaud, Menez**:"High-frequency regeneration of base-band vocoders by multipulse excitation". ICASSP - Dallas - Apr. 87.

[4] **Gray R., Suzo, Gray A., Matsuyama** : "Distorsion measures for speech processing" - IEEE ASSP vol 28 n°4 - Aug. 80.

[5] **Jesus** :"Estimation d'un signal répétitif bruité par sommation synchrone et lissage adaptatif : application à la structure fine du signal cardiaque". Thèse de Docteur en Sciences-Université de Nice - Juin 86.

[6] **Max** :"Méthodes et techniques de traitement du signal et application aux mesures physiques" - Ed Masson 72.

[7] **Mayoran** : Rapport de DEA - Université de Nice - Juin 87.

[8] **Zerubia** :"Modélisation d'un signal à partir d'observations bruitées, application à la réduction du bruit pour des signaux de parole". Thèse de Docteur-Ingénieur - Université de Nice. Oct. 86.

[9] **Zerubia, Mathieu, Menez** :"Using synchronous averaging to enhance noisy speech" - Internoise - Pékin - Sept 87.

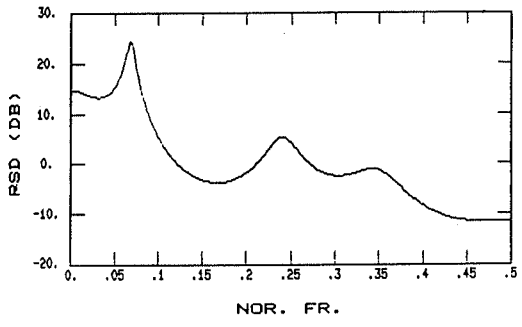


Figure 2: DSP du signal synthétique non bruité.

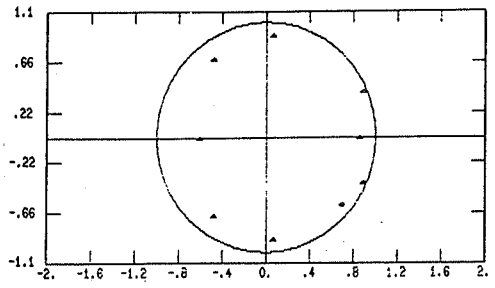


Figure 3: Poles du signal synthétique non bruité.

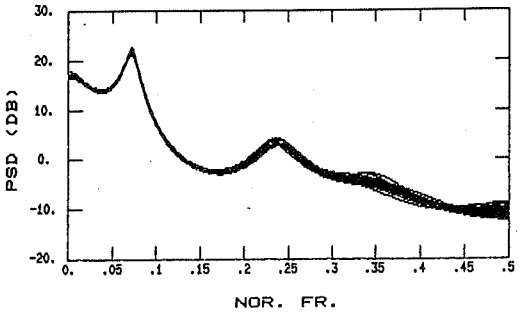


Figure 4: Tests statistiques. DSP après sommation synchrone - S/B = 15 dB.

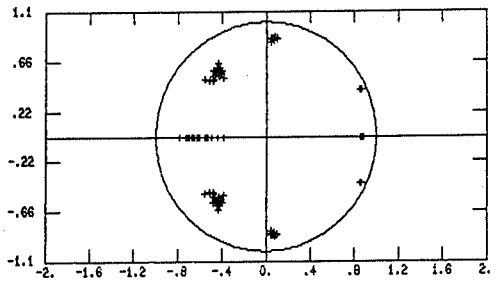


Figure 5: Tests statistiques. Poles aprés sommation synchrone - S/B = 15 dB.

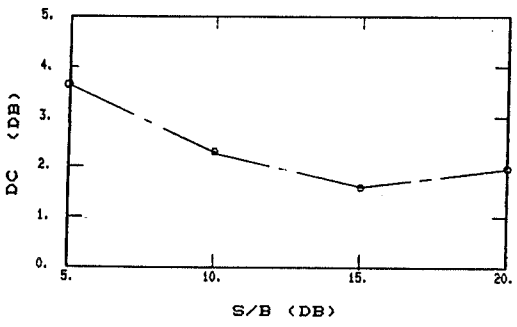


Figure 6: Distance cepstrale (dB).

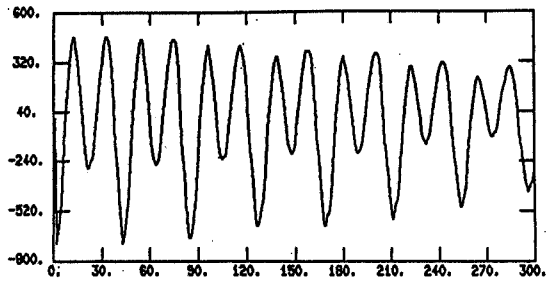


Figure 7: Signal vocal non bruité.

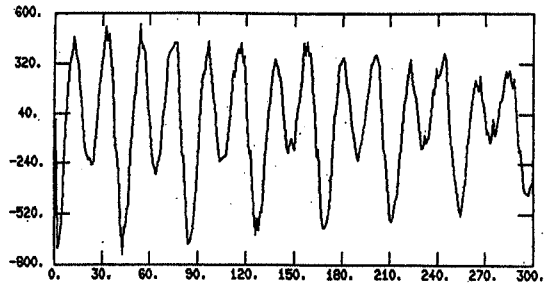


Figure 8: Signal vocal bruité -
S/B = 20 dB.

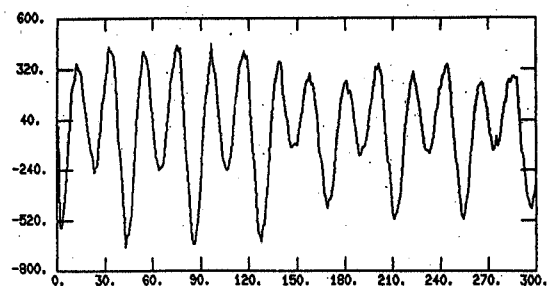


Figure 9: Signal vocal obtenu après
sommation synchrone.

INTRODUCTION AU TRAITEMENT MORPHOLOGIQUE DE SIGNAUX DE PAROLE

B. ZOUABI & N. ELLOUZE

LSTS - ENIT - TUNISIE

SUMMARY - This paper presents a new approach to speech signal processing based on mathematical morphology techniques used on image processing. First we present the morphological operations which are adapted for temporel signals. Then we introduce the idea of morphological filtering and describe its fondamental properties. Finally we present a concrete examples dealing whith pitch detection, the smoothing of noisy signal and the segmentation of continuous speech.

RESUME - Ce papier présente une nouvelle approche de traitement de signal de parole basée sur les techniques de la morphologie mathématique qui sont des techniques d'analyse d'image. On présentera tout d'abord les principales transformations qu'on a adaptées sur les signaux temporels. On abordera ensuite la notion du filtrage morphologique en présentant ses propriétés fondamentales. La dernière partie sera consacrée aux diverses applications du traitement morphologique du signal de parole liées en particulier à la détéction du fondamental, au lissage des signaux bruités et à la ségmentation de parole continue.

I - INTRODUCTION

La morphologie mathématique est une théorie d'analyse d'image. Elle est née au milieu des années soixante avec les travaux de G. Matheron et J. Serra [1],[2],[3]. Fondamentalement la morphologie mathématique est une théorie ensembliste car elle s'appuie sur les opérations de base de la théorie des ensembles. Elle emploie des transformations en tout ou rien par élément structurant appelées transformations morphologiques telles que: l'**érosion**, la **dilatation**, l'**ouverture**, la **fermeture**...

L'idée de base de la morphologie mathématique est de comparer les objets d'une image à une forme géométrique connue appelée **élément structurant.** Le principe consiste à déplacer un élément structurant dans le plan image en effectuant des tests relatifs à l'union, l'intersection ou à l'inclusion de cet élément avec ou dans les objets qui constituent l'image. L'ensemble des réponses positives forme une nouvelle image que l'on appelle image transformée. Cette démarche, qui est à la base de toutes les transformations morphologiques, permet de ressortir de l'image initiale tous les détails pertinents et d'éliminer par conséquent tous les détails unitiles.

Les techniques de la morphologie mathématique étaient appliquées au départ sur les images binaires dans l'analyse des textures et la reconnaissance de formes. L'extension aux images à niveau de gris n'a été rendue possible qu'avec le développement de la théorie du filtrage morphologique [4],[5],[6]. Le traitement morphologique couvre actuellement toutes les applications d'analyse d'image.

L'idée fondamentale de notre approche est d'étendre les possibilités du traitement morphologique à l'analyse temporelle des signaux. Des méthodes ont été élaborées tant sur le plan du filtrage que sur le plan de la détection et la caractérisation de certaines formes d'ondes du signal de parole.

Aprés avoir défini les principales transformations morphologiques qu'on a adaptées sur les signaux temporels à savoir : l'érosion, la dilatation, l'ouverture et la fermeture, nous présentons une nouvelle transformation de type adaptative appelée **transformation par les extréma** [9]. Dans la deuxiéme partie, on abordera la notion de filtrage morphologique en donnant ses propriétés importantes . Enfin, nous terminons par présenter quelques applications de traitement morphologique liées à la détection du fondamental, au lissage des signaux bruités et à la segmentation de parole continue.

II-TRANSFORMATIONS MORPHOLOGIQUES

II . 1 - Erosion et Dilatation

Soit un signal échantillonné représenté par x_k ; k=1, ..., N. Définissons une fenêtre temporelle de largeur P=L+M ayant son origine placée au point j=L+1. On fait déplacer cette fenêtre sur le signal au pas de l'échantillonnage. Dans chaque position on affecte au point j la valeur minimale ou maximale trouvée dans la fenêtre. L'ensemble des valeurs relatives aux points j et auxquelles on aurait rajouté P-1 valeurs nulles placées aux extrêmités, forme un nouveau signal y_k qu'on appelle **signal transformé**.

Cette démarche conduit à deux transformations différentes, une transformation par le minimum et une transformation par le maximum, qui ne sont

autres que les transformations morphologiques de base à savoir respectivement l'érosion et la dilatation exprimées dans un espace unidimensionnel. La fenêtre qui a été définie correspond à l'élément structurant et la largeur P=L+M définit la taille de la transformation.

L'érosion et la dilatation sont définies respectivement par les expressions [9] :

$$y_k = \min [x_{k-L},\dots,x_{k+M}], \qquad k = L,\dots,N-M$$

$$y_k = \max [x_{k-L},\dots,x_{k+M}]$$

Pour simplifier les représentations, on adoptera selon le cas les notations suivantes :

$y_k = E_P[x_k]$ ou $y_k = E_{L,M}[x_k]$, pour l'érosion

$y_k = D_P[x_k]$ ou $y_k = D_{L,M}[x_k]$ pour la dilatation

Selon la position de l'origine dans la fenêtre, le signal x_k peut être érodé ou dilaté d'une manière symétrique lorsque L = M ou d'une manière asymétrique lorsque L ≠ M. Il est à noter toutefois que quelque soit la manière dont le signal x_k a été transformé, le signal résultant est le même à un retard près, affecté par les échantillons nuls. La figure (1) donne des exemples de transformations symétriques et asymétriques opérées sur un signal échantillonné avec une fenêtre de largeur égale à dix pas d'échantillonnage.

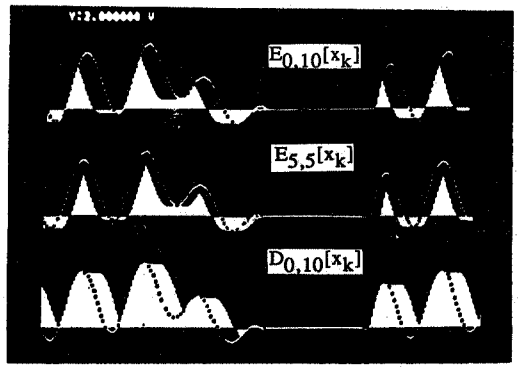


Fig. 1. Transformations par érosion et dilatation.

D'aprés les cas de figure présentés, on remarque que l'érosion et la dilatation agissent d'une manière conséquente sur le signal. En effet, l'érosion réduit l'amplitude en différents points et dilate les minima du signal en leur affectant un palier de largeur égale à la taille de l'érosion. Avec la dilatation il se produit la situation inverse, une réduction s'opére sur les valeurs minimales et une dilatation se crée autour des maxima.

- **Remarque** : Toutes les formes d'onde du signal dont la largeur de base est inférieure ou égale à la taille de l'érosion ou de la dilatation sont les premières à être affectées et ceci indépendamment de leur amplitude.

- **Propriété d'iterativité** : l'érosion et la dilatation sont des transformations itératives. En effet une érosion ou une dilatation de taille P peut être obtenue en appliquant successivement P fois la transformation de taille unitaire :

$$E_P [x_k] = E_1 [E_1 \dots [x_k]] \qquad P \text{ fois}$$

$$D_P [x_k] = D_1 [D_1 \dots [x_k]] \qquad P \text{ fois}$$

II - 2 - Ouverture et Fermeture

Il est possible d'opérer sur un signal x_k la transformation qui consiste à éffectuer une érosion du signal suivie d'une dilatation, dans ce cas en parlera de transformation par ouverture, ou à lui appliquer une dilatation suivie d'une érosion auquel cas on parlera d'une transformation par fermeture.

Ces transformations sont définies respectivement par les expressions :

$$O_P [x_k] = E_{L,M}[D_{M,L}[x_k]]$$

$$F_P [x_k] = D_{L,M}[E_{M,L}[x_k]]$$

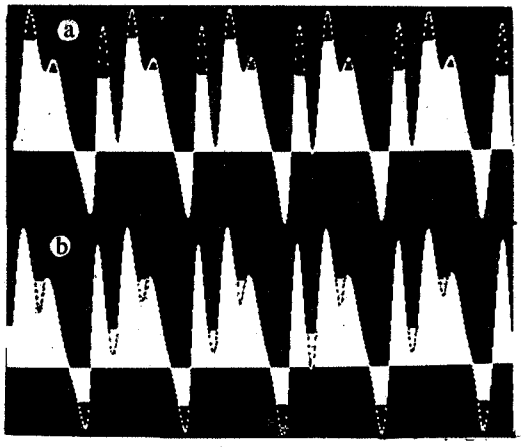


Fig. 2. Transformations par ouverture et fermeture.

La figure (2) montre le résultat de ces transformations. On remarque que l'ouverture arase les portions du signal situées autour des maxima (Fig. 2-a) et la fermeture agit de la même façon au niveau des minima (Fig. 2-b), et ceci independamment de leur amplitude relative. Ces transformations permettent d'envisager des applications importantes qui seront mentionnées par la suite notamment en matière de filtrage morphologique.

- **Proprietes D'idempotence** : L'ouverture et la fermeture ne sont pas des transformations itératives, elles sont par contre idempotentes. On peut montrer en effet que :

$$O_P[O_P[x_k]] = O_P[x_k]$$

$$F_P[F_P[x_k]] = F_P[x_k]$$

Signalons que l'idempotance est une propriété fondamentale des filtres morphologiques.

II - 3 - Transformation par les extrema

La transformation par les extrema est un cas particulier des transformations morphologiques adaptatives [9]. Pour réaliser cette transformation on choisit comme élément structurant une fenêtre symétrique définie par L=1 et M=1 qu'on déplace le long du signal. Pour chaque position on affecte à l'echantillon se trouvant au milieu de la fenêtre la valeur du premier minimum ou maximum rencontré. Cette valeur est gardée jusqu'au prochain extrémum. Le signal résultant de cette transformation est un signal multiniveaux dont les paliers correspondent à tous les extréma du signal. (fig3)

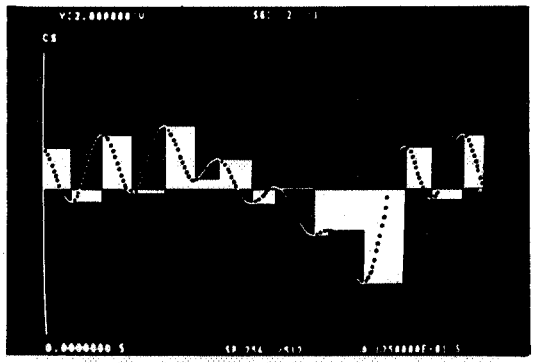


Fig. 3. Transformation par les extrema.

La tranformation par les extrema peut être réalisée par une suite d'opérations d'érosion et de dilatation opérées sur le signal jusqu'à l'idempotence. En effet, appelons T_E la transformation par les extrema relative à une seule itération. Le signal résultant de n itérations est donné par l'expression :

$$x_k^n = T_E^n [x_k]$$

tel que: $x_k^n = T_E^1 [x_k^n]$; relation qui traduit l'idempotence. L'algorithme associé à une itération est le suivant :

- Si xk est croissant :

$$T_E [x_k] = E_{1,1} [x_k]$$

- Si xk est décroissant :

$$T_E [x_k] = D_{1,1} [x_k]$$

- Si xk est un extrema ou un point stationnaire:

$$T_E [x_k] = x_k$$

III - FILTRAGE MORPHOLOGIQUE

III - 1 - Définitions et Propriétés

Signalons tout d'abord que le filtrage morphologique est essentiellement de type non linéaire. En effet, aucune transformation morphologique ne vérifie la propriété fondamentale des filtres linéaires à savoir l'additivité. Considérons deux signaux x_k et y_k et T une transformation morphologique. On aura généralement :

$$T[x_k + yk] \neq T[x_k] + [y_k].$$

Par contre, pour que la transformation T soit appelée filtre morphologique, elle devra vérifier les propriétés suivantes :

- Multiplication par un scalaire : $T[\lambda x_k] = \lambda T[x_k]$

- Continuité : si $x_k \dashrightarrow y_k \Rightarrow T[x_k] \longrightarrow T[y_k]$

- Croissance : $x_k \leq y_k \Rightarrow T[x_k] \leq T[y_k]$

- L'idempotence : $T[T[x_k]] = T[x_k]$

L'érosion et la dilatation ne verifient pas toutes les propriétés énoncées puisqu'elles sont des transformations itératives ce qui exclut évidemment la propriété d'idempotence. D'autre part elles ne peuvent pas consituer de bons filtres morphologiques pris séparément car elles agissent d'une manière conséquente sur le signal. A l'inverse l'ouverture et la fermeture ne modifient le signal qu'en certains points. Ainsi, seules les variations du signal qui correspondent à la taille de ces transformations sont lissées. Il en résulte que le signal transformé, compte tenu de la taille choisie, présente une structure plus régulière que le signal initial; cf. fig. 6. L'ouverture et la fermeture sont à la base du filtrage morphologique. Ceci est dû au fait que ces transformations vérifient toutes les proprietés des filtres morphologiques [8].

III-2- Les filtres médians

Il est possible d'autre part d'utiliser une suite d'opérations d'ouverture et de fermeture de façon à réaliser des filtres medians tels que : les filtres **FO, OF, OFO, FOF**...qui ont des

propriétés remarquables dans le filtrage de formes d'ondes particulières d'un signal.

D'autres transformations permettent d'extraire les contours lissés par l'ouverture et la fermeture. En effet, la transformation :

$$T\,[x_k] = x_k - O_p\,[x_k]$$

permet d'extraire les régions des maxima (fig. 4-a) et la transformation :

$$T\,[x_k] = F_p\,[x_k] - x_k$$

permet d'extraire les régions des minima (Fig. 4-b). Toutefois ces transformations ne présentent pas les propriétés des filtres morphologiques [8].

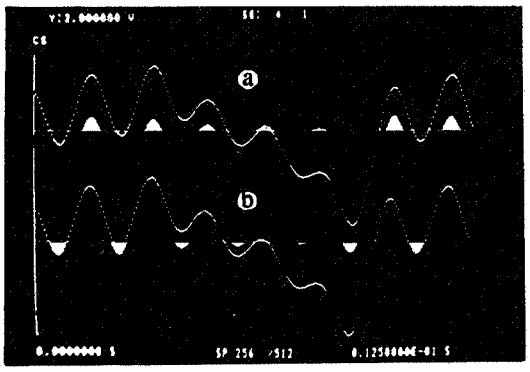


Fig. 4. Détection des régions des extrema

IV - APPLICATIONS

Le traitement morphologique du signal de parole peut être à la base de diverses applications. En effet le signal de parole, particulièrement riche en applications, présente des caractéristiques propres qui s'adaptent à ce type de traitement. Nous présentons quelques exemples d'applications liées à la détection du fondamental, au lissage des signaux bruités et à la segmentation de parole continue.

IV -1 - Détection morphologique du fondamental

L'observation du signal temporel de parole voisée et préfiltré dans la bande du pitch permet de distinguer des maxima ou minima importants liés aux instants d'excitation du conduit vocal. Les méthodes morphologiques permettent la détection de ces extréma indépendemment de leur amplitude et de la variation de la valeur moyenne du signal.

Jusqu'alors l'approche de la détermination du fondamental est basée sur une double transformation par érosion et dilatation suivie d'une transformation par les extéma. Un système de décision permet de déterminer la présence ou l'absence du pitch et de valider le résultat [9].

Une deuxième approche similaire à la précédente utilise une seule transformation, soit de type ouverture soit de type fermeture selon la polarité du signal recueilli [11] [12]. La figure (5) donne le résultat d'une érosion aprés transformation par les extéma.

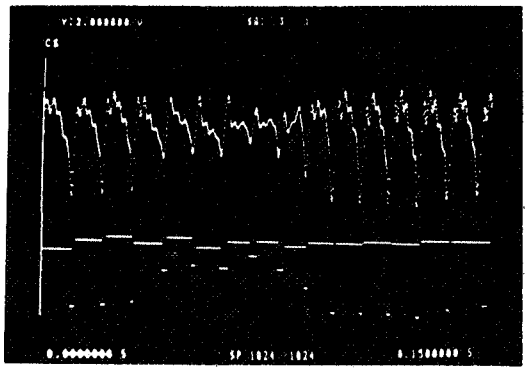


Fig. 5 . Détection morphologique du pitch.

Ces méthodes nécessitent bien évidemment une connaissance à priori de la bande dans laquelle évolue le pitch. En effet la taille de ces transformations est choisie en fonction du locuteur : homme, femme ou enfant. Une nouvelle méthode indépendante du locuteur, est en cours d'élaboration. Elle est basée sur l'utilisation des filtres médians.

IV-2- Lissage morphologique

Le lissage morphologique des signaux de parole bruités s'opére à l'aide des filtres médians. La taille des opérateurs est choisie en fonction de l'occupation spectrale du bruit. La figure (6) donne le résultat d'une application d'un filtre médian de type FO sur un signal relatif à une fricative voisée.

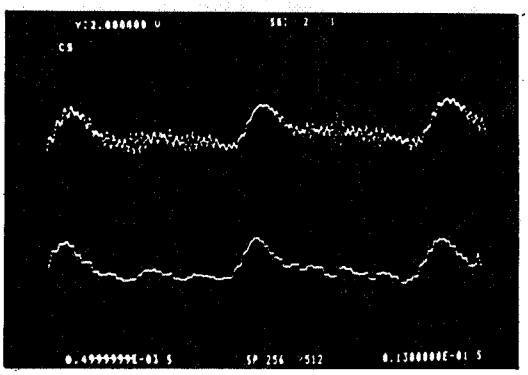


Fig. 6. Lissage morphologique d'un signal bruité.

IV-3- Segmentation de parole continue

Le principe de la segmentation du signal de parole continue peut être obtenu à partir d'un filtrage morphologique multicanal suivi d'une estimation de l'amplitude moyenne du signal dans chaque canal. Généralement le premier canal est utilisé pour représenter

l'évolution du signal. Les autres canaux peuvent servir à la caractérisation de certaines formes d'ondes du signal et entre autres à affiner la méthode de segmentation préconisée.

La figure (7) donne le résultat d'une éstimation de la variation de l'amplitude moyenne d'un signal de parole arabe obtenue à partir d'un détecteur morphologique qu'on a réalisé [10] Il est à remarquer que l'examen des deux cas de figures, correspondants à un échantillon de parole arabe, doit s'effectuer de droite à gauche.

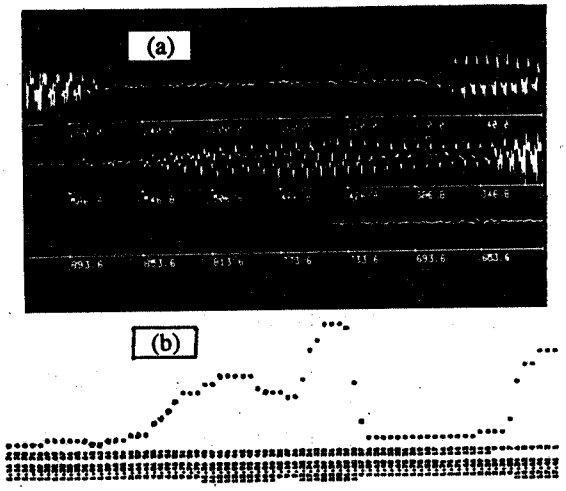


Fig. 7. Estimation de l'amplitude moyenne instantannée par détection morphologique, (a) Signal de parole, (b) amplitude moy. instantannée.

CONCLUSION

La contribution que nous avons apportée dans l'application des techniques de la morphologie mathématique a permis de présenter une nouvelle approche de traitement de signaux temporels. Les premiers résultats recueillis montrent que les techniques morphologiques s'adaptent bien aux diverses applications de traitement de signal de parole. Toutefois les méthodes présentées, notamment en matière de filtrage morphologique, gagneraient à être affinées en définissant des éléments structurants appropriés de type non linéaire tels que : les élements semi-élliptiques ou semi-circulaires.

Notons que l'originalité de l'approche morphologique réside dans le choix de la technique, dont la mise en oeuvre est facile ce qui permet d'élaborer des traitements en temps réel au vu du développement de processeurs spécialisés.

REFERENCES

[1] G. Matheron, "Les variables regionalisées et leur estimation", Masson, 1965.

[2] G. Matheron, "Elements pour une théorie des milieux porcux", Masson, 1967.

[3] J. Serra, "Introduction à la morphologie mathématique" Cahiers du Centre de Morphologie Mathématique, Ecole des Mines, Fontainebleau, N° 3, 1969.

[4] F. Meyer, Thèse de Docteur Ingénieur, Ecole des Mines de Paris, 1979.

[5] J. Serra, "Image analysis and mathematical morphology," Academic Press, 1982.

[6] G. Matheron ,"Les applications idempotentes," Rapport du CGMM, Fontainebleau, 1982.

[7] S. Beucher, "Extrema of gray tone functions and mathematical morphology," Ecole des Mines, Fontainbleau, 1983.

[8] M. Coster, J. Chermant, "Précis d'analyse d'images," CNRS, Paris , 1985.

[9] B. Zouabi, N. Ellouze et A. Ben Slimane, "Traitement morphologique de signaux unidimensionnels", Onziéme Colloque GRETSI, Nice Juin 1987.

[10] A. Ben Slimane, B. Zouabi, PREMIERE APPROCHE DE SEGMENTATION PAR FILTRAGE MORPHOLOGIQUE 16e JEP, Tunisie, 1987.

[11] A. BEN SLIMANE ET E.SELLAMI DETECTEUR MORPHOLOGIQUE DU PITCH 16e JEP, Tunisie, 1987.

[12] A. Ben Slimane, "Traitement du signal de parole par les méthodes de morphologie mathématique", Rapport ENIT, Tunisie, 1985.

[13] A. Ben Slimane, "Contribution au traitement du signal de parole par des techniques de morphologie mathématique", D.E.A, ENIT, Tunisie,1986.

[14] N. Ellouze, B. Zouabi, "Detecteur Electronique de Pitch", 3e JTEA, Tunisie, 1982.

[15] N. Ellouze , B. Zouabi , M. Folsched "Modulation Delta Synchrone", 3e JTEA, Tunisie, 1982.

[16] B. Zouabi, A. Gacem ,"Introduction à la Reconnaissance de Parole Arabe", 4e JTEA, Tunisie, 1983

[17] B. Zouabi, N. Rehif, "Système de Reconnaissance Globale de Mots Isolés", Arab School on Sciences and Technology, Maroc, 1983.

COEFFICIENTS DE REFLEXION ET SYNTHESE PAR POINTS-CLES

AGNES MANTOY

Laboratoire Image et Parole et Laboratoire de Phonétique
- Université PARIS 7 -

ABSTRACT

LPC analysis provides sets of reflection coefficients regularly up-dated during the processing of the speech signal. Synthesis is usually based on the whole set of coefficients. But here, only the parameters associated with remarkable portions of the signal ("turning-points") are retained. The search for these turning-points is performed on a graphic representation of the time function of the first two reflection coefficients. Between these frames, intermediate coefficients are recalculated via a non linear interpolation and are then used to reconstruct, through LPC synthesis, a new signal.

INTRODUCTION

On peut repérer sur le signal de parole des segments à peu près stables, d'autres transitoires à évolution lente, ainsi que des discontinuités plus ou moins marquées associées généralement à des événements articulatoires bien précis [2]. Certains auteurs ([4], [5]) ont exploité ces propriétés acoustico-phonétiques pour réaliser des synthèses à partir de paramètres attachés à ces seuls points remarquables du signal.

Nous avons entrepris de reprendre cette étude sur du français parlé standard. L'originalité de notre travail repose sur la méthode de recherche des "points-clés" du signal, méthode qui fait appel à l'étude du comportement des coefficients de réflexion issus de l'analyse LPC. Elle est mise en œuvre actuellement sur un corpus de quarante mots environ CVC/a/ insérés dans la phrase porteuse " C'est ____ ça ". Les voyelles étudiées sont /i/, /u/, /a/ et /ã/ et les consonnes /t/, /k/, /b/, /d/, /n/, /s/, /ʃ/, /v/, /z/, /l/ ainsi que le glide /j/.

SYNTHESE PAR POINTS-CLES : METHODE

Le signal, numérisé à 16 kHz sur 12 bits, est analysé, sans préemphase, par tranches de 16 ms. L'analyse fournit pour chaque tranche un ensemble de 14 coefficients de réflexion qui sont stockés avec l'intensité et le piteh de la tranche considérée. Cet ensemble de coefficients, après lissage et correction éventuelle des paramètres prosodiques, est traité pour :

- la synthèse LPC du signal "original" auquel les signaux reconstitués à partir des points-clés seront comparés,
- la synthèse par points-clés proprement dite.

La synthèse par points-clés consiste à reconstituer un signal à partir des coefficients attachés à certaines tranches d'analyse (les points-clés) dans lesquelles sont situés des événements associés à des changements articulatoires considérés a priori comme essentiels. Les points-clés ayant été déterminés, une interpolation sur les 14 coefficients de réflexion est effectuée entre ces tranches, introduisant ainsi un nouveau jeu de paramètres pour la synthèse LPC du signal.

LA RECHERCHE DES POINTS-CLES : METHODE

On sait que les coefficients sont directement liés à la fonction d'aire du conduit vocal lorsque l'analyse est précédée d'une préemphase [3] : un changement articulatoire doit se répercuter sur leur évolution. On a donc été conduit à rechercher des événements sur la représentation temporelle des coefficients de reflexion, recherche qui se déroule en trois étapes :

(i) la segmentation

La segmentation approximative en phones, sur le signal de parole, est reportée sur la représentation temporelle des deux premiers coefficients de réflexion k_1 et k_2 qui sera seule utilisée dans les étapes suivantes sauf lors de contrôle a posteriori dans les cas douteux.

(ii) l'examen des "parties stables"

A l'intérieur de ces segments, on marque les intervalles où k_1 et k_2 sont à peu près constants : on a en effet observé qu'ils coincident avec les parties stables des phones.

On a par ailleurs constaté, sur l'ensemble du corpus étudié, certaines régularités . Le coefficient k_1 est toujours proche de +1 sur les parties silencieuses ou voisées donc sur les voyelles, sur toutes les consonnes voisées ainsi que sur la tenue des occlusives sourdes. Toutefois, sa valeur est légèrement plus élevée sur les consonnes que sur les voyelles ce qui permet de distinguer des voyelles les consonnes vocaliques comme /l/. Sur les fricatives sourdes, k_1 est généralement compris entre 0 et 1.

Le coefficient k_2 est toujours négatif sur les phones voisés, voyelles et consonnes, et est en particulier très proche de -1 sur les voyelles /a/ et /ã/. Sur les fricatives sourdes étudiées, la valeur de k_2 évolue, selon la consonne, entre -0.5 et +0.5 mais cette valeur semble

dépendre, pour une même consonne, du contexte vocalique antérieur (les valeurs de k_1 et k_2 sur le /s/ de la syllabe /sa/ sont sensiblement différentes selon la voyelle, /i/,/u/,/a/ ou /ã/, qui le précède). De plus, on observe quelquefois des pics pendant la tenue de certaines fricatives mais plusieurs essais de synthèse nous amènent à penser qu'ils ne sont pas pertinents.

Ces marques constituent les premiers points-clés. Il y en a donc en principe deux par phone, sauf cas de son très bref (voyelle non accentuée par exemple) ou transitoire (certains /v/ et /l/ par exemple).

(iii) l'examen des "transitions"

Ce sont les zones où k_1 et/ou k_2 ne sont pas constants. Les coefficients varient en général de façon monotone de la fin de la partie stable précédente au début de la partie stable suivante sauf lorsqu'une occlusive est en jeu. On observe alors systématiquement une chute brusque de k_1 pendant la durée du V.O.T. (du relâchement de l'occlusive jusqu'au début du voisement de la voyelle suivante) suivie d'une remontée jusqu'au début de la partie stable de la voyelle à une valeur proche de +1. Le mouvement est le même, mais de très faible ampleur, sur les occlusives voisées.

Sur les occlusives sourdes, k_2 évolue pendant toute la tenue de la consonne (qu'on ne peut donc qualifier de partie stable) de sa valeur sur la voyelle précédente jusqu'à un point culminant dont la hauteur semble comparable à celle prise par k_2 sur les fricatives correspondantes. Au relâchement de la consonne, il redescend jusqu'à la valeur prise sur la voyelle suivante. Sur les occlusives voisées, le mouvement est à peu près semblable mais là encore de beaucoup plus faible ampleur et également plus tardif.

Ces pics de k_1 et de k_2 dans les transitions occlusive-voyelle constituent autant de nouveaux points-clés.

Les extrémités des parties stables ainsi que les pics rencontrés dans certaines transitions constituent donc l'ensemble des points-clés recherchés. Les jeux de coefficients associés à ces points sont les noyaux du système entre lesquels tous les autres coefficients sont recalculés, tranche par tranche, par interpolation.

INTERPOLATION DES COEFFICIENTS DE REFLEXION

Les coefficients étant toujours compris entre +1 et -1, nous sommes assurés que toute interpolation est stable.

Plusieurs auteurs ([2],[6]) ont montré que la sensibilité spectrale de la fonction de transfert ne varie pas linéairement en fonction d'un coefficient de réflexion k_i donné et diminue quand $|k_i|$ décroît. Différentes fonctions non linéaires ont donc été utilisées pour coder les coefficients, parmi lesquelles les fonctions

$\text{Log}[(1 + k_i) / (1 - k_i)]$ ("Log Area Ratio") et $\text{Arcsin}\ k_i$.

Nordstrand et Öhman [4], qui ont comparé divers types de codage, sont arrivés à la conclusion que l'interpolation linéaire de la fonction $\text{Arcsin}\ k_i$ est la plus satisfaisante pour la qualité de la synthèse. C'est donc cette interpolation que nous avons utilisée pour recalculer les coefficients des tranches intermédiaires.

SYNTHESE ET RESULTATS

Le signal "interpolé" est alors reconstruit, par synthèse LPC, à partir de ces nouveaux coefficients, du pitch et de l'intensité du signal d'origine (nous avons choisi dans un premier temps de ne pas faire intervenir les paramètres associés à la source).

La comparaison des signaux "interpolés" et des signaux "originaux" permet d'apprécier la pertinence de la localisation et du nombre des points-clés retenus a priori : la perception d'un net défaut à un endroit donné indique vraisemblablement l'oubli d'un point-clé. La comparaison des signaux "interpolés" entre eux peut conduire par contre à éliminer des points-clés dont la présence n'est pas justifiée par une qualité meilleure de la synthèse.

Ce contrôle a posteriori nous a conduit quelquefois à faire des corrections. Il nous a fallu par exemple introduire dans certains cas un point-clé supplémentaire dans la transition de /s/ à /a/ sur une tranche où le coefficient k_1 remonte légèrement. Sans ce point-clé, le /s/ est perçu comme palatalisé. Dans les syllabes occlusive-voyelle, caractérisées par des pics sur les deux coefficients, il semble par contre que l'on puisse quelquefois supprimer l'un ou l'autre de ces points-clés. Il faut toutefois que la pente de la variation, sinon son ampleur, soit préservée.

Dans tous les cas cependant, les signaux sont intelligibles et leur qualité est généralement comparable à celle des signaux "originaux".

La figure 1 montre la phrase " c'est dansa ça " synthétisée avec les 74 jeux de coefficients de réflexion originaux ainsi que l'évolution temporelle des deux premiers coefficients. La même phrase synthétisée à partir de 20 points-clés ainsi que les coefficients interpolés sont représentés sur la figure 2.

CONCLUSION

Notre corpus est trop limité pour que l'on puisse tirer des conclusions définitives sur le comportement des coefficients de réflexion. Cependant, certaines régularités observées sur les phones les plus étudiés (/a/, /d/, /t/, /s/, /u/) ne peuvent être dues au hasard et l'étude systématique de ces coefficients sur l'ensemble des voyelles d'abord puis sur les consonnes, en particulier les consonnes continues, devraient nous apporter de nouvelles informations.

Dans les transitions, certains pics apparaissent comme non pertinents dans la qualité voire l'intelligibilité de la synthèse. Une analyse fine sur le signal et sur sa représentation spectrale sera nécessaire pour réaliser une meilleure prévision du comportement des coefficients de réflexion dans ces zones.

Ces études complémentaires devraient permettre dans un premier temps, de réduire, en les choisissant mieux, le nombre de points-clés nécessaire à la synthèse qui est actuellement de deux à trois par phonème.

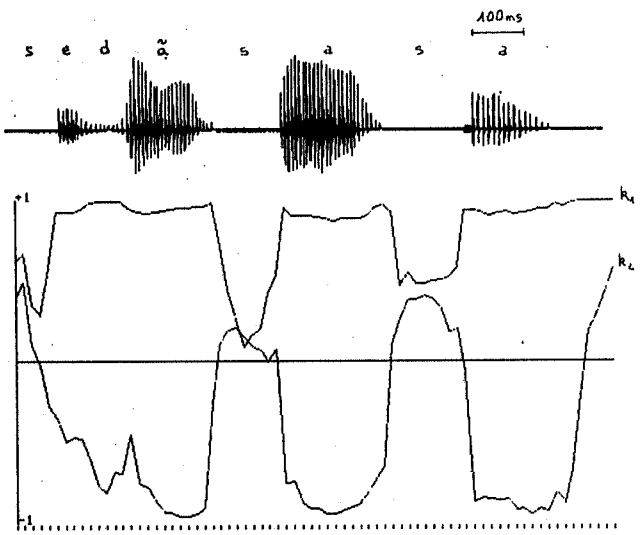


fig. 1 signal "original" /sedãsasa/

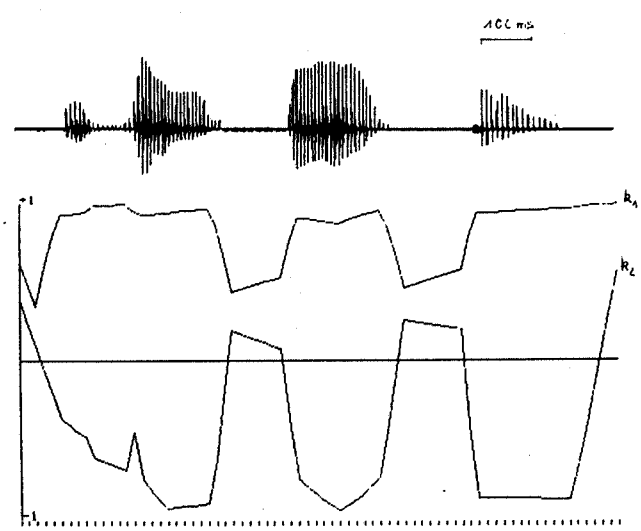


fig. 2 signal "interpolé" à partir de 20 points-clés : 1 à chaque extrémité des parties stables, 2 pour le relâchement de l'occlusive /d/, 1 au milieu du premier /a/ et 1 en fin de phrase.

A terme, si les observations déjà effectuées se confirment, la représentation temporelle des coefficients de réflexion pourrait s'avérer une aide utile à la segmentation, voire à la reconnaissance des phones.

REFERENCES

[1] Abry C. & al. - Propositions pour la segmentation et l'étiquetage d'une base de données des sons du français - 14èmes JEP GALF p 156-163, 1985.

[2] Gray A.H. & Markel J.D. - Quantization and bit allocation in speech processing - IEEE ASSP vol. 24 p 459-473, 1976.

[3] Markel J.D. & Gray A.H. - Linear prediction of speech - Springer-Verlag New-York, 1976.

[4] Nordstrand L. & Öhman S.E.G. - Computer resynthesis of speech on phonetic principles - Lund University WP n° 19 p 74-79, 1980.

[5] Olive J.P. & Spickenagel N. - Speech resynthesis from phoneme-related parameters - JASA vol. 59 n° 4 p 993-996, 1976.

[6] Viswanathan R. & Makhoul J. - Quantization properties of transmission parameters in linear predictive systems - IEEE ASSP vol. 23 p 309-321, 1975.

ANALYSE ET SYNTHESE DE VOIX PARLEE ET CHANTEE PAR MODELISATION DE L'ENVELOPPE SPECTRALE ET DE L'EXCITATION

Xavier RODET (1) (2), Philippe DEPALLE (1) (3), Gilles POIROT (1)

(1) IRCAM, 31, rue Saint-Merri, 75004 Paris, France (42 77 12 33, poste 48-27 et 48-14)
(2) LAFORIA,Univté Paris 6, 4 place Jussieu, 75005 Paris, France (43 36 25 25, poste 47-57)
(3) ESE, 2, rue Ed. Belin, 57078 Metz Cedex 3, France (87 36 29 38, poste 333).

ABSTRACT

We describe analysis, coding and synthesis technics developped at IRCAM for speech and singing voice. Power spectral density are estimated with an adaptive autoregressive modeling technique (LPC). LPC residual are coded into a voicing index for each of the frequency bands centered around harmonic partials of the signal. We give some examples of speech and singing voice synthesis. The synthetic voice is found of high quality, without the usual LPC defaults.

I - INTRODUCTION

Ces dernières années, l'IRCAM a développé un ensemble de programme d'analyse et de synthèse de la voix. Le but est de fournir des outils pour la création musicale contemporaine en voix parlée, en voix chantée, ou pour tout autre son. Il est donc nécessaire que les analyses soient très précises et les synthèses de très haute qualité.

Le cadre théorique de ces études est le modèle classique de la production vocale source-filtre linéaire. L'analyse porte donc, d'une part sur une estimation de la fonction de transfert globale du système de production par prédiction linéaire, d'autre part sur une modélisation de l'excitation en termes de fréquence fondamentale et du caractère harmonique/bruité (voisé/non-voisé) dans des bandes de fréquences centrées sur les partiels du signal.

Pour les applications sophistiquées de traitement et de synthèse, nous tirons bénéfice d'une représentation originale en formants du spectre LPC; le filtre linéaire correspondant peut être reconstitué à partir des paramètres formantiques.

Diverses méthodes de synthèse ont été implantées et comparées: fonctions d'onde formantiques (FOF), filtrage dans le domaine fréquentiel (FFT), synthèse additive, filtre en treillis. L'excitation est reconstruite par diverses méthodes : filtrage d'un train d'impulsion et de bruit blanc, FFT inverse, synthèse additive.

II - METHODES D'ANALYSE

II 1) Estimation de l'enveloppe spectrale

Nous avons cherché une représentation spectrale du signal simultanément très précise et sur une fenêtre temporelle aussi réduite que possible pour suivre rapidement les évolutions des propriétés statistiques du signal. Pour cela nous avons comparé différentes méthodes d'analyse par prédiction linéaire : autocorrélation, covariance, burg, lattice-covariance, et des méthodes adaptatives. Par exemple, la fig.1 montre les spectres LPC obtenus sur 18,75 ms d'un segment stable de parole , d'une part (A) en autocorrélation avec fenêtre de Hamming, d'autre part (B) grâce à une méthode lattice-adaptative [Viswanathan 78] avec une fenêtre demi Blackman-Harris (Cf. fig. 2) [Rodet et al 87]. Les spectres LPC sont comparés au spectre FFT pris sur 32 ms avec fenêtre de Blackman-Harris [Harris 78]. Le signal est échantillonné à 16 KHz et l'ordre de prédiction est 30. On remarque en particulier sur A l'erreur de +6 à +10 db autour du premier formant et de -6dB au-dessus de 6000 Hz. Par contre, la modélisation B est tout à fait satisfaisante, présentant très peu d'erreurs quelle que soit la fréquence.

On peut noter sur la figure 2 que seules 13 ms de signal analysé ont une amplitude non négligeable, c'est-à-dire en général entre 1 et 2 périodes fondamentales. De plus nous effectuons une analyse synchrone au pitch. Les transitions, même très rapides, peuvent donc être suivies sans perte de précision. Ceci est confirmé notamment par l'analyse de signaux synthétiques: la

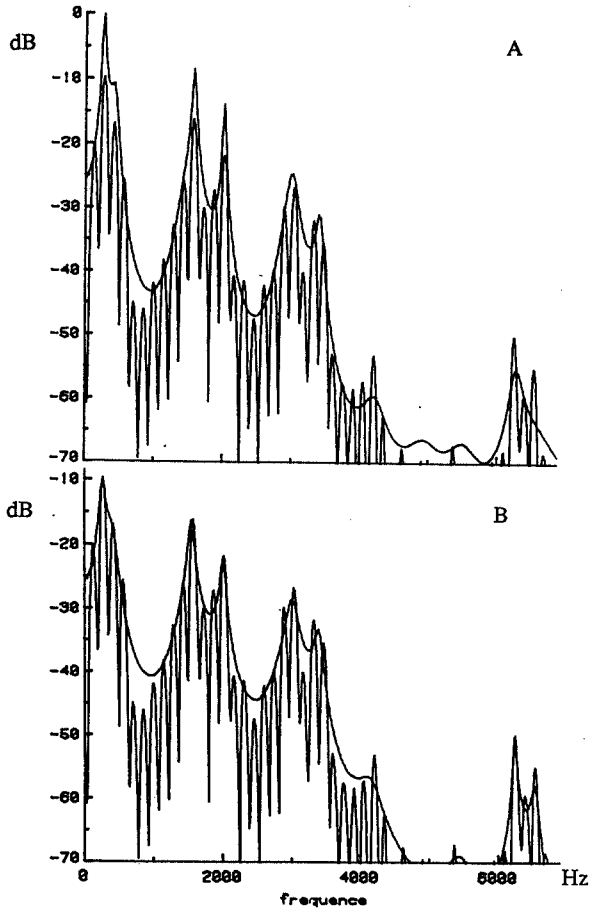


Fig. 1. Spectre FFT et spectres LPC par la méthode d'autocorrélation (haut) et par une méthode lattice adaptative (bas) sur le signal de la figure 2.

figure 3 présente le résultat de l'analyse d'une transition de fréquence et d'amplitude instantanée. Sur ce *fofogramme*, les formants trouvés sont représentés par des triangles dont la hauteur indique l'amplitude du formant.

II 2) Détection voisée/non-voisée

La détection automatique de voisement est connue pour être un problème difficile. De nombreuses "erreurs" apparaissent qui dégradent considérablement la qualité. De plus, de nombreux signaux de parole montrent un spectre à court terme qui est à la fois harmonique (voisé) dans certaines bandes de fréquence, et aléatoire (non voisé) dans d'autres. Ceci est bien connu pour les fricatives voisées mais est vrai aussi pour les voyelles. C'est une des raisons de la sonorité "buzzy" des synthétiseurs LPC: l'excitation par un train d'impulsions identiques est "trop périodique", particulièrement en hautes fréquences. Nous avons donc développé une méthode de détection de voisement dans des bandes de fréquences harmoniques de la fréquence fondamentale couvrant l'ensemble du spectre [Rodet-87]. Cette méthode est analogue à celle proposée par [Griffin 85] mais présente plusieurs améliorations. En particulier, puisqu'il s'agit de modéliser l'excitation, la méthode est appliquée au signal résiduel LPC et

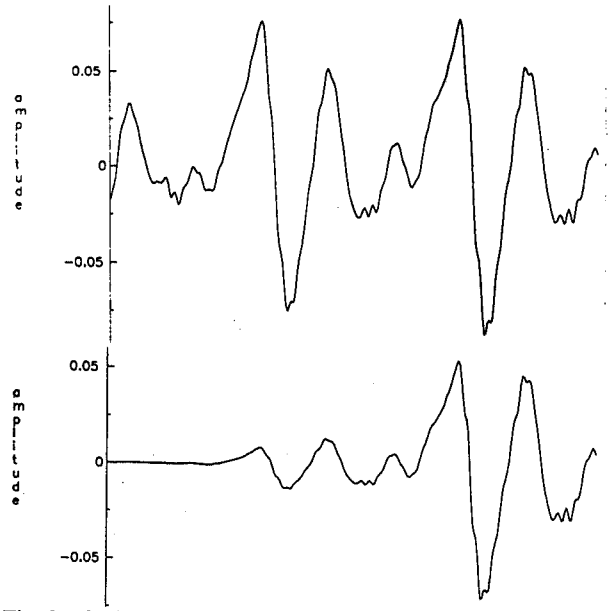


Fig. 2. 18,75 ms de signal de parole utilisés pour l'analyse LPC de la figure 1, avant fenêtrage (haut) et après fenêtrage (bas) par une fenêtre demi Blackman-Harris.

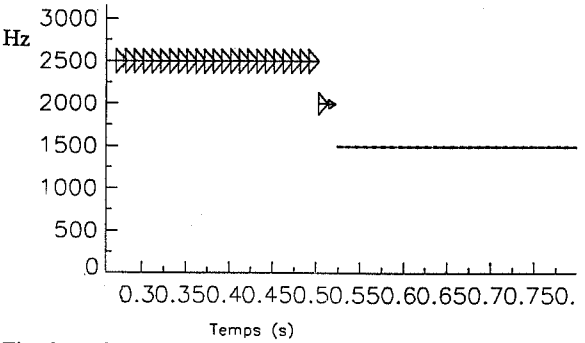


Fig. 3. *Fofogramme* d'un signal synthétique de transition de formant instantanée en fréquence et en amplitude. L'amplitude des formants détectés est représentée par la hauteur des triangles.

non au signal vocal. On évite ainsi les "non-périodicités" qui seraient dues aux variations brusques de fonction de transfert ou d'énergie.

III - CODAGE EN PARAMETRES FORMANTIQUES

Les deux méthodes de synthèse par règles les plus répandues ont des avantages et inconvénients opposés :

- la synthèse par diphone LPC [Stella 85] est relativement facile à implanter et reproduit bien les transitions consonantiques. Mais il est difficile d'adapter par règles les fonctions de transfert en fonction du contexte et la qualité sonore en général paraît limitée.

- La synthèse par trajectoires de formants [Klatt 82] [Rodet 79] nécessite un ensemble de règles définissant ces trajectoires pour toutes les séquences de phonèmes (diphones ou mieux, triphones /V_1CV_2/ [Rodet 77]). Cette approche est donc plus longue et plus difficile. De plus, l'évolution spectrale naturelle n'est pas toujours bien respectée par les règles. Par contre, le rythme, les effets de coarticulation, les allophones sont plus facilement pris en compte par des règles portant sur les paramètres formantiques.

Pour réunir les avantages de ces deux méthodes, nous commençons par coder le spectre LPC suivant ses maxima (appelés abusivement formants). Ils sont caractérisés par leurs fréquences centrales, leur amplitudes et leur largeurs de bande à -6dB. La méthode développée à l'Université Paris 6 [Montacié 87] utilise l'algorithme de Bairstow pour trouver les racines de la fonction de transfert A(z) du filtre inverse LPC. A chaque étape, la valeur d'initialisation de l'algorithme est une estimation d'un pôle calculée à partir du maximum de l'enveloppe spectrale. Finalement, les racines de A(z) sont séparées en deux groupes : le premier reçoit les pôles réels dont la contribution globale représente l'enveloppe du spectre de la source. Les p paires de pôles trouvées dans le second groupe sont réparties en m classes correspondant à m maxima, (m≤p), car un maximum peut être décrit par plusieurs paires de pôles regroupées en un "formant" unique.

La fig. 4 montre un spectre LPC et les enveloppes spectrales correspondant aux m "formants" ainsi trouvés. On remarque que ce codage représente de façon très précise les caractéristiques essentielles du spectre LPC. On en déduit une représentation de type sonagramme: le *fofogramme* d' un segment de parole est présenté à la figure 5.

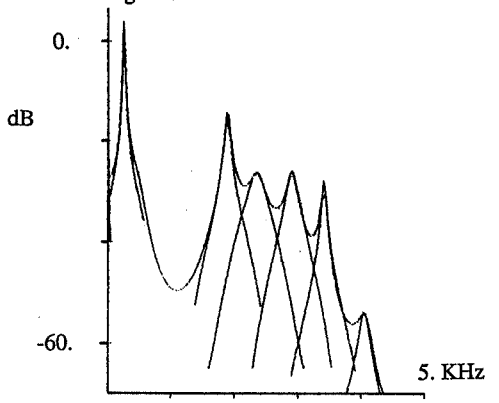


Fig. 4. Supperposition d'un spectre LPC et des spectres de Fonctions d'Onde Formantiques codant les maximas de ce spectre LPC en fréquence, amplitude et largeur de bande.

IV - TRAJETS FORMANTIQUES

Les étapes précédentes fournissent des données formantiques de la forme

$$j, t^j, F_1^j, A_1^j, BW_1^j, F_2^j, A_2^j, BW_2^j, \ldots, F_m^j, A_m^j, BW_m^j$$

où $5 \le m \le 8$ et F_i^j, A_i^j, B_w^j sont respectivement la fréquence, l'amplitude et la largeur de bande du $i^{ème}$ formant trouvé à l'analyse de la $j^{ème}$ période détectée au temps t^j.

Le problème d'extraction automatique des trajets formantiques pour un synthétiseur de type *vocodeur à formants*, est très difficile à résoudre car ces trajectoires peuvent, apparemment, se croiser, se dédoubler etc... [Chafcouloff 80]. Nous avons donc utilisé un programme graphique interactif qui affiche sur un écran les points de coordonnées (t^j, F_i^j). L'opérateur utilise une "souris" pour dessiner une courbe $f_k(t)$ qui semble suivre le trajet du formant k. Pour cela, il indique un écart sf qui permet au programme de sélectionner les points F_i^j considérés comme appartenant au trajet $f_k(t)$ si :

$$f_k(t_j) - sf < F_i^j < f_k(t_j) + sf$$

La trajectoire k est alors constituée de la séquence des $f_k(t)$ sélectionnée. Les trajectoires sont mémorisées sous forme des fonctions tabulées dans le format <temps> <valeur> pour chaque paramètre formantique.

La procédure est répétée pour chaque k de 1 au nombre n de trajectoires formantiques désirées. Finalement, un son analysé (typiquement /VC≠/ ou /≠CV/) est représenté par un ensemble de 3.n trajectoires tabulées FT_1, AT_1, BWT_1, FT_2 AT_2, BWT_2,..., FT_n, AT_n, BWT_n et par la trajectoire de la fréquence fondamentale. Les résultats obtenus en synthèse à partir des trajets formantiques ainsi obtenus sont discutés au § VII.

V - METHODES DE SYNTHESE

V 1) Synthèse par fonction d'onde formantique (FOF) [Rodet 79b]

Ce type de synthétiseur s'apparente à un vocodeur à formants en parallèle. Mais, au lieu d'utiliser une impulsion d'excitation et un filtre du second ordre pour chaque formant, on calcule directement le signal de sortie du filtre comme une sinusoïde avec une enveloppe d'amplitude exponentielle amortie. Cette méthode a donné d'excellents résultats en voix chantée pour les parties voisées. Elle a été étendue aux excitations bruitées au LAFORIA, Université Paris 6 [d'Alexandro 87].

V 2) Filtrage dans le domaine fréquentiel

On calcule la FFT de l'excitation à filtrer, puis on multiplie le spectre à court terme par la fonction de transfert en amplitude du

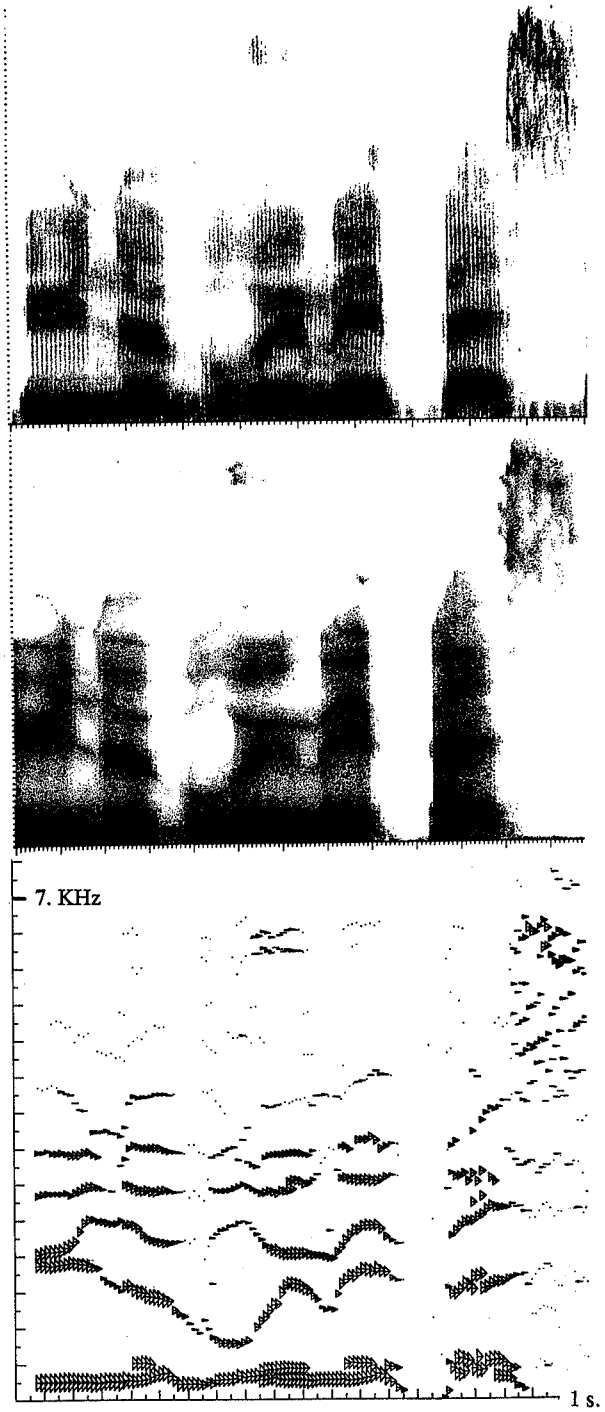


Fig. 5. Analyse du segment de parole *"une brume épaisse"*. Sonagramme numérique (haut), sonagramme LPC (milieu), et *fofogramme* (bas).

filtre. Le signal filtré est alors obtenu par FFT inverse. On peut ainsi filtrer par n'importe quelle fonction de transfert, par exemple déduite des caractéristiques formantiques obtenues à l'analyse.

Une variante de cette méthode évite la FFT directe. Le signal d'excitation étant harmonique dans certaines bandes de fréquence, bruité dans d'autres, il n'est pas difficile de construire directement son spectre à court terme d'après la décision voisée/non voisée dans chaque bande de fréquence.

V 3) Synthèse additive

Les partiels harmoniques sont calculés simplement comme des sinusoïdes dont la fréquence est harmonique de la fréquence fondamentale voulue $f_0(t)$ et l'amplitude déterminée par la fonction de transfert au temps t : H[t,f]. Ainsi le kième harmonique est:

$$x_k(t) = H[t, k.\ f_0(t)].\sin\ [t.2\pi k\ f_0(t)]$$

Pour les bandes de fréquence bruitées, elles sont obtenues par modulation du partiel harmonique par un signal aléatoire B[t; f_0] passe bas . Soit $¥_B(\nu)$ la densité spectrale de B[t; f_0]:

$$¥_B(\nu) = 1 \quad |\nu| < \frac{f_0(t)}{2}$$

$$¥_B(\nu) = 0 \quad |\nu| > \frac{f_0(t)}{2}$$

alors $x_k(t) = H[t, k.\ f_0(t)].B\ [t, f_0(t)].\sin\ [t.2\pi k f_0(t)]$

V 4) Synthèse par un filtre linéaire

Suivant l'application, on peut utiliser directement les coefficients de réflexion Ki trouvés lors de l'analyse LPC, soit reconstituer le filtre d'après les paramètres formantiques.

Dans le premier cas, de trame à trame, les Ki sont interpolés linéairement à chaque échantillon. Une interpolation linéaire des coefficients LAR (Log Area Ratio) n'a pas semblé donner à l'écoute de meilleurs résultats; il faut noter que, en raison de l'analyse synchrone du pitch, les trames sont distantes de 5 à 10 ms environ.

Dans le second cas, un filtre linéaire est estimé à partir de la fonction de transfert en amplitude déduite des paramètres formantiques (figure 4). Si l'ordre du filtre est assez élevé, l'estimation est quasi-identique à la fonction de transfert voulue.

VI - CALCUL DES PARAMETRES DE SYNTHESE

Le calcul des paramètres de synthèse à chaque instant se fait soit classiquement par concaténation de diphones soit par calcul de trajets formantiques. La méthode présentée au § IV permet d'extraire les trajets formantiques de diphones naturels avec une bonne précision. Lors de la synthèse, nous recalculons les trajets dans un contexte de cycle vocalique (typiquement /V_1CV_2/) de façon à mieux prendre en compte les phénomènes de coarticulation, ce qui paraît difficile avec une simple concatenation de diphones [Rodet 85]. Pour chaque séquence /V_1CV_2/, les trajectoires sont obtenues par une sorte d'interpolation entre les trajectoires tabulées de /$V_1C\neq$/ et /$\neq CV_2$/ [Rodet 77 et 79].

Expliquons le calcul de $f_k(t)$, trajectoire en fréquence du k^{ème} formant à partir des trajectoires tabulées $FT_{k,1}$ de /$V_1C\neq$/ et $FT_{k,2}$ de /$\neq CV_2$/. Nous calculons d'abord :

$$f'_k(t) = [1-r\ (t)]\ FT_{k,1}\ (t) + r\ (t)\ FT_{k,2}(t - T_k)$$

où - r(t) est une fonction de pondération croissant de 0 à 1 pendant la transition /V_1CV_2/ et qui représente la contribution de $FT_{k,2}$

- T_k, qui dépend du contexte, permet de positionner les transitions /$V_1C\neq$/ et /$\neq CV_2$/ l'une par rapport à l'autre et détermine donc la durée de la consonne.

Pour permettre une articulation plus ou moins accentuée, nous calculons aussi $f''_k(t)$ qui, durant la transition /V_1CV_2/, varie linéairement entre la première valeur de $FT_{k,1}$, c'est à dire $FT_{k,1}(t_{debut})$ et la dernière valeur de $FT_{k,2}$, $FT_{k,2}(t_{fin})$.

Finalement: $f_k(t) = Q\ f'_k\ (t) + (1-Q)\ f''_k\ (t)$

où Q règle en quelque sorte le degré d'articulation.

De plus, nous utilisons d'autres paramètres pour accélérer ou ralentir chaque trajet, ou chaque transition globalement, et également pour les retarder ou les avancer.

Cet algorithme nous permet de spécifier une articulation plus ou moins accentuée, rapide ou lente, avec n'importe quelle durée de consonnes et de voyelles, indépendamment des vitesses de transition, elles-mêmes variables selon le contexte. Ces caractéristiques sont nécessaires pour obtenir une bonne reproduction des transitions naturelles de la parole continue.

VII. RESULTATS

L'analyse synthèse par LPC et décision voisée/non voisée en bandes de fréquence a été testée dans des conditions de haute qualité : 16 KHz, écoute au casque. La parole de synthèse, quoique légèrement différente de l'original est exempte des défauts habituels en LPC, tel que le caractère "buzzy". De plus, il est possible de varier la fréquence fondamentale d'un octave au-dessus ou au-dessous de l'original sans perte de qualité. Ainsi la méthode est-elle adaptée à la synthèse par règles.

Parmi les techniques de synthèse, la synthèse par filtre linéaire donne la meilleure qualité et la synthèse additive est à peine moins bonne. Le filtrage dans le domaine fréquentiel, par contre, semble un peu inférieur.

La synthèse par règles sur les trajets formantiques extraits par la méthode du § IV, donne un résultat extrêmement variable : certaines transitions /V_1CV_2/ sont parfaitement reproduites (par exemple pour C = l), d'autres ne sont pas satisfaisantes (par exemple pour /i d i/). Nous pensons que ces erreurs proviennent de l'étape de codage en "trajets formantiques". En effet, si l'on effectue seulement le codage en paramètres formantiques, puis directement la synthèse à partir de ces paramètres, la qualité est préservée. Nous sommes donc provisoirement enclins à penser que lors de transitions complexes, la définition de "trajets" individuels conduit à une dégradation de l'information.

Pour remédier au problème évoqué ci-dessus, nous étudions une méthode qui permet de calculer les transitions /V_1CV_2/ suivant la méthode du § VI, à partir des transitions /$V_1C\neq$/ et /$\neq CV_2$/ codées en paramètres FOF, mais ne nécessite pas l'extraction de "trajets" individuels.

VIII. IMPLANTATIONS

Certains algorithmes d'analyse et de synthèse ont été écrits sur Array Processeur FPS 100 [Potard 84]. Les règles de synthèse et le contrôle de la synthèse sont développées dans le système FORMES [Cointe 84], un environnement de programmation orienté objet et écrit en Le Lisp à l'IRCAM. Certains algorithmes ont également été portés sur TMS-320 (OROS AU 20).

Nos développements actuels s'effectuent de préférence sur station de travail SUN-III et Array Processeur Mercury [Eckel 87] avec interface OROS-AI et PCM Sony pour la conversion numérique-analogique

IX. CONCLUSION

Des méthodes d'analyse précises des signaux de parole et de voix chantée ont été developpées à l'IRCAM. Elles constituent un environnement d'analyse très riche et très complet.

De nombreuses méthodes de synthèse ont été étudiées en vue de la synthèse par règles ou pour d'autres applications musicales. Les meilleurs résultats semblent actuellement pouvoir être obtenus:

- en analyse, par une prédiction linéaire adaptative synchrone au pitch avec fenêtre demi Blackman-Harris, avec une décision voisée/non-voisée par bandes de fréquences réparties sur l'ensemble du spectre et détermination simultanée de la fréquence fondamentale.
- En synthèse, par un synthétiseur LPC avec interpolation des paramètres, avec une excitation voisée/non voisée obtenue par construction du spectre à court terme et FFT inverse.
- Par règles de transitions sur les paramètres formantiques sans extraction de trajets formantiques individualisés.

Une expérience d'application à la synthèse par règles de la voix parlée est en cours (ce travail est supporté partiellement par le CNET-Lannion).

REFERENCES

[Depalle 84] Ph. Depalle, *Analyse numérique des sons & codage par prédiction linéaire, extraction de formants*, Université du Maine (septembre 84). Mémoire de DEA.

[Potard 84] Y. potard et J. Vandenheede, "Implantation d'un synthétiseur CHANT sur un processeur vectoriel FPS100", *Rapport interne IRCAM* (1984).

[Rodet 79] X. Rodet, J.B. Barrière et Y. Potard, "The Chant Project : from the synthesis of the sung voice to synthesis in general", *Computer Music Journal* (octobre 1984).

[Viswanathan 78] R. Viswanathan et J. Makhoul, "Adaptive lattice methods for linear predictive signal processing - Apr. 78", *IEEE Int. conf. Acoustics, Speech, Signal Proc.*, (April 78).

[Chafcouloff 80] M. Chafcouloff, G. Chollet, P. Durand, J. Guizol, et X. Rodet, "Observation and modeling of formant transitions using ISASS", *Proc. IEEE Int. Conf. ASSP, Denver, Colorado, April 1979* (April 1980).

[Cointe 84] P. Cointe et X. Rodet, *Formes : an Object and Time Oriented System for Music Composition and Synthesis*, Conference Record of the 1984 ACM symposium on LISP and Functional Programming, Austin (Texas) (5-8 August 84).

[Dechelle 84] F. Dechelle et C. d'Alessandro, *Synthèse en temps réel sur un microprocesseur TMS 320*, Université Paris 6, Paris (septembre 84). Rapports de DEA.

[Klatt 82] D. Klatt, "The Klattalk text-to-speech conversion system", *Proc. IEEE Int. Conf. ASSP pp 1589-1692, Paris, France, mai 1982* (mai 1982).

[Rodet 77] X. Rodet, Analyse du signal vocal dans sa représentation amplitude-temps. Synthèse de la parole par règles", LIF, Université Paris VI, Paris (juin 1977). Thèse d'état.

[Rodet 79a] X. Rodet et J.L. Delatre, "Time-domain Speech Synthesis by Rule using a flexible and fast signal management system", *Proc. IEEE Int. Conf. ASSP, Washington, D.C., April 1979* (April 1979).

[Rodet 79b] X. Rodet et J.L. Delatre, "Time-domain Formant-Wave-Function Synthesis", *Acte du NATO-ASI, Bonas, France, juillet 1979* (juillet 1979).

[Schwartz 79] R. Schwartz, J. Klovstad, J. Makhoul, D. Klatt et V. Zue, "Diphone synthesis for phonetic vocoding", *Proc. IEEE Int. Conf. ASSP pp 891-894, Washington D;C;, April 1979,* (April 1979).

[D'Alessandro 87] C. D'Alessandro, X. Rodet, Fonctions d'Onde Formantiques, Extraction des Paramètres et Synthèse vocale, à paraitre dans: Actes des 16èmes journées d'étude sur la parole de la Société Française d'Acoustique, Hammamet, Tunisie, Octobre 87.

[Harris 78] F. Harris, On the Use of Windows for Harmonic Analysis with Discrete Fourier Transform Proceedings of the IEEE 66(1) : 51-83.

[Griffin 85] D.W. Griffin, J.S. Lim, A New Model-Based Speech Analysis/Synthesis System, IEEE-ICASSP, Tampa, Fl., March 85.

[Montacie 87] C. Montacie, Une Détection plus sûre des Formants, Rapport interne LAFORIA, Université Paris 6, mars 1987.

[Poirot 86] G. Poirot, P. Willequet, Vocodeur à Prédiction Linéaire, Rapport de DEA, Université du Maine, France, novembre 1986.

[Rodet 85] X. Rodet, P. Depalle, Synthesis by Rule : LPC Diphones and Calculation of Formant Trajectories, IEEE-ICASSP, Tampa, Fl., March 85.

[Rodet 87] X. Rodet, P. Depalle, G. Poirot, Speech Ananlysis & Synthesis methods based on Spectral Envelopes and Voiced/Unvoiced Functions, *to appear in* European Conf. Speech Tech., Edinburgh, 2-4 Sep. 87.

[Eckel 87] G. Eckel, X. Rodet, Y. Potard, A Sun-Mercury Workstation, *to appear in* Proc. Int. Computer Music Conf., Urbana, Champain, Aug. 87.

DECODAGE
ACOUSTICO - PHONETIQUE
IA

UTILISATION D'UN SYTEME MIXTE EN DECODAGE ACOUSTICO-PHONETIQUE: PRESENTATION ET PREMIERS RESULTATS

P.DELEGLISE

U.A. 820 département SYC, ENST 46 barrault 75634 Paris Cedex 13

Résumé :

Cet article présente la collaboration entre un logiciel déclaratif et un logiciel procédural pour le décodage acoustico-phonétique de la parole continue. Cette architecture a été utilisée avec succès pour la classification en zones "voisé"-"non voisé "et pour la détection et l'identification des fricatives sourdes.

Abstract :

A collaboration of declarative and procedural systems is developed in this research. This architecture system has been successfully used for voiced-unvoiced classification and for voiceless fricative identification.

I- INTRODUCTION

Cet article décrit rapidement un système mixte déclaratif-procédural utilisé pour le décodage acoustico-phonétique de la parole continue [1]. Les premières expériences, qui ont porté sur la détection des zones voisées et non voisées puis sur la détection et la reconnaissance des fricatives sourdes, sont ensuite exposées en détail.

II- PRESENTATION DU SYSTEME MIXTE.[2]

Les deux logiciels déclaratif et procédural opèrent sur des objets munis d'attributs valués. (le mot objet est utilisé ici dans le sens usuel et non dans le sens des représentations centrées objet). Ces deux logiciels communiquent alors par échange de messages. Dans le sens déclaratif -> procédural peut être transmise une demande de caractéristique secondaire d'objet ou une demande de traitement particulier sur un objet. En réponse la partie procédurale transmet les caractéristiques précédemment demandées ou une liste d'objets munis de leurs caractéristiques principales.

Le logiciel déclaratif est un système de règles de production avec variables. Plusieurs raisonnements peuvent être conduits en parallèle par ce système.

Le logiciel procédural se compose d'un noyau et d'un ensemble d'unités de traitement.
Le noyau comprend deux parties :
-la partie gérant les objets connus des deux logiciels
-la partie qui réalise l'appel à l'unité de traitement adaptée : elle décode donc les messages provenant du logiciel déclaratif.
Les unités de traitement secondaires réalisent des algorithmes de traitement du signal ou de reconnaissance des formes.

III- CORPUS ET PROTOCOLE EXPERIMENTAL.

Le corpus expérimental est obtenu avec les trente premières phrases phonétiquement équilibrées [3] prononcées par un locuteur et deux locutrices choisis dans le corpus BDSONS du GRECO. Ce corpus est divisé en deux parties. Les quinze premières phrases de chaque locuteur forment le corpus A, les quinze autres forment le corpus B.

Le protocole suivant a été adopté :
1) détermination des algorithmes, règles et seuils sur le corpus A, appelé corpus d'apprentissage.
2) résultats obtenus sur le corpus B, appelé corpus test. Ce corpus de test est bien évidemment inconnu lors de la phase 1.

IV- PARTIE EXPERIMENTALE.

Deux expériences ont été effectuées. La première porte sur la détection des zones voisées, la deuxième sur la reconnaissance des fricatives sourdes.

A) Classification en zones "voisé"-"non voisé" [4,5].

Chaque fenêtre de signal d'une durée de 20 millisecondes est caractérisée par les paramètres suivants : cepstre, erreur résiduelle normalisée, énergie, énergie moyenne, densité de passage par zéro.

La classification s'effectue en deux étapes. Dans la première, les fenêtres sont séparées grossièrement en deux classes "voisé" "non voisé". Une fenêtre quelconque est affectée à une de ces deux classes à l'aide d'une décision logique portant sur les valeurs seuillées des paramètres choisis. Il est impossible d'obtenir à ce niveau une classification exacte, nous imposons donc à cette pré-classification de satisfaire les conditions suivantes :
- toutes les fenêtres non voisées sont classées correctement.
- toutes les zones voisées de parole produisent au moins une fenêtre classée "voisé".

Dans la deuxième étape de la classification, les zones classées "non voisé", qui sont entourées de zones classées "voisé "sont à nouveau examinées. Selon leurs durées, elles sont affectées à l'une des quatres catégories possibles où des règles spécifiques sont appliquées :
-zones courtes : la conclusion finale de présence d'une zone voisée est obligatoire. La durée est en effet trop faible pour que ce puisse être une consonne sourde.

-zones moyennes : si cette zone est effectivement voisée, l'erreur commise par la décision précédente est obligatoirement due à la présence d'une transition voyelle-consonne sonore. La conclusion finale de présence d'une zone voisée est donc effectuée si l'énergie minimale de cette zone est suffisamment importante par rapport à son contexte et si la courbe d'énergie est monotone à cet endroit. (Fig 1a)

- zones longues : le seul cas possible de zone longue voisée identifie comme "non-voisé" par la décision précédente est le cas de certains phonèmes "ʒ". Pour vérifier la présence de ce phonème, un calcul de spectre sur la partie centrale de la zone est effectué. La présence ou l'absence d'un pic dans les basses fréquences de ce spectre permet de prendre la décision finale sur le caractère voisé de la zone.
(Fig 1b,c)

- zones très longues : la décision finale reste "non-voisé".

La classification ainsi réalisée est parfaite sur le corpus testé. Toute zone voisée est reconnue "voisé" et toute zone non-voisée est reconnue "non-voisé".

B) Détection automatique et reconnaissance des fricatives sourdes.
1) Détection.

Puisque notre système mixte permet une reconnaissance absolue des zones voisées et non-voisées, il est possible de rechercher les fricatives sourdes uniquement dans les zones reconnues précédemment comme"non-voisé". Pour chacune de ces zones, un calcul de spectre lissé, sur des fenêtres d'une durée de 10 millisecondes décalées de moitié, est effectué. A chacune de ces fenêtres est associée la fréquence du maximum spectral. Pour chaque zone "non-voisé", une courbe C1 représentant une fonction f(i) est ainsi obtenue : f(i) est la fréquence du maximum spectral de la ième fenêtre de la zone.
Les zones se répartissent en trois catégories selon la durée d de la partie où f(i) est supérieure à 1000 Herz
- d>0.20 secondes : présence d'une fricative sourde et d'une autre consonne sourde ou d'un autre phonème dévoisé.
- 0.07<d<0.20 : présence d'une fricative
- 0.04<d<0.07 : présence d'une fricative si l'énergie des basses fréquence est décroissante au début de la zone.

2) Identification [6,7].

Nous considérons la partie Zf formée des fenêtres i où f(i)>=1000. Zf est une zone du signal. Nous disposons de l'information de la courbe C2, restriction de la courbe C1 à la zone Zf. Nous caractérisons cette courbe par une fréquence représentative Fr et par le nombre Nr de fenêtres où f(i) est supérieur ou égal à 3500 Herz. Nous calculons également un spectre lissé S1 sur une fenêtre de 32 millisecondes centrée sur la partie Zf. Ce spectre est fourni en bande critique.
La discrimination s-f se fait sur la valeur de Fr.
La discrimination s-ch peut se faire à partir de la valeur de Fr : si cette valeur est élevée, la décision "s" est prise. Des cas d'ambiguité subsistent pour des valeurs moyennes de Fr. La discrimination s'opère alors en comparant le minimum de l'énergie des basses-fréquences (500-1000 Herz) à l'énergie des moyennes fréquences (1500-3500) dans le spectre S1.
La discrimination f-ch suit le même shéma que celle de s-ch, avec des valeurs différentes.
A l'aide de ces règles le système mixte analyse le corpus test qui comprend 3*21 fricatives. Il fournit les résultats suivants :
- 51 détections et identifications exactes
- 1 erreur d'identification
- 2 refus de décision à l'identification
- 5 omissions à la détection
- 2 fausses détections
Ces erreurs ou omissions ne sont pas très inquiétantes car elles proviennent en fait de cas non rencontrés dans l'ensemble d'apprentissage.

V- CONCLUSIONS

Le système présenté ici a obtenu d'excellents résultats sur la séparation des zones voisées et non-voisées et sur la reconnaissance des fricatives sourdes. L'ensemble test étant inconnu du système et de l'utilisateur au moment de l'apprentissage, nous pouvons espérer que les résultats seront robustes lors de l'augmentation du nombre de locuteurs. Pour compléter ce système, il nous faut dans un proche avenir le coupler à des modules d'apprentissage symbolique et statistique (8).

FIGURES

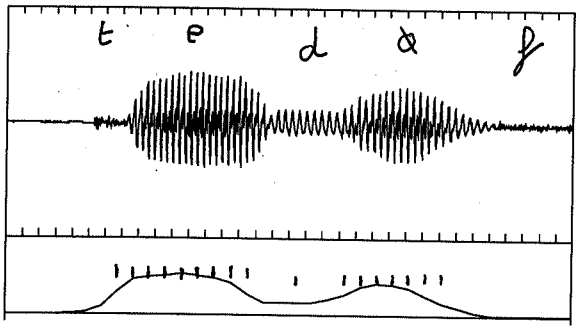


figure 1a

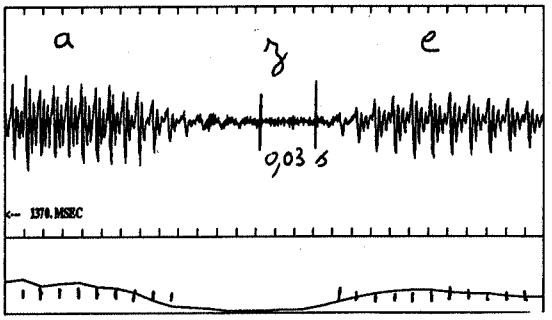


figure 1b

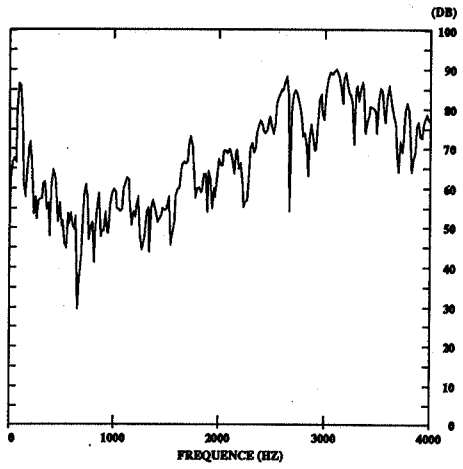


figure 1c

figures 1a et 1b partie supérieure: signal de parole, partie inférieure: courbe d'énergie avec marques de voisement indiquant la fréquence fondamentale.
figure 1a: l'erreur de la préclasification est due à une transition voyelle -consonne et vice versa.
figure 1b: l'erreur de la préclassification est due à la présence du phonème ʒ.
figure 1c: spectre T.F.D de la partie centrale de la zone longue.

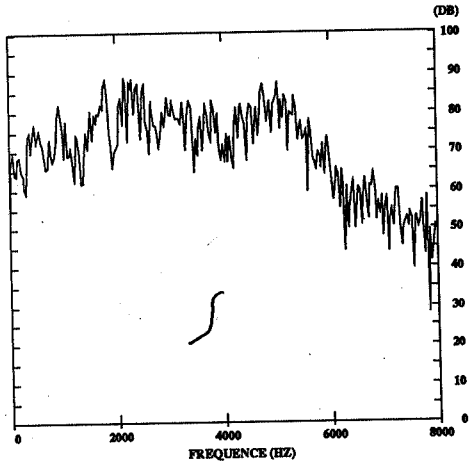


figure 2a

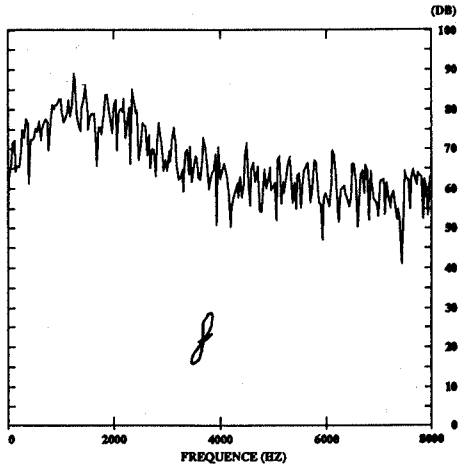


figure 2b

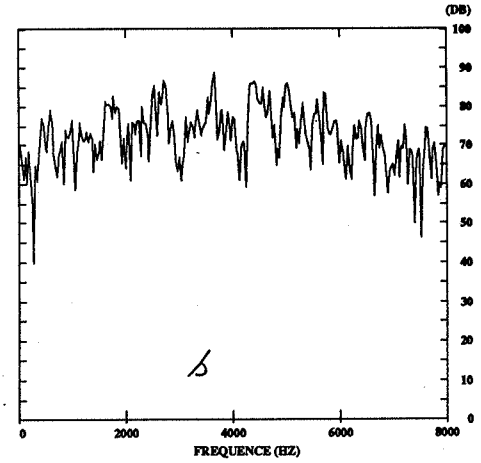


figure 2c

figures 2a,2b,2c: spectre T.F.D d'un exemple de chacune des trois fricatives sourdes

BIBLIOGRAPHIE

1- J.P. HATON : "Les Systèmes à base de Connaissances" TSI, vol 4, n°3, 1985.
2- P. DELEGLISE : "Décodage Acoustico-Phonétique de la parole continue par un système déclaratif-procédural : description et résultats" AFCET-RFIA,nov 1987.
3- P. COMBESCURE : " Vingt Listes de Dix Phrases Phonétiquement Equilibrées" Revue d'Acoustique, 14, n°56, 1981.
4- H. PIGOT, &al : " Recherche du Fondamental par une Méthode AMDF avec Programmation Dynamique" Séminaire GALF GRECO. PARIS. 1983.
5- B.GOLD, L.R. RABINER : " Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain"
6- M.CHAFCOULOF, A. DI CRISTO : " Les Indices Acoustiques et Perceptuels des Consonnes Constrictives du Français, Application à la synthèse" JEP, Lannion, 1978.
7- A. BONNEAU, G. MERCIER, M. GERARD, M. ROSSI : " Le décodage acoustico-phonétique à l'aide du système expert SERAC-IROISE". JEP, 1986.
8- H. MELONI, R. BULOT : " Décodage Acoustico-Phonétique en Prolog", JEP, 1986.

LES ACCES : INTERFACE ENTRE LE DECODAGE ACOUSTICO-PHONETIQUE ET LE LEXIQUE

Jacques Gispert

GIA - Faculté de Luminy - 70 Route Léon Lachamp 13288 Marseille Cedex 9

ABSTRACT

This paper presents a system for accessing words in a large Lexicon, written in PROLOG II for Word Identification in Continuous Speech.

The processed sentence is stated as a list of segments described by valued phonetic features.

Data used to compute accesses is constituted by the phonetic form of the words in the Lexicon, and by some informations about the behaviour of the acoustico-phonetic decoding system which produces the input phrase.

These informations constitutes a set of declarative rules setting the average valuation of each feature, depending on its context.

Using these rules and the phonetic description of words in the Lexicon, the system produces a hierarchical set of clauses which verifies successively the occurence of acoustic features in the sentence to split the set of candidates words.

1 - INTRODUCTION

Une étude préliminaire des possibilités d'utilisation de l'intelligence artificielle en R.A.P.C. a été présentée récemment [1]. Il s'agissait d'un système de reconnaissance analytique de mots à partir de segments de parole, découpés et étiquettés par un programme de traitement du signal écrit en FORTRAN. Il reprenait une structure de treillis déjà décrite dans [2]. Les divers problèmes y étaient abordés simplement, et les résultats obtenus étaient assez encourageants.

On a décidé de reprendre cette méthode pour développer en détail un nouveau sytème de reconnaissance analytique de mots dans le discours continu : décodage acoustico-phonétique (D.A.P.) [3], génération de règles de caractérisation phonémique par apprentissage [4], et constitution du Lexique.

Le codage du Lexique a été envisagé par ailleurs [5], et comprend la description complète des mots : formes graphique et phonologique, catégories syntaxiques, usage et particularités, renseignements divers. Il constitue une représentation du vocabulaire totalement indépendante du système de reconnaissance et du locuteur. Les verbes sont fournis à l'infinitif.

Le problème traité ici est celui de la détermination et de l'utilisation des accès lexicaux, qui constituent un module à part faisant correspondre le D.A.P. et le lexique.

L'ensemble du système, lexique, codage des accès et utilisation en reconnaissance est décrit sous forme déclarative en PROLOG II.

Le Lexique contient actuellement un peu plus de mille mots cités essentiellement dans l'Elaboration du Français Fondamental [6], ouvrage destiné à l'apprentissage de la langue en particulier par les étrangers. Il contient beaucoup de mots lexicaux courts. Ce vocabulaire sera progressivement enrichi de mots répondant à des thèmes.

2 - REPRESENTATION DE LA PHRASE

Le D.A.P. utilise une arborescence de traits, ayant les traits vocalique, fricatif, interrompu et consonnantique (qui regroupe ce qui n'a pas été reconnu fricatif ou interrompu) au niveau supérieur. La nature et la détermination d'un trait dépendent des traits supérieurs du même segment.

La fiabilité des traits obtenus en reconnaissance à partir du signal est décroissante en descendant le long d'une branche. Ceci correspond à une augmentation de la difficulté de caractérisation avec la finesse des détails, et à la propagation des erreurs. La méthode de reconnaissance doit en tenir compte. En particulier, le choix des accès doit refléter la prépondérance des traits supérieurs.

Le "e" muet est codé de façon particulière, ses traits n'étant pas définis par suite de son extrême variabilité. Il est considéré comme un segment indéfini dont la longueur peut être nulle.

La phrase à analyser est donnée sous la forme d'une liste de segments, porteurs de traits valués par les règles qui les ont déterminés.

3 - DIVERS TYPES D'ACCES

On cherche à obtenir un accès remontant, qui propose un treillis de mots susceptibles d'être reconnus.

Un accès prédictif est également nécessaire, par les catégories syntaxiques, et éventuellement par d'autres attributs des mots.

La définition précise des accès remontants doit réaliser un équilibre entre :

- l'accessibilité des mots, qui pour être exhaustive risque de conduire à des reconnaissances erronées ;
- la pertinence des propositions, qui incite à rendre le filtre plus sélectif, et qui est donc susceptible d'entraîner des omissions.

4 - CRITERES DE CHOIX DES ACCES

Pour obtenir des accès efficaces, il faut tenir compte des caractéristiques du D.A.P. (provenant de ses caractères propres et des propriétés intrinsèques des phonèmes). Contrairement au Lexique lui-même, le module d'accès est dépendant du D.A.P., et dans une certaine mesure du locuteur. Cette dernière dépendance sera précisée ultérieurement.

L'adaptation des accès aux propriétés du D.A.P. est réalisée automatiquement à l'aide de connaissances relatives à son comportement.

Elles décrivent la qualité prévisible des traits d'un phonème calculée en fonction de son contexte. Ces connaissances ont été pour l'instant déterminées manuellement; mais elles seront obtenues automatiquement par apprentissage lorsque l'ensemble des phrases du corpus de test pour le D.A.P. aura été analysé [4].

Ces connaissances permettent de traiter les mots du lexique en imitant les performances du D.A.P. Elles prévoient que certaines configurations n'ont qu'une probabilité très faible d'être reconnues. On peut les éliminer dans le codage des accès, sans risque de perte notable.

Au moment de la comparaison entre un mot trouvé dans le lexique et la phrase donnée, elles permettent de moduler les valuations associées à certains traits, et évitent ainsi de rejeter quelques mots présentant une anomalie sur un trait dans le signal.

Les accès aux verbes sont construits en conjuguant à tous les temps et toutes les personnes. Les terminaisons phonétiquement identiques sont regroupées.

5 - STRATEGIES

La probabilité pour qu'un mot reconnu ait été réellement prononcé dépend du nombre de phonèmes qui le composent, et de leur qualité de détermination. Pour chaque mot du lexique, on calcule donc le score de reconnaissance de ses phonèmes, et on lui applique une fonction dépendant linéairement des scores, et non linéairement du nombre de phonème. On a choisi de multiplier la somme des scores par une fonction empirique du nombre de phonèmes, qui privilégie les mots longs. Cette valeur est comparée à un seuil au-dessous duquel aucun accès remontant n'est créé pour le mot.

Le seuil a été choisi de manière à ne pas remonter les mono-syllabiques (voyelles : à, y, au, eu... ; consonnes : qu', d', l'...) qui ne sont pas assez sélectifs pour être reconnus avec certitude en l'absence de frontières de mots prédéfinies. Ils seront déterminés par accès descendant à partir de l'ossature syntaxique de la phrase en cours de construction.

On peut envisager de privilégier les mots longs, peu nombreux en moyenne dans une phrase, mais plus surement reconnus et qui peuvent par conséquent servir de points d'ancrage. Il suffit de considérer d'abord ceux pour lesquels cette fonction prend la plus grande valeur.

6 - DETERMINATION DES ACCES

Dans un premier temps, on a réuni en une seule classe "consonnantique" les phonèmes ayant pour premier trait : fricatif, interrompu ou consonnantique. On retient pour chaque mot du lexique sa structure C/V, qui est la suite des classes consonnantique et vocalique de ses phonèmes. Les différentes structures trouvées dans le lexique sont numérotées :

structure(consonnantique.vocalique.consonnantique.nil,12) -> ;

déterminée entre autre par le mot : "donc".

A chaque trait d'un phonème est associé un score à l'aide des connaissances énoncées sur la valeur contextuelle des traits reconnus en fonction des deux phonèmes voisins. Si un trait a une probabilité de reconnaissance trop faible, sa valuation est nulle, et les traits qui en dépendent dans la hiérarchie sont ignorés (valeur nulle). Le score a une valeur comprise entre 0 et 5, c'est donc un chiffre.

Exemple

règle : le trait compact du phonème /a/ est reconnu avec un score 1 ;
en contexte gauche labial, son score est 2 ;
en présence d'un /r/ à gauche ou à droite, sa valuation est 3.

Certains des ensembles déterminés par une telle structure contiennent un grand nombre de mots. On procède de manière semblable pour les traits de niveau 2, i.e. placés immédiatement au-dessous de consonnantique et vocalique. Toutefois, si la valuation d'un trait est insuffisante, on remplace ce trait par une variable à laquelle on impose des restrictions provenant des autres valeurs de traits bien reconnues dans cette position. Ainsi, si aucun trait n'a pu être déterminé à ce niveau, mais qu'une seule possibilité est mal reconnue dans ce contexte, on pourra accéder à la cohorte des mots ayant ce trait.

Dans le cas où il reste plusieurs possibilités, ceci agrandit la cohorte des mots atteints, par fusion de celles dont le trait est mal défini, mais correspond à une situation dans laquelle l'information remontante n'est pas assez fiable pour prendre des décisions correctes.

A chaque structure C/V correspondent quelques structures de niveau 2 qui fragmentent l'ensemble des mots déterminés par la première. On refait la même chose aux niveaux inférieurs. On a au plus 5 niveaux correspondant à la profondeur maximale de la hiérarchie des traits. Ceci constitue un filtre arborescent simple. Chaque niveau donne accès au niveau immédiatement inférieur.

Plus on descend dans la hiérarchie, moins les traits sont fiables (certains ne sont pas définis du tout, comme par exemple pour /v/ ou pour les voyelles nasales). Le filtre correspondant est de moins en moins sélectif, mais il s'applique à des ensembles restreints.

On arrête ce processus de découpage si la cohorte ne contient plus qu'un seul élément. L'accès aux mots se fait donc d'après l'arborescence des traits, conduisant à une fragmentation successive des cohortes.

L'ensemble des accès constitue un véritable programme d'analyse du signal, obtenu automatiquement par traitement de chacun des mots du Lexique, et adapté au D.A.P. Dans certains cas, ce programme demande des traits qui n'ont pas été trouvés de prime abord dans le signal. On peut alors entreprendre une recherche plus fine pour évaluer la possibilité ou l'impossibilité de le reconnaître. Ceci n'est pas encore réalisé, mais le D.A.P. est prévu pour le permettre.

Tous les mots proposés ont ensuite leur description théorique comparée avec le signal, ce qui permet d'éliminer certaines erreurs, et fournit une valuation de la reconnaissance.

Les accès syntaxiques ne sont pas codés pour tous les mots afin de diminuer l'encombrement. Ils sont définis pour les mots n'ayant aucun accès phonétique, ou n'ayant qu'un accès phonétique peu fiable. Dans ce cas, ils assurent une certaine redondance qui permet de faire un choix cohérent.

7 - VALUATION DE LA PHRASE

La phrase à analyser est donnée sous la forme d'une liste de segments porteurs de traits valués. On détermine les portions les plus intéressantes de la phrase de la manière suivante :

- pour chaque segment, on construit une liste de traits (dans l'ordre de la hiérarchie des traits), en ne retenant que ceux ayant la meilleure valuation ;
- on fait correspondre à cette liste la liste des valuations ;
- on remplace dans cette dernière les entiers par des chiffres représentant des classes de valuations ;
- on construit un entier en base 10 en accolant les chiffres de la liste.

Chaque phonème de la phrase est ainsi caractérisé par un nombre qui reflète la qualité de chaque trait. Il est par conséquent immédiat de comparer deux phonèmes entre eux, par la simple relation d'ordre sur les entiers. Cette comparaison tient compte de la hiérarchie des traits, le chiffre de poids fort correspondant au trait de niveau le plus élevé. Pour les phonèmes ayant moins de traits distinctifs, on complète à droite par des zéros.

Exemple

Sur un segment, les meilleurs traits obtenus sont :

vocalique(100%), oral(72%), fermé(84%)

ils donnent les listes : vocalique(100).oral(72).fermé(84).nil

les classes de valuation étant 4 : de 90 à 100 ;
3 : de 80 à 89 ;
2 : de 70 à 79

on en déduit la liste : 4.2.3.nil

qui donne le nombre 42300 pour score du segment.

Chaque phonème d'une phrase étant valué de cette façon, on obtient une liste d'entiers caractéristiques. On peut mettre en évidence des îlots favorables à l'intérieur d'une phrase en calculant une seconde liste d'entiers obtenue à partir de la première par pondération de trois valeurs successives : par exemple, somme de la valeur centrale et de la moitié des deux valeurs extrêmes . Ceci donne une vision plus globale de la phrase, et la plus grande des valeurs obtenues désigne la partie du signal qui porte les plus grandes potentialités de reconnaissance.

8 RECHERCHE DES MOTS POSSIBLES

C'est à partir des îlots les meilleurs que la recherche des mots commence. Chaque îlot détermine trois segments, dont on retient les traits consonnantique et vocalique. Ce triplet de traits donne un accès direct (précalculé) aux structures C/V qui le contiennent. Ces structures sont alors appliquées sur la liste des traits C/V de la phrase à analyser, et celles qui lui correspondent sont retenues.

On passe alors à l'examen de la liste des traits de niveau 2 de la phrase, et la comparaison donne accès au niveau 3, etc... Ce processus s'arrête au bout de la branche, ou bien lorsque les accès n'ont plus été calculés à cause d'un score trop faible. A ce moment, on obtient la cohorte proposée pour cette portion de la phrase donnée.

Les liaisons sont calculables en fonction du caractère consonnantique ou vocalique du premier phonème du mot suivant. Elles consistent en la présence faculative d'un phonème en fin de mot. Cette présence est également conditionnée par le style de conversation, les habitudes du locuteur etc... Ces informations ne seront pas traitées, la présence ou l'absence de liaison sera simplement constatée.

9 DETERMINATION DES MOTS RETENUS

Ensuite, la description phonétique exacte de chaque mot proposé, trouvée dans le Lexique, est comparée avec le signal, ce qui constitue un filtre ne laissant subsister que des mots qui ne peuvent être exclus par le module de D.A.P..

Tous les mots du lexique ne sont pas accessibles en remontant, soit parce qu'ils sont trop courts et donc trop ambigus, soit parce que leur constitution phonétique conduit à une espérance de caractérisation trop faible. Ces mots-là ne seront accessibles que par dessus, grace à la syntaxe.

La classification des mots par leur structure C/V. a été évaluée par Shipman [7] ; elle détermine un peu moins d'une centaine de groupes. Certains d'entre eux contiennent très peu de mots, et suffisent presque à leur détermination. D'autres sont au contraire très importantes, et nécessitent une analyse beaucoup plus fine. Les mots obtenus avec le plus petit nombre de traits sont plus fiables, puisque les erreurs de détermination des traits se cumulent en descendant dans la hiérarchie.

Le remplacement des structures C/V par les structures vocalique, fricatif, interrompu et consonnantique ne changera rien au principe, mais multipliera le nombre des classes. Il doit corrélativement diminuer la taille des cohortes, en étant plus sélectif.

Malheureusement, la méconnaissance des frontières entraîne la détermination de mots à cheval sur deux mots réels. Il arrive que ces mots fantômes soient bien reconnus, et leur élimination n'est possible que par le renforcement des valuations individuelles des mots corrects par leur proximité temporelle dans le signal et leur appartenance à une même structure syntaxique.

10 RESULTATS

Les résultats dont on dispose actuellement sont fragmentaires, établis sur un trop petit nombre de mots pour avoir une valeur statistique. On peut toutefois noter que :

- le volume des accès est raisonnable ; il comprend les structures de traits des différents niveaux, qui représentent quelques centaines de termes, les "bons" triplets permettant de les atteindre, et enfin, pour chaque mot, un accès direct lui faisant correspondre la dernière structure. Le codage des structures n'est pas directement lié à la taille du lexique, et ne croit que très lentement avec elle (dans un lexique de base, les structures les plus fréquentes sont déjà représentées, et les mots ajoutés par la suite en créent rarement de nouvelles). En gros, on peut donc dire que le volume du codage est linéaire par rapport à la taille du lexique.
- les temps d'accès aux mots sont très raisonnables, de l'ordre d'une dizaine de mots par seconde sur Macintosh Plus (non compris le calcul des traits par le D.A.P.). Il faut toutefois noter que ces temps s'allongeront forcément lorsque le lexique dépassera la dizaine de milliers de mots ; on peut envisager une vitesse de l'ordre d'un mot par seconde sur un ordinateur de même catégorie.

On peut indiquer qu'actuellement les ensembles proposés ne contiennent que quelques mots, mais ceci dépend fortement de la taille du lexique. Il est raisonnable de penser que ces ensembles ne dépasseront que rarement quelques dizaines de mots.

Enfin, il faut signaler qu'à ce jour les accès n'ont pas encore été utilisés sur des phrases, mais seulement sur des exemples construits à la main. Ceci est dû à la réalisation en parallèle des différents modules du système de reconnaissance de mots.

CONCLUSION

La méthode exposée est simple à la fois dans sa définition et dans sa mise en œuvre. Il s'agit d'un principe probablement perfectible.

L'un des aspects importants tient dans la symétrie entre le traitement du lexique et celui de la phrase à analyser. Cette ressemblance vient des connaissances décrivant le comportement moyen du D.A.P.

La dépendance vis-à-vis du locuteur doit être précisée par un traitement des phrases du corpus de test prononcées par plusieurs personnes. Elle induira de nouvelles règles indiquant quels sont les traits fiables indépendemment du locuteur, ce qui doit donner des accès à des cohortes communes à tous les locuteurs. Ensuite, des accès personnalisés définis toujours de la même manière pourront fragmenter ces ensembles. Il reste à espérer que le D.A.P. fournisse des traits assez résistants...

REFERENCES

[1] H. Méloni, J. Gispert, J.Guizol
"Traitement de connaissances déclaratives pour l'identification analytique de mots dans le discours continu"
14ème JEP Paris 1985

[2] H. Méloni
"Etude et réalisation d'un système de Reconnaissance Automatique de la Parole Continue"
Thèse de Doctorat d'Etat, Université d'Aix-Marseille II, Faculté des Sciences de Luminy Février 1982

[3] H. Méloni, R. Bulot
"Reconnaissances des Formes et Segmentation"
16ème JEP Hammamet, Tunisie, Octobre 1987

[4] J. Guizol
"Inférence automatique de règles : quelques résultats"
16ème JEP Hammamet, Tunisie, Octobre 1987

[5] J.Gispert
"Représentation d'un lexique à l'aide de connaissances de Phonologie Générative en R.A.P.C."
15ème JEP Aix en Provence, Mai 1986

[6] G. Gougenheim, R. Michéa, P. Rivenc, A. Sauvageot
"L'élaboration du Français Fondamental"
Didier Editeur, 1964

[7] D. Shipman, V.W. Zue
"Properties of large Lexicons : implications for advanced isolated word recognition systems"
IEEE 1982

[8] L.F. Lamel, V.W. Zue
"Properties of consonnants sequencies within words and accross words boundaries"
Congrès I.C.A.S.S.P. 1984

[9] W.A. Woods
"Optimal search strategies for speech understanding control"
Artificial Intelligence 18, 1982, 295-326

[10] G. Perennou
"Base de données lexicales Rapport scientifique"
GRECO communication parlée Juin 1984 CRIN Nancy

[11] G. Adda, M. Eskénazi, P.E. Stern
"Reconnaissance de grands vocabulaires : utilisation et évaluation de traits grossiers"
15ème JEP Aix en Provence, Mai 1986

INFERENCE AUTOMATIQUE DE REGLES : QUELQUES RESULTATS

Jacques Guizol

GIA - Faculté de Luminy - 70 Route Léon Lachamp 13288 Marseille Cedex 9

Abstract

The obtaining of expert rules in any field and especially in speech recognition always comes up against several problems, such as the extraction of implicit knowledge, formalization, exhaustivity, etc. Numerous methods can be used but these often produce disappointing results regarding the difficulty of implementation.

For these reasons, we have taken an interest in automatic learning using examples which describe various concepts (accoustic events, phonemes, syllables, features, cues, prosodic marks,...) in symbolic form.

In this paper we present some of the results we have obtained concerning voiced fricative sounds in particular.

1 - Introduction

Le décodage acoustico-phonétique, qui constitue un module-clé dans un système de reconnaissance automatique de la parole, peut être abordé de diverses façons (clustering, schémas d'affectation procéduraux, règles expertes ...). En tout état de cause, cette phase s'appuie sur des observations et connaissances issues d'études menées essentiellement par les phonéticiens ([1] pour ce qui nous interesse ici), et concernant la détermination d'indices acoustiques caractéristiques des sons considérés, et dont la pertinence peut être vérifiée à postériori par des tests perceptifs.

Ces indices sont eux-mêmes déterminés grâce à des paramètres divers qui bien souvent constituent une représentation idéale du phénomène par sa longueur, son intensité, la répartition fréquentielle de son énergie spectrale, certaines micro-variations à caractère discriminant, etc. Le modèle ainsi décrit doit ensuite être adapté à la réalité en fonction de l'accentuation, du contexte, du débit, du locuteur ...

Du point de vue conceptuel, le problème est donc de trouver une "bonne idéalisation" d'un phénomène étudié, puis les règles transformationelles pertinentes permettant de passer du modèle aux réalisations effectives.

Du point de vue opérationnel, certains des ajustements nécessaires seront difficilement réalisables par manque d'informations suffisantes dans un schéma remontant de la reconnaissance [2]. Dans ces conditions, doit-on appliquer ces ajustements de façon indéterministe pour essayer de faire "coller" avec la réalité (au risque de se heurter à l'inévitable problème de l'explosion combinatoire plus ou moins maitrisée) ?

Vu l'investissement nécessaire au premier point et l'incertitude qui réside dans le second, nous avons essayé d'automatiser l'obtention de règles de caractérisation prenant implicitement en compte l'influence du contexte, du locuteur, etc.

Nous ne reviendrons pas ici sur la présentation du système d'apprentissage utilisé décrit dans [3], notre propos se limitant à exposer quelques-uns des résultats obtenus. Toutefois, nous insisterons sur l'intérêt que revêt le caractère systématique du processus dont le rôle est de mettre en évidence des invariants contenus dans des échantillons réels et permettant, le cas échéant, de dégager l'influence contextuelle, celà pour un locuteur quelconque.

Le danger de cette méthode tient au caractère non exhaustif des configurations du domaine. Les règles obtenues vérifieront donc une complétude relative aux seuls exemples considérés. Néanmoins, le traitement étant automatique, ceux-ci pourront être choisis en grand nombre, de façon à ce que leur ensemble ait une forte probabilité de représentativité des diverses possibilités d'occurrences

2 - Les données

Les exemples considérés par le système représentent des événements grâce à des configurations temporelles de formes décrivant les variations des paramètres considérés. Ces paramètres sont essentiellement des résultats bruts de l'analyse du signal, donnant des informations sur le comportement spectral (énergie utile, énergie basses-fréquences, hautes fréquences, position des centre de gravité hautes et basses fréquences...), des indications temporelles (densité de passages par zéro, premier coefficient de la fonction d'auto-corrélation), les évolutions des maxima du spectre, etc.

Ils peuvent aussi être obtenus par combinaison linéaire des précédents (distance entre centres de gravité, différence d'énergie entre hautes et basses fréquences, entre basses fréquences et spectre utile ...).

Les exemples sont des conjonctions d'événements primaires associés aux paramètres considérés et pour chacun desquels sont indiqués l'amplitude, le positionnement temporel et le degré de recouvrement par rapport à chacun des autres [4].

Si l'on appelle :

- I l'ensemble des instances fournies au système,
- E l'ensemble des événements primaires possibles,
- V_e l'ensemble des valeurs associées à chaque $e \in E$,
- N_i l'ensemble des noms identifiant les divers événements primaires d'une instance $i \in I$,
- T l'ensemble des relations de positionnement,

alors tout exemple peut être défini de la façon suivante :

$$\forall i \in I \{ \forall n \in N_i \{ \exists e \in E \text{ et } \exists v \in V_e / e(n, v) \} \wedge \forall p \in N_i \{ \exists t \in T / t(n, p) \} \}$$

- Chacune des règles trouvées sera une conjonction d'événements éventuellement disjonctifs -la disjonction indiquant un indéterminisme sur le caractère (*Exemple : vallée d'énergie basse fréquence ou d'amplitude du premier pic*) ou sur l'amplitude- et de relations de situation,

- La généralisation sur V_e et T sera opérée en introduisant sur ces ensembles une relation d'ordre total,

- Chacun des exemples devra être cohérent avec une au moins des règles obtenues. L'ensemble R des règles trouvées devra donc satisfaire la condition suivante :

$$\forall i \in I \{ \exists r \in R \text{ et une substitution } \sigma / i \supseteq r_\sigma \}$$

3 - Résultats obtenus sur les fricatives voisées

Dans les données dont nous disposions pour faire cette étude, les événements considérés étaient des collines et vallées décrites

dans le temps par le premier coéfficient d'auto-corrélation (*er0*), les énergies hautes et basses fréquences (*ehf* et *ebf*), l'energie spectrale (*esp*), la densité de passages par zéro (*dpz*), la fréquence du premier pic (*fp1*) et son amplitude (*ap1*).

Les distinctions globales entre les divers phonèmes de la catégorie considérée ici, sont fournies par l'ensemble d'événements invariants et pour chacun de ceux-ci, par sa position relative par rapport aux autres et au noyau consonantique.

Plus précisément et à titre d'exemple, alors que la réalisation des phonème /ʒ/ ou / z / laissera systématiquement apparaître une *colline-dpz* avec un niveau d'intersection avec le noyau plus grand en moyenne pour le premier que pour le second, cela ne sera pas toujours le cas pour / v /.

Nous avons choisi de présenter les résultats essentiellement grâce à des figures, pensant que ce moyen offrirait la possibilité d'une interprétation plus aisée que ne l'autoriseraient des règles. Nous donnerons donc pour chaque règle fournie par le système un tableau indiquant les niveaux de recouvrement minimum pour chaque couple d'événements primaires (ceux de faible amplitude apparaitront sur fond grisé), leur position mutuelle (dans le cas où elle est constante) et un exemple visualisant une instance vérifiée.

Le tableau TR délivrera pour chaque événement, un taux de recouvrement par rapport à chacun des autres :
- si X est un événement-ligne, Lg(X) la longueur de X, et Y un événement-colonne, on aura :

$$Lg(X \cap Y) \geq Lg(X) \times TR(X, Y)$$

- de la même façon on aura :

$$Lg(Y \cap X) \geq Lg(Y) \times TR(Y, X).$$

Ce tableau fournit donc en définitive les événements indispensables à la caractérisation d'un phénomène et un système d'inéquations associé.

Un individu, pour être affecté à une classe devra donc :

- contenir les événements apparaissant dans la règle,
- vérifier les contraintes de positionnement,
- vérifier le système d'inéquations.

3.1 - Phonème / v /

La difficulté de caractérisation du phonème / v / se concrétise par un nombre de règles plus important que pour / z / et /ʒ/, preuve d'une plus grande dispersion dans les réalisations.

- Dans le cas où le contexte droit est de nature fermée, un / v / se caractérise par une *vallée-er0* de forte amplitude et une *vallée-esp* d'amplitude moins importante, la première ayant tendance à être plus centrée que l'autre.

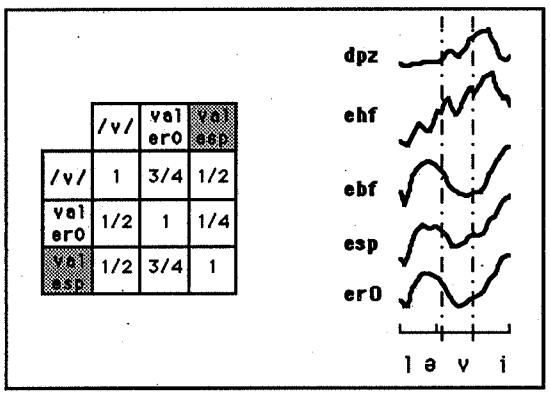


Figure 1 : /v/ en contexte droit fermé

A noter qu'une règle-exception (de pertinence plus faible) autorise l'apparition d'une *vallée-ebf*.

- Dans le cas où le contexte droit est non fermé, on obtient une caractérisation faisant intervenir des vallées *esp*, *er0*, *ebf* et *ehf*, celle coïncidant le mieux avec le noyau étant la *vallée-esp*. En effet, par rapport au cas précédent, celle-ci est limitée dans son expansion à droite du fait de la nature même du contexte. Les relations de positionnement sont beaucoup mieux précisées.

A noter là encore une règle-exception autorisant l'apparition d'une*colline-ehf* dans la partie droite du noyau.

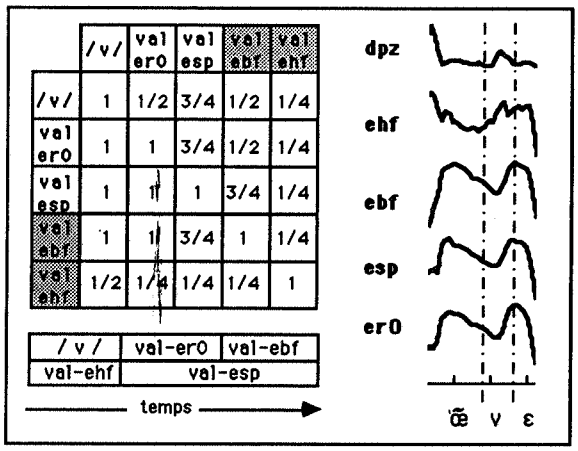


Figure 2 : /v/ en contexte droit non fermé

- Deux règles ont été obtenues pour les cas où le contexte droit est aigu. La première s'applique lorsque le contexte droit est fermé et a donc peu d'énergie, la deuxième vérifiant les autres cas.

L'une se caractérise par un léger pic d'*ehf* qui sera interrompu à la transition pour se rétablir de façon évidente ensuite. C'est en particulier le cas où le contexte droit est la voyelle / i / ou la semi-consonne / j / suivie d'une voyelle aiguë.

L'autre ne comportera pas ce pic d'*ehf* qui n'atteindra son maximum que sur la voyelle adjacente. Par contre, on notera une *vallée-ebf* qui n'apparaissait pas précédemment, le minimum étant atteint sur le phonème suivant.

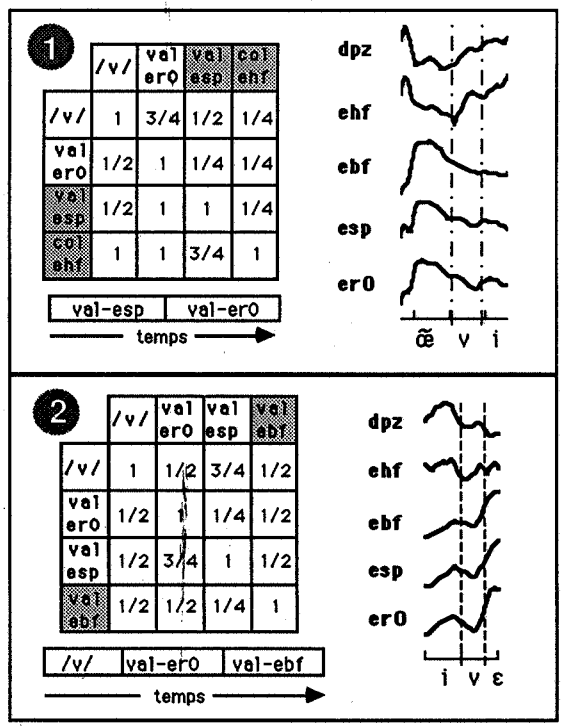


Figure 3 : /v/ en contexte droit aigu

- Dans le cas où le contexte droit est non aigu, la caractérisation est unique sur les exemples traités et fait apparaître un resserrement très net des vallées *ebf*, *esp* et *er0* par rapport au noyau du phonème étudié.

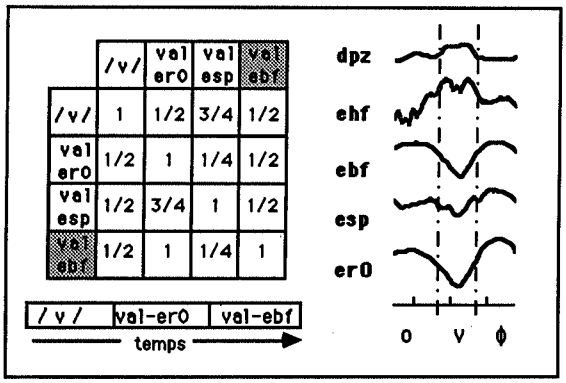


Figure 4 : /v/ en contexte droit non aigu

• Enfin, dans le cas où le contexte droit est la semi-consonne / j / ou la liquide / r /, une règle nous indique l'apparition systématique d'une *colline-dpz* très nette.

Phonème / z /

A l'inverse du phonème / v /, / z / se caractérise par l'apparition plus ou moins nette mais systématique d'une *colline-dpz*. Toutefois, celle-ci ne sera jamais la mieux centrée ou la plus large par rapport au noyau comme c'est le cas pour / ʒ / que nous verrons ensuite.De plus, la discrimination avec / v / se fera grâce à une *colline-ehf* présente dans toutes les règles sauf une.

• Dans le cas où / z / est en contexte droit fermé, nous obtenons une seule règle mettant en jeu essentiellement une *vallée-er0* et une *vallée-ebf* (qui sont les mieux centrées) ainsi qu'une *colline-ehf* et une *colline-dpz*.

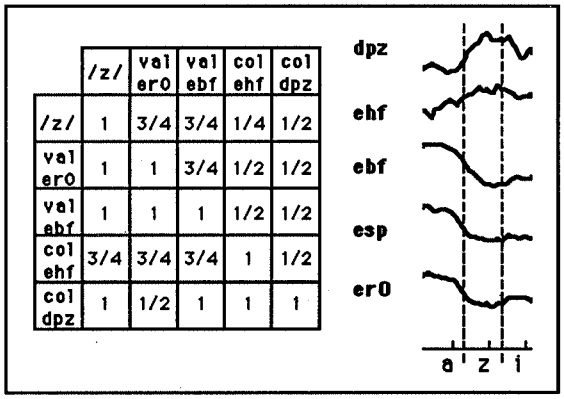


Figure 5 : /z/ en contexte droit fermé

• En contexte droit non fermé, la caractérisation s'opère par deux règles selon que le contexte gauche est plus ou moins ouvert.

Dans le premier cas, cinq paramètres interviennent avec une grande amplitude : une *vallée-esp*, une *vallée-ebf* et une *vallée-er0*, bien centrées sur le noyau, ainsi qu'une *colline-dpz* et une *colline-ehf*. Le phonème est donc bien défini et "limité".

Dans le cas où le contexte gauche est moins ouvert, voire consonantique ou à l'attaque, la caractérisation perd en précision. En particulier, on a une double possibilité d'événement concernant la chute d'énergie (*esp* ou *er0*), moins bien centrée et moins nette que précédemment ; d'autre part, la *vallée-ebf* n'étant plus systématique, elle disparaît de la règle.

• Mise à part une exception où le phonème / z / apparaissait au niveau d'une frontière prosodique entre deux voyelles aiguës, la caractérisation obtenue en contexte droit aigu ne nécessite qu'une règle dans laquelle, outre une *vallée-er0* très nette et bien centrée sur le phonème, on dispose d'une *colline-ehf* et d'une *colline-dpz* tendant à se prolonger vers la droite.

• Enfin, nous avons obtenu une règle unique caractérisant / z / en contexte droit non aigu grâce à cinq formes ayant la propriété d'être toutes contenues dans le noyau. Parmi celles-ci, nous

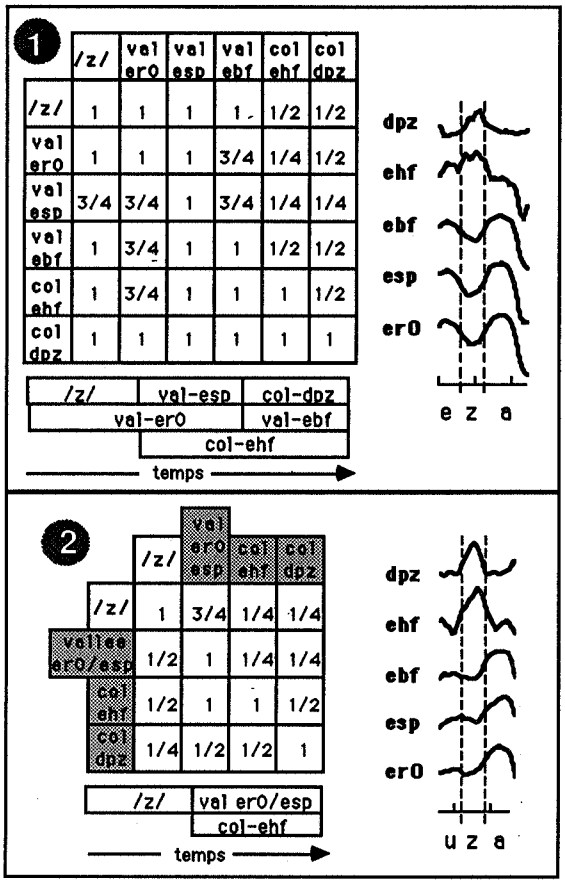


Figure 6: /z/ en contexte droit non fermé

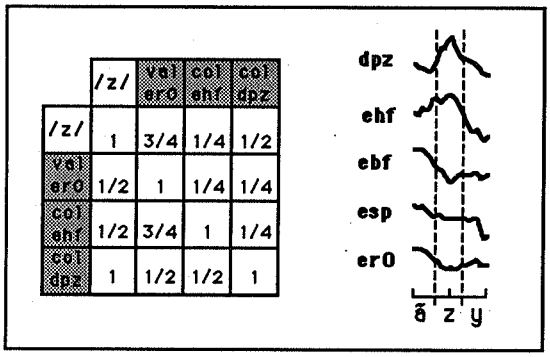


Figure 7: /z/ en contexte droit aigu

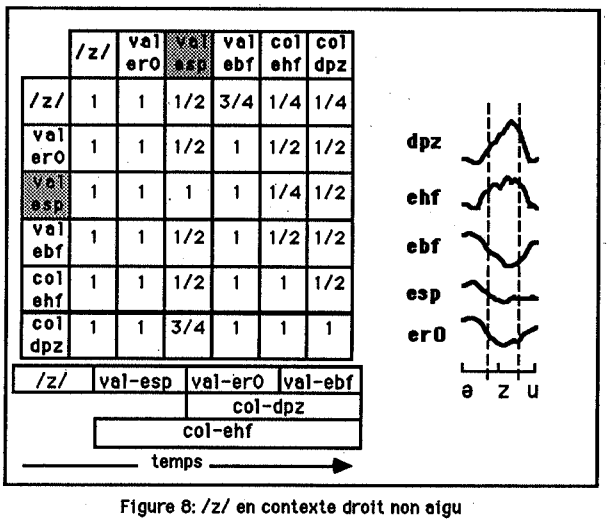


Figure 8: /z/ en contexte droit non aigu

trouvons, outre une *colline-ehf* de forte émergence, une *vallée-esp* et une *colline-dpz* éventuellement moins nettes, deux profondes vallées (*er0* et *ebf*) recouvrant bien le noyau.

3.1 - Phonème /ʒ/

La différence essentielle dans la caractérisation de ce phonème par rapport au deux précédents réside dans la présence systématique d'une *colline-dpz* contenue dans le noyau et plus ou moins large selon que le contexte est aigu ou non. Autre constatation semblant se dégager d'une caractérisation grossière que nous avons obtenue, le regroupement des instances comportant un contexte droit nasal et/ou aigu. Une analyse plus fine (en particulier en étant plus exigeant sur les contraintes imposées aux règles) nous a permis d'obtenir les résultats suivants :

• Les réalisations de / ʒ / en contexte droit aigu ou fermé sont regroupées et donnent naissance à deux règles.

La première concerne les instances dont le contexte gauche possède une grande énergie dans les basses fréquences (/ i /, / r /, / u /, / o /, etc. et le "buzz" d'attaque). La caractérisation qui en découle comporte outre une *colline-ehf* et une *colline-dpz* de grande amplitude, une *vallée-ebf* de moindre importance.

Dans les autres cas, cette *vallée-ebf* n'est plus caractéristique et disparaît donc de la description. D'autre part, une *colline-dpz* moins importante que dans le premier cas est tolérée. Enfin, les deux collines sont bien contenues dans le noyau, leur largeur étant plus réduite.

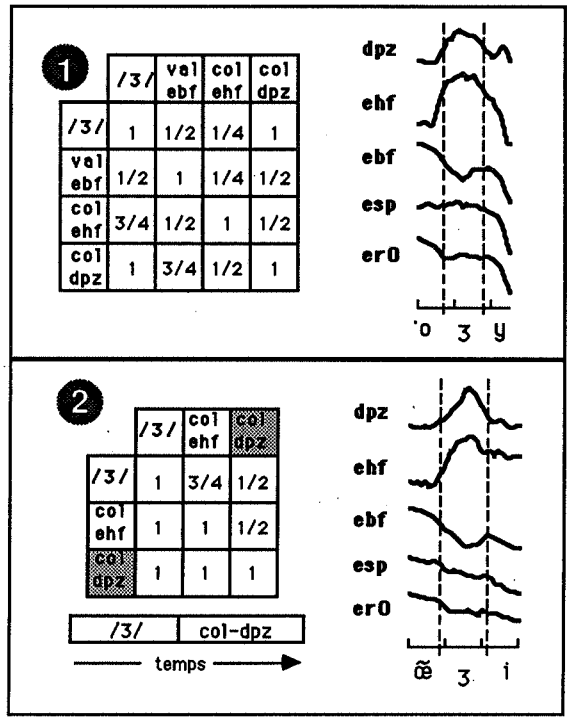


Figure 9: /3/ en contexte droit fermé ou aigu

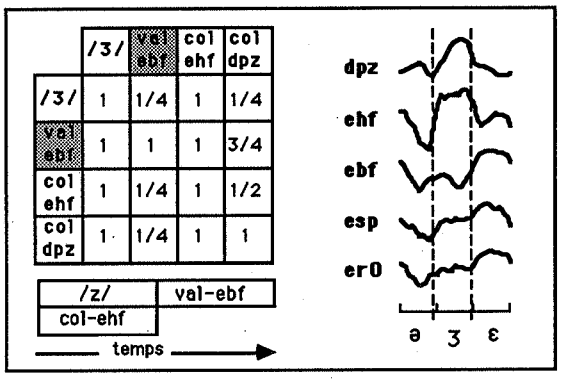


Figure 10: /3/ en contexte droit non fermé

• Lorsque le contexte contexte droit est ouvert, on retrouve comme précedemment une *colline-ehf* et une *colline-dpz* bien marquées ainsi qu'une *vallée-ebf*, mais dans le cas présent, la *colline-ehf* occupe en permanence toute la largeur du noyau. A l'inverse, l'ouverture du conduit vocal peut provoquer un raccourcissement du temps de friction et par conséquent de la longueur de la *colline-dpz*. D'autre part, les indications temporelles montrent bien le déport de celle-ci vers la droite

• Enfin, l'influence d'un contexte droit grave se manifeste par une prédominence d'une *colline-ehf*, mais aussi par la présence d'une*vallée-ebf* significative alors que la *colline-dpz* perd en hauteur et en largeur.

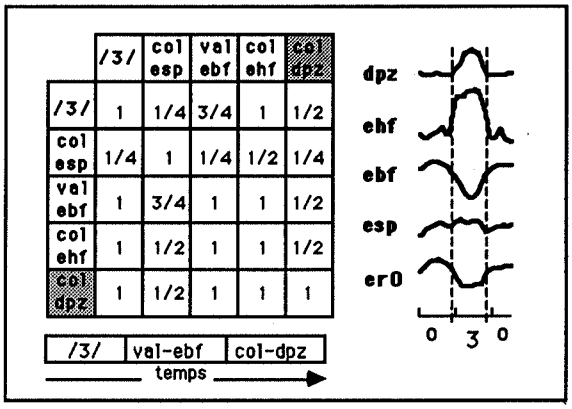


Figure 11: /3/ en contexte droit grave

4 - Conclusion

Les résultats que nous venons de présenter, qu'ils soient déjà connus ou inattendus, sont, rappelons le, issus d'un ensemble d'exemples non choisis et non triés. Les règles obtenues sont donc de ce fait corrélées à leur répartition. Ce que nous voulions montrer plus précisément, était la faisabilité de cette approche symbolique libérée de contraintes de seuils utilisés "au premier degré", prenant en compte de façon implicite les caractéristiques du locuteur et permettant de supprimer toute information reconnue non pertinente pour la description des concepts considérés.

Toutefois, le problème des ambigüités, apparaissant lors de la caractérisation de deux objets proches, reste à traiter.

Références

[1] Chafcouloff M., Di Cristo A., Seimandi L.
"*Effets de la Coarticulation sur les Caractéristiques Acoustiques des Contoïdes Fricatives du Français*"
T. I. P. A. 3, pp.61-113, (1976).

[2] Gispert J.
"*Les Accès: Interface Entre le Décodage Acoustico-Phonétique et le Lexique*"
16èmes J. E. P. , Hammamet : 3-10 octobre 1987.

[3] Guizol J.
"*Apport de l'apprentissage au décodage acoustico-phonétique en reconnaissance automatique de la parole*"
7èmes J. I. S. E., pp. 1489-1510, Avignon : 13-15 mai 1987

[4] Meloni H., Bulot R.
"*Reconnaissance des Formes et Segmentation*"
16èmes J. E. P. , Hammamet : 3-10 octobre 1987.

REMORA Représentation orientée objet pour la collaboration de connaissances procédurales et déclaratives

T. MARTELLI, J-P TUBACH, L. MICLET

ENST - Dept SYC - UA 820 CNRS -
46 rue Barrault 75013 PARIS

ABSTRACT

REMORA, (**RE**presentation et **M**odelisation **O**bjet pour la **R**econnaissance **A**coustico phonetique: object representation and modelling for the acoustic phonetic recognition) is a tool for helping phonetic decoding. The basic idea of this project, realized at ENST, is to include various kinds of knowledge to improve accuracy and efficiency. The procedural knowledge may come from several sources such as centisecond segment recognition, diphone spotting or segmentation. The declarative knowledge may come from expertise on speech signal processing or from phonetic knowledge. In order to connect these two types of knowledge and to satisfy modularity and efficiency, we use an object-oriented formalism. The programming environment is a set of interactive menus, designed to be easily employed by the experts.

1 - INTRODUCTION :

Dans sa version actuelle, le système REMORA est une architecture appliquée au décodage acoustico phonétique de la parole continue. Si son but final est la production d'ur treillis phonétique, à partir de connaissances procédurales et déclaratives et la sélection d'heuristiques permettant d'obtenir les meilleures performances, au fur et à mesure de l'enrichissement du système, son objectif à court terme est de constituer une station de travail pour l'expert, au sens où nous le décrirons dans la suite.

Au stade de développement actuel, l'accent est mis sur l'acquisition des connaissances et leur utilisation, ce qui consiste donc à les structurer et les évaluer à l'aide de REMORA. La priorité est donnée à l'utilisateur, le système constituant en premier lieu, un outil efficace pour l'expert: Il lui fournit un formalisme puissant de représentation des connaissances, qu'il peut utiliser par l'intermédiaire de menus interactifs, afin d'enrichir la base de connaissance, et de soumettre, à son gré, des problèmes.

Le décodage acoustico phonétique est un problème clé en reconnaissance de la parole continue, en effet, les erreurs du niveau phonétique, se reportant dans les niveaux supérieurs, vont augmenter les ambiguïtés et erreurs sur les mots et les phrases (5). Pour améliorer les performances, il est nécessaire de réduire l'indéterminisme provenant du niveau acoustico phonétique. Des systèmes s'intéressant particulièrement à ce niveau sont donc développés (1), (3), (4), (9), (12), (16), il est également pris en compte, dans des systèmes complets de reconnaissance (11).

On présente tout d'abord le système: les objectifs et les motivations qui ont conduit au choix d'une représentation centrée-objet, et à la construction de son architecture. Le formalisme et les connaissances utilisées ne seront pas détaillées dans cet article, pour plus de détails on se reportera à (6 - 8). Le système est présenté du point de vue de l'utilisateur, en s'intéressant à l'interface.

2 - LES OBJECTIFS :

L'amélioration des performances en décodage acoustico phonétique requiert la collaboration de sources de connaissances diverses tant procédurales que déclaratives de façon à bénéficier de l'expérience déjà acquise dans le domaine (13). En effet les taux de performances des algorithmes actuels pour la reconnaissance se situent au mieux entre 70 et 80%.

D'une part, l'intérêt de regrouper dans un même système divers algorithmes de reconnaissance permet de les évaluer et de comparer leurs performances dans le contexte d'un problème à résoudre. On obtient ainsi de meilleurs taux de performances en liant par exemple le choix d'un algorithme au contexte de reconnaissance pour lequel ses performances sont les meilleures ou bien grâce à la confrontation de résultats de divers algorithmes sur les mêmes données, on confirme ou infirme les taux de confiance des hypothèses obtenues.

D'autre part, il est intéressant de prendre en compte des connaissances, qui ne peuvent s'exprimer de façon algorithmique: ce sont celles d'un expert humain, nous distinguons:

- l'expertise en traitement du signal de parole, qui permet d'associer à un segment de signal des traits acoustico-phonétiques.
- l'expertise d'un programmeur sur ses propres programmes, en particulier l'interprétation qu'il fournit des résultats , qu'il s'agisse des échecs ou des succès, en fonction du contexte.

En conclusion: la nécessité de prendre en compte ces types de connaissances variés a conduit à définir un formalisme adapté qui permet de représenter et d'utiliser avec efficacité les connaissances de façon évolutive, d'assister l'expert en l'aidant à formaliser son expertise et à la structurer. Le système constitue un outil de travail, de mise au point et d'intégration de connaissances portant sur le décodage acoustico phonétique (10). Dans une optique d'auto amélioration, le système peut tester ses propres heuristiques de façon à améliorer leurs performances, il en va de même tant pour les connaissances procédurales que pour les connaissances déclaratives. D'autre part le système se propose de favoriser l'émergence d'une caractérisation de traits acoustiques liés à des événements phonétiques.

3 - UN OUTIL POUR L'EXPERT :

L'examen d'une part des connaissances à représenter, spécifiques au traitement de la parole, ainsi que leur mode d'acquisition et d'autre part leur mise en œuvre ont préludé à l'élaboration du formalisme de Remora et à la définition de son architecture.

Le but premier du système étant de constituer un outil pour l'utilisateur, la base de connaissance est repertoriée en fonction de ses critères. On est donc amené à distinguer cinq types de connaissances à représenter, ce qui constitue ce que l'on appelle la structuration externe.

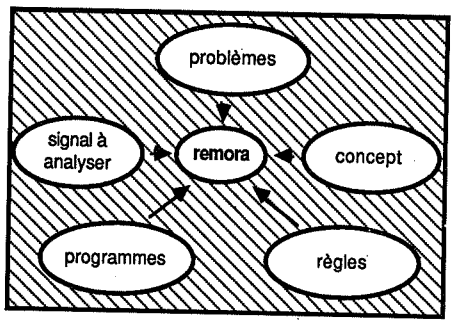


Figure 1
la structuration externe des connaissances

On dispose d'une bibliothèque de programmes et de leurs résultats, de fichiers de signaux à analyser, et de l'expertise de l'utilisateur. Nous avons déja défini ce que nous entendons par expertise de l'utilisateur. Il est néanmoins nécessaire de remarquer que cet expert a besoin d'exprimer d'une part des concepts, qui représentent des connaissances acoustiques et phonétiques, (comme par exemple une fenêtre d'analyse, un formant, un burst) et d'autre part des heuristiques ou des méta connaissances, qui expriment une série de raisonnement sur les connaissances elle mêmes, des heuristiques d'amélioration sur les résultats délivrés par ses programmes etc... Ce qui nous donne les deux types de connaissances, que l'on appelle pour plus de commodité "entités" et "règles" . L'utilisateur désire également pouvoir représenter les problèmes qu'il veut voir résolus par le système, et qui sont en fait des actions à réaliser. La structuration externe des connaissances, consiste donc à associer à chaque type de connaissances un module d'interface spécifique qui dispose d'un contrôle intégré, sous forme d'un objet particulier, qui lui permet de proposer à l'utilisateur un cadre de description réutilisable et extensible. La tâche de l'utilisateur consiste à ajouter ou à supprimer des fonctionnalités, c'est à dire à compléter le modèle de description proposé par le système.

De fait, les connaissances sont restructurées de façon interne dans le formalisme du système, qui souscrit à une représentation centrée-objet. Donnons en une brève présentation: tout d'abord, l'intérêt d'une telle représentation est de façon générale la facilité de création et de modification d'une base de connaissance (14). Dans le cas présent, pour donner au système une grande efficacité, toute entité, qu'elle soit procédurale ou déclarative, aussi bien que les entités nécessaires au fonctionnement du système, sont représentées sous forme d'objets. Ceci permet de considérer toute connaissance de la même manière. Il est également intéressant pour un problème donné de regrouper en un objet toutes les informations le concernant, et également pour un programme, de donner des indications le concernant qui ne peuvent être traduites sous forme de code.

Objet est un terme générique qui désigne aussi bien les classes que les instances. Une classe est caractérisée par ses attributs et ses méthodes. Une instance est générée par une classe. Un attribut peut contenir des valeurs ou des contraintes ou des sélecteurs qui précisent le domaine de définition des valeurs et permettent d'établir la structuration interne. Les méthodes sont des procédures d'exploitation des objets,

elles désignent les actions que peut réaliser un objet sur réception d'un message provenant d'un autre objet, elles ne sont pas obligatoirement créées en même temps que leur classe, elles peuvent être soit des programmes soit des règles, et seront de toute façon représentées par des objets. Les objets communiquent entre eux par l'intermédiaire de messages. Une instance est créée par affectation de valeurs aux attributs de la classe. Un tel formalisme permet de réaliser de manière aisée l'attachement de procédures et de règles, il permet de modifier avec facilité les composantes internes du système, l'indépendance des connaissances étant intrinsèquement assurée. L'héritage entre classes concerne les méthodes et les attributs d'une classe, ils sont transmis à ses sous classes. Une classe conserve la liste de ses propres instances (2).

Définir un mode de représentation pour que l'expert décrive son expertise de façon simple, est un problème d'acquisition de connaissances, qui nous a amené à élaborer une interface utilisateur (5), composée de menus. L'utilisateur est guidé dans sa tâche d'élaboration et de test de nouvelles connaissances par un ensemble d'objets qui lui permettent de décrire et de structurer ses connaissances indépendamment de la structure interne du système.

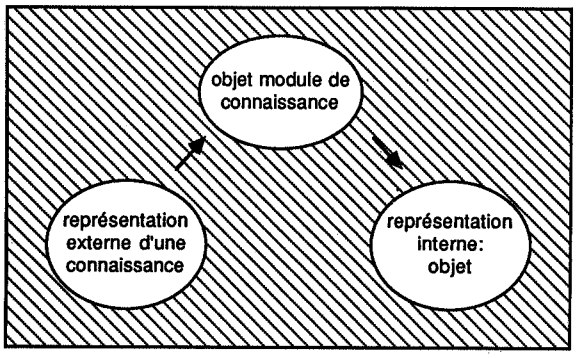


figure 2 :
la transformation dans le formalisme interne

L'expert, conformément aux modèles proposés, construit des objets où il stocke ses connaissances, qui se réfèrent à des entités phonétiques abstraites, constituant un vocabulaire de mots clés; ces derniers éclairent un aspect spécifique des connaissances (parties stables, formants etc...) ou se réfèrent à un traitement particulier. Ces objets constituent un véritable vocabulaire de base: chaque objet posséde ses procédures ou règles attachées et représente un package ou groupe de connaissances associé à un domaine spécifique d'expertise. D'autre part, l'intérêt de ne pas introduire la connaissance en utilisant des prédicats consiste pour l'expert, à ne pas avoir à mémoriser ni l'ordre des paramètres, ni l'arité des prédicats. Il dispose par l'intermédiaire des menus d'un langage simple de représentation de connaissances.

moule de description d'un programme:		
nom de l'auteur	domaine	chaine de caracteres
nom de programme	domaine	chaine de caracteres
données	domaine	fichier-signal
etc...		

Figure 3 :
un exemple de langage de description

La figure 3 donne un exemple du langage de description mis à la disposition de l'utilisateur. L'indication de domaine sert à indiquer le type autorisé, il peut s'agir d'un entier, d'une chaîne de caractères ou d'une classe d'objet déjà définie; le rôle de ce sélecteur est de réaliser une structuration interne, en effet lors de l'instanciation de la valeur de l'attribut possédant une telle facette, il y aura un contrôle de type.

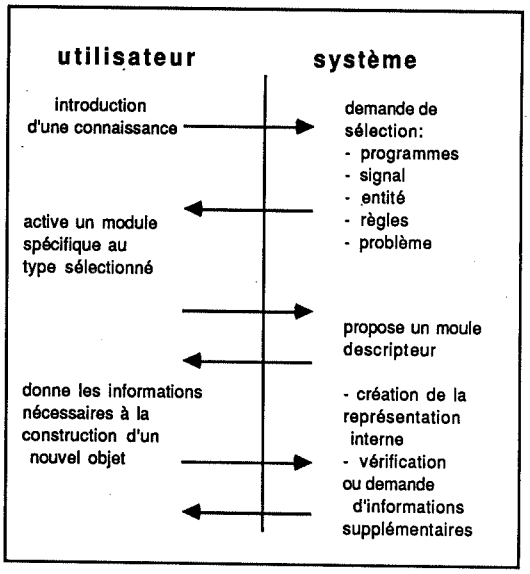


figure 4:
un exemple d'introduction de connaissance

On propose deux principaux services à l'utilisateur par l'intermédiaire de l'interface : d'une part un service de gestion de connaissances, d'autre part un service de création, modification et destruction et un service de soumission d'une tâche

Il est à noter que la structuration interne, qui n'est pas détaillée ici (6), est faite de façon transparente pour l'utilisateur; elle permet de prendre en compte des liens qui ne sont pas des liens d'héritage entre objets. Il s'agit par exemple de lier un objet et les parties qui le constituent, c'est le cas, par exemple d'un fichier de signal ou de résultats et de leurs contenus. D'autre part, par l'intermédiaire des facettes de

sélection, on peut lier un programme à la classe de problèmes qu'il peut résoudre.

Considérons maintenant un exemple d'utilisation:

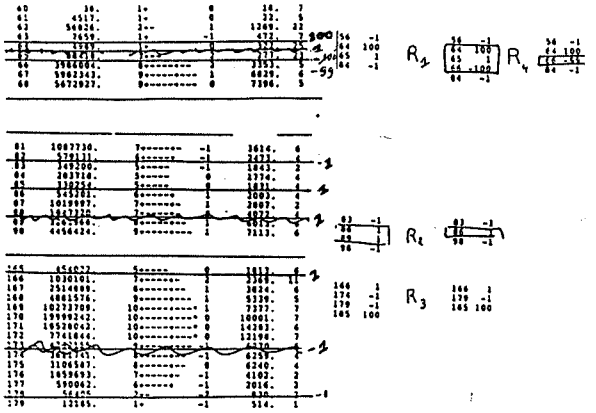


Figure 5:
un exemple d'utilisation

un utilisateur veut donner des règles d'amélioration sur les résultats de son programme. Il introduit des règles d'amélioration, qui converties dans le formalisme interne lui permettent d'optimiser les performances. Les règles sont vues comme des méthodes, associées aux résultats du programme. Ceci permet, par exemple, à l'utilisateur, dans le cas d'un programme de segmentation d'éliminer les segmentations redondantes ou de rectifier des indices de segmentation.

4 - CONCLUSION :

Le système REMORA constitue en fait, un poste de travail qui permet, par exemple, à partir de phrases connues, de valider un étiquetage provenant de différentes sources déclaratives ou procédurales (en exploitant leur aspect complémentaire) : l'utilisateur peut donc tester ces différentes sources et si nécessaire dégager la nécessité éventuelle d'un traitement complémentaire ou le recours à l'utilisation d'autres algorithmes.

Ceci nous conduit donc à définir une nouvelle dénomination pour notre système: nous dirons qu'il consistue un poste d'expertise. Le système, en cours de réalisation, sera appliqué et validé en reconnaissance sur un corpus plus important intégrant éventuellement l'aspect multilocuteur.

Références :

(1) Damestoy J.P. ; Réalisation d'un système à base de prototypes pour le contrôle du décodage acoustico-phonétique de la parole;Thèse de doctorat d'université de NANCY I , janvier 1986.

(2) Ferber J. ; Mering un langage d'acteur pour la représentation et la manipulation des connaissances, Thèse 83 Paris VI

(3) Gilloux M., Mercier G., Tarrideg C., Un système expert pour la reconnaissance de la parole; Colloque International d'Intelligence Artificielle Marseille Oct 84, pp 30-40

(4) Guizol J. ; Apprentissage inductif de règles pour le décodage acoustico-phonétique ; 15ème JEP, Aix en Provence 27-29 MAI 1986 .

(5) Haton J. P. ; Intelligence artificielle en compréhension automatique de la parole ; TSI vol 4 n°3 mai 85

(6) Martelli T., Miclet L., Tubach J-P, REMORA A software architecture for the collaboration of different knowledge sources in phonetic decoding of continuous speech, ICASSP 87, vol 1 pp 387-390

(7) Martelli T., Rapport au Gréco-CNRS; Avril 87

(8) Martelli T., Miclet L., Tubach J-P, Configuration de base du système REMORA RFIA nov 87 (à paraitre)

(9) Meloni H. , Bulot R. ; Décodage acoustico-phonétique en Prolog ; 15ème JEP,Aix en Provence 27-29 MAI 1986 .

(10) De Mori ; Extraction of acoustic cues using a grammar of frames, Speech Communication 2, 1983

(11) Pierrel J.M ; Etudes et mises en oeuvre de contraintes linguistiques en compréhension automatique du discours continue : le système Myrtille I et le système Myrtille II; Thèse d'Etat 81 Nancy I

(12) Stern P.E., Eskenazi M., Memmi D. An expert system for speech spectrogram reading, ICASSP 86, vol 2

(13) Vicard D., Miclet L.; Reconnaissance de parties stables de parole continue pour le décodage acoustico phonétique ; 15ème JEP, Aix en Provence 27-29 MAI 1986 .

(14) Niwa K. , Sakasaki K. , Ihara H. ; An experimental comparison of knowledge representation schemes; The AI Magasine, vol 5, n°2, 1984

(15) Zue V.W and Lamel L.F. An expert Spectrogram Reader: A knowledge based approach to speech recognition, ICASSP 86, vol 2

Détection d'indices par quantification vectorielle et réseaux markoviens

G. Bailly, J.P. Liu

Laboratoire de la Communication Parlée - ICP, unité associée au CNRS
INPG/ENSERG
46, Av. F. Viallet, 38031 Grenoble Cedex

Abstract

Classical approach to acoustic-to-phonetic decoding in feature-based recognition systems consists in using cue detectors in parallel . These detectors are tuned in order to simulate the properties of the underlying linguistic cues. The gap between acoustic correlates and phonetic cues could be filled by intensive use of knowledge-based systems which attempt to modify heuristics and rules according to context, noise level. Another approach is described here which describes the congruency between an acoustic space and a cue space by means of a Hidden Markov Model (HMM). Application to formant detection is described which opens this approach to acoustic-to-articulatory congruency.

A- INTRODUCTION

En Décodage Acoustico-Phonétique, l'attribution du titre d'indice à un corrélat acoustique est monnaie courante bien que la communauté phonéticienne ait, très tôt, fait la distinction entre les objets clairement définis dans les niveaux de l'analyse phonétique (propriétés, indices et traits) et les objets issus du traitement automatique du signal révèlant une distinction linguistique pertinente [Abry-Boë 81]. Pourtant la distinction entre corrélats et indices en reconnaissance automatique est confuse car souvent la quantification de ces corrélats en divers niveaux ou formes permettent de définir des traits distinctifs à l'intérieur d'un système rendu cohérent par l'emploi d'algorithmes déterministes. Ces corrélats [Caelen- Caelen-Haumont 81, Rossi 83...] grâce à l'ajustement de seuils et de règles de décision ah-hoc permettent de simuler le comportement des indices sous-jacents. C'est donc en fait une expertise humaine guidée par une batterie d'algorithmes de traitement de signal, de classification automatique et utilisant une base de connaissances qui permettra d'ajuster les paramètres du modèle intuitif aux faits de parole et aux contraintes du système linguistique [Caelen-Vigouroux 85].

Une autre approche d'appariement faits- connaissances est proposée ici et évaluée grossièrement sur une application classique: la détection de formants. La correspondance entre deux espaces l'un acoustique, l'autre indiciel est vue dans cette approche, comme le résultat du décodage d'une séquence d'observations acoustiques par un processus markovien qui se charge de reproduire d'une manière cohérente l'analyse faite par l'expert lors d'une séance d'apprentissage.

B- CADRE GENERAL DE L'ETUDE

Cette étude reprend la démarche générale suggérée par [Baker 75, Ney 83] et plus récemment par [Kopec 86] pour la détection automatique de formants et par [Atal & Al 78] pour modéliser un système d'inversion d'un modèle articulatoire. Dans ces quatre études, l'utilisation de la programmation linéaire, dynamique ou du chaînage markovien avait pour but de ne plus définir les correspondances entre deux espaces par une simple algorithmique et donc par un modèle dont les lois sont explicites mais par un modèle stochastique dont le fonctionnement est implicite.

L'outil que nous avons construit est largment inspiré de ces ouvrages et plus particulièrement de [Kopec 86]. Il présente l'avantage de pouvoir à tout moment réactualiser les connaissances après correction par l'expert des décisions incohérentes grâce à une quantification vectorielle à seuil.

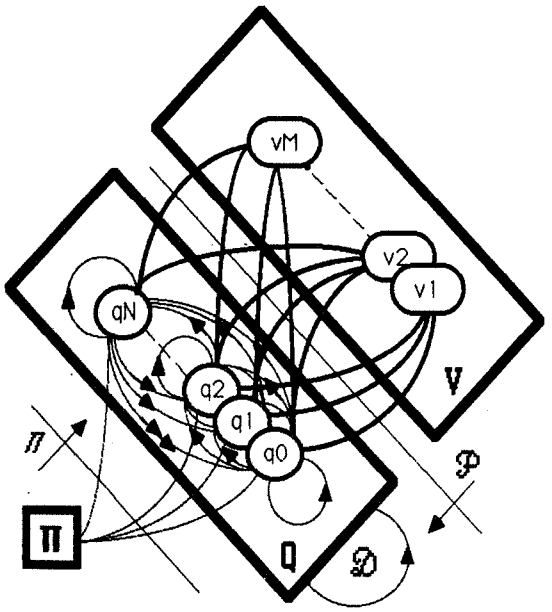


Fig 1: Description du modèle de congruence entre deux espaces V et Q

Le shéma de modèle de congruence entre deux espaces quantifiés est présenté Fig.1. On peut remarquer que ce shéma est symétrique à cela près que chaque quantification scalaire privilégie un état noté q_0, qui assume le rôle de modèle d'incertitude maximale ou d'absence d'indice (cf SC). Ainsi l'espace d'entrée est indifférent et l'on pourrait envisager aisément de reconnaître des configurations acoustiques à partir d'indices (synthèse à partir d'indices) ou passer d'un espace acoustique à un autre (application au codage).

B-1. formalisme

Un Modèle de Markov Caché (MMC) est un 5-uplet (Q,V,Π,D,P) qui consiste en un ensemble d'états Q (q_i / i=0..N) et un ensemble de symboles V (v_j / j=1..M), un vecteur de probabilités à-priori Π (Π_i/ i=0..N), une matrice de probabilités de transition D (d_{ij} / i=0..N, j=0..N) et une matrice de probabilités d'observation P (p_{ij} / i=0..N, j=1..M).

$$\Pi = (\Pi_i) == (Pr(q_i))$$
$$D = (d_{ij}) == (Pr(q_j(t)/q_i(t-1))$$
$$P = (p_{ij}) == (Pr(q_j(t)/v_i(t))$$

La matrice D sera appelée par la suite *matrice dynamique* car elle modélise la dynamique de l'espace Q. Dans le même ordre d'idée, la matrice P sera appelée *matrice de passage* de l'espace V à l'espace Q.

On voit donc que les contraintes qui guideront le suivi dynamique du décodage de la chaîne markovienne sont sur l'espace de sortie. Cette méthode est donc particulièrement adaptée au suivi de corrélats acoustiques tels que formants, paramètres de modèles articulatoires où les contraintes de continuité sont très fortes.

B-2. quantifications

L'utilisation de ce MMC suppose que les deux espaces soient quantifiés préalablement: une quantification scalaire doit s'appliquer aux indices et une quantification vectorielle aux formes acoustiques.

B-2.1. quantification scalaire

La bande de variation (V_m, V_M) d'un indice est divisée à priori en N niveaux suivant une fonction de quantification f adaptée: linéaire, logarithmique... La valeur de l'indice $val(q_i)$ associé à un symbole q_i sera donc:

$$val(q_i) = f\left(\frac{f^{-1}(b)+f^{-1}(h)}{2}\right), b=(2i-1).\left(\frac{V_M-V_m}{2N+2}\right), h=(2i+1).\left(\frac{V_M-V_m}{2N+2}\right)$$

B-2.2. quantification vectorielle

Les algorithmes classiques de constitution de dictionnaire de quantification vectorielle [Linde & Al 80, Wong & Al 82] consistant par des procédures itératives tendant à converger vers une distribution des représentants minimisant localement une distorsion globale, présentent deux inconvénients majeurs: le coût en temps de calcul et le nécessité de reconstruire complètement le dictionnaire si l'on augmente la séquence d'apprentissage. Nous utiliserons donc une quantification vectorielle en deux temps: une première *quantification* dite *à seuil* [Tou & Al 74, Dabouz & Miclet 83] lors de la phase d'apprentissage. Puis on pourra réorganiser ce dictionnaire grâce à un algorithme de classification afin d'obtenir un nombre de classes inférieur et fixé à priori.

L'avantage de la quantification à seuil est d'être *en ligne*, c'est-à-dire que les vecteurs d'apprentissage sont traités dans l'ordre où ils apparaissent et une seule fois. On pourra donc facilement constituer dictionnaire de formes spectrales et MMC.

B-3. apprentissage

L'apprentissage s'effectue *en ligne* (voir plus haut): dictionnaires de formes spectrales et modèles de Markov sont constitués en même temps suivant le shéma fig.1.a. Les paramètres des modèles markoviens sont estimés à partir de la fréquence des événements dans la phase d'apprentissage. Les matrices Π,P,D sont en fait des histogrammes d'observation et les probablités correspondantes seront calculées selon les formules:

$$p_{ij}=\frac{n^P_{ij}+\varepsilon^P_j}{\sum_{m=0}^{N}\left(n^P_{mj}+\varepsilon^P_j\right)}, \quad n_{ij}=\frac{n^D_{ij}+\varepsilon^D_j}{\sum_{m=0}^{N}\left(n^D_{mj}+\varepsilon^D_j\right)}, \quad \pi_i=\frac{n^{\pi}_i+\varepsilon^{\pi}}{\sum_{m=0}^{N}\left(n^{\pi}_m+\varepsilon^{\pi}\right)}$$

où n^P_{ij} est le nombre de trames dans l'état j de Q sachant qu'elle est dans l'état i de V. n^D_{ij} est le nombre de trames dans l'état j de Q sachant que la trame précédente était dans l'état i de Q. n^{π}_i est le nombre de trames dans l'état i de Q.

Les biais notés $\varepsilon^P,\varepsilon^D,\varepsilon^{\pi}$ sont destinés à assurer un dénominateur non nul à ces formules et sont calculés à partir de la somme des nombres n_* correspondants par la formule: $\varepsilon = 10^{-6}.\Sigma n_*$.

B-4. reconnaissance. décodage de la séquence d'observations

Le décodage de la séquence de symboles $\mathcal{V}(\mathcal{V}(t)/t=0..T)$ appartenant à l'espace V peut s'effectuer grâce aux algorithmes classiques de décodage de MMC [Bahl & Al 83]. L'algorithme de Viterbi présente le désavantage de ne déterminer que le meilleur chemin de parcours du MMC soit la séquence de symboles $\mathcal{Q}$ $(\mathcal{Q}(t)/t=0..T)$ de l'espace Q: la solution respecte donc les pas de quantification de l'espace Q et la procédure de décision est définie implicitement dans l'algorithme [Kopec 85]. L'algorithme de Forward-Backward présente quant à lui l'avantage de donner un estimé de la distribution de probabilité de présence de chaque trame dans l'espace Q. La séquence décodée $\mathcal{Q}$ sera alors un barycentre des symbôles de Q affectés des probablités de présence.

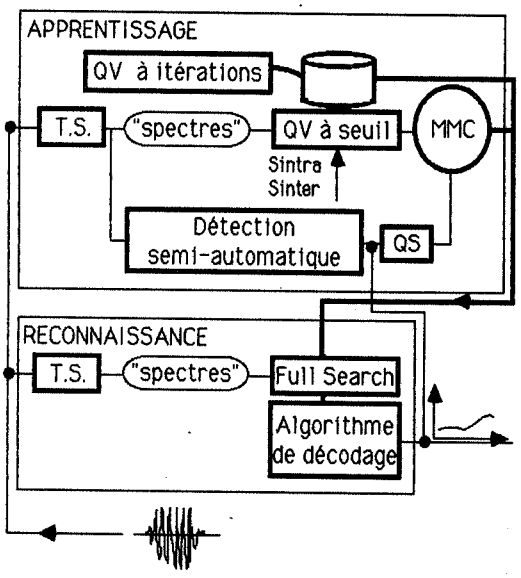


Fig.1.a.: Diagramme de fonctionnement des phases d'apprentissage et de reconnaissance. On voit apparaître la possibilité de renvoi de la sortie sur l'entrée.

L'algorithme de Forward-Backward permet de calculer l'ensemble des probabilités de présence à l'état q_i à la trame t sachant que l'on a la séquence de décodage $\mathcal{V}$ dans l'espace V par la formule:

$$Pr(\mathcal{Q}(t)=q_i/\mathcal{V}) = Pr(\mathcal{Q}(t)=q_i \,\&\, \mathcal{V}) \,/\, Pr(\mathcal{V})$$

avec $Pr(\mathcal{Q}(t)=q_i\&\mathcal{V}) = \alpha_i(\mathcal{V},t).\beta_i(\mathcal{V},t)$

$$\alpha_i(\mathcal{V},t)=\sum_{j=0}^{N}\alpha_j(\mathcal{V},t-1)).d_{ji}.p_{\mathcal{V}(t)i}$$

$$\beta_i(\mathcal{V},t)=\sum_{j=0}^{N}\beta_j(\mathcal{V},t+1)).d_{ij}.p_{\mathcal{V}(t)i}$$

$$Pr(\mathcal{V})=\sum_{i=0}^{N}\alpha_i(T)$$

Les probabilités α et β sont respectivement les probabilités *Forward* et *Backward*. Les conditions aux limites sont: $\alpha_i(\mathcal{V},0) == \pi_i$ et $\beta_i(\mathcal{V},T+1) == 1/(N+1)$.

La distribution de probabilité des symboles de l'espace Q au temps t étant ainsi calculée, on calcule la probabilité $\delta(t)$ d'avoir un symbole de sortie différent de q_0 par la formule:

$$\delta(t) = \sum_{i=1}^{N} Pr(\mathcal{Q}(t)=q_i/\mathcal{V})$$

Si aucun symbole q_0 n'a été détecté dans la phase d'apprentissage, la valeur de δ_{min} est sans effet et l'algorithme détectera des valeurs $val(\mathcal{Q}(t))$ en continu. Cependant cet état vide q_0 permet à l'expert d'avouer son indécision lors de la détection des indices sur le corpus d'apprentissage. La perte de visibilité d'un indice sera alors retraduite par la même émisson de q_0 lors du décodage.

Un symbole différent de q_0 sera donc alors émis à l'instant t si $\delta(t) > \delta_{min}$. La valeur de $\mathcal{Q}(t)$ sera alors donnée par un simple calcul de barycentre des valeurs des symboles de Q différents de q_0 affectés des probabilités $Pr(\mathcal{Q}(t)=q_i/\mathcal{V})$ suivant l'expression :

$$val(\mathcal{Q}(t)) = \frac{\sum_{i=1}^{N} val(q_i).Pr(\mathcal{Q}(t)=q_i/\mathcal{V})}{\sum_{i=1}^{N} Pr(\mathcal{Q}(t)=q_i/\mathcal{V})}$$

C- APPLICATION A LA DETECTION DES FORMANTS

C-1. Motivations

L'application de cette modélisation à la détection de formants permet de comparer quantitativement l'outil ainsi créé à d'autres méthodes classiques exploitant des techniques de prétraitement comme l'analyse LPC, le cepstre ou le zero-crossing [McCandless 74, Olive 71, Niederjohn 75]. Elle permet de suivre facilement l'évolution de la performance du modèle par apprentissage progressif.

C-2. Corpus

Les phrases utilisées pour l'apprentissage et le test des modèles markoviens sont extraites de la base de données des sons du français BDSON comprenant pour l'instant les réalisations d'une dizaine de locuteurs sur divers corpus. L'évaluation présentée ici porte sur les réalisations d'une locutrice du corpus de phrases phonétiquement équilibrées (PEQ01..5). Ce corpus comporte 50 phrases. Les signaux originaux de BDSON ont été échantillonnés à 16 KHz sur une dynamique de 16 bits. Ces sons sont ensuite sous-échantillonnés numériquement à 10 KHz en réduisant la dynamique à 12 bits.

L'espace vectoriel de départ C est constitué de la quantification vectorielle de trames de vecteurs cepstraux obtenus toutes les 5 ms. 10 coefficients cepstraux (c_i/i=1..10) sont obtenus [Atal 74] avec une fenêtre de Hamming de 25.6 ms et 16 pôles. La distance cepstrale utilisée est la suivante:

$$d(C_1,C_2) = \sum_{i=1}^{10} i.(C_{1i}-C_{2i})^2$$

Le seuil de quantification inter-classe est de 1.1 et intra-classe de 0.3. Dans le cas de comparaison entre 2 vecteurs cepstraux de distance inférieure au seuil intra-classe, on ne conserve que celui associé à l'erreur de prédiction la plus faible (spectre le plus résonant) [Miclet-Dabouz 85]. Aprés la quantification à seuil, le nombre de représentant à été réduit à 512 par une quantification itérative [Linde & Al 80].

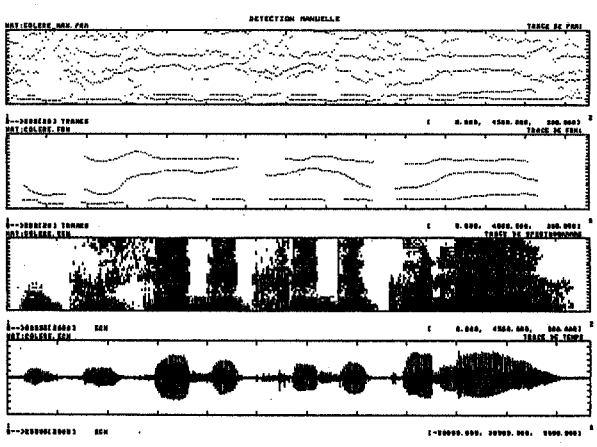


Fig.2. Exemple de détection manuelle des 3 premiers formants sur la phrase: "vous poussez des cris de colère". Les 4 cadres figurent respectivement de bas en haut: le signal, son spectrogramme large bande, les 3 formants détectés et les maxima du spectre LPC à 16 coefficients sur la bande [0,4500Hz].

Les espaces indiciels de sortie sont les espaces de quantification scalaire des trois premiers formants notés F1,F2,F3. Les bandes de quantification sont choisies respectivement égales à [50-1500], [500-3500], [1500-4000] pour notre locutrice et sont divisée en 20 bandes sur une échelle linéaire.

C-3. Méthodologie d'étiquetage de la base.

Les 3 premiers formants des phrases du corpus PEQ ont été détectés manuellement grâce à un logiciel interactif de traitement de signal PTS développé à l'ICP. Les représentations de référence étaient constituées d'un spectrogramme large-bande et du tracé des divers pics du spectre lissé par une LPC à 16 coefficients. Les trajectoires des 3 premiers formants ont été acquis alors à l'aide d'une table à digitaliser.

Les critères d'étiquetage prépondérants ont été des critères de continuités et de proéminence. Cependant lors du croisement de 2 formants les modèles formantiques n'ont pas été échangés et les courbes présentent alors des points de rebroussement. Un module d'interprétation sera ajouté par la suite afin d'associer maxima du spectre et fréquences de résonance de cavités. Pour les consonnes et voyelles nasales, la résonance naso-pharyngale autour de 1000Hz n'a pas été étiquetée. De même le premier formant n'a pas été détecté lorsqu'il se confond avec le formant glottique (cas des occlusives et fricatives voisées).

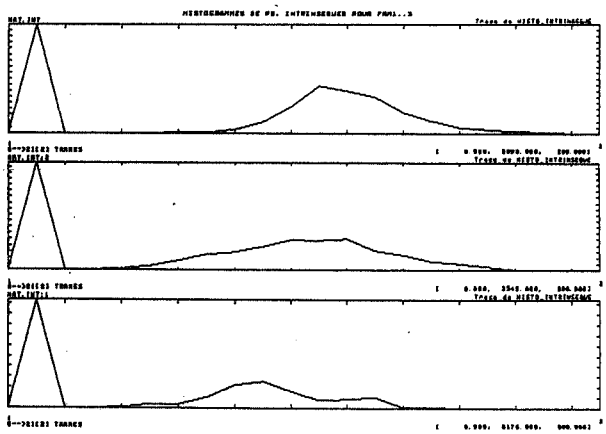


Fig.3. Histogrammes intrinsèques $\pi(F1..3)$.

Un exemple de détection manuelle est présentée fig.2. La phrase est extraite du corpus PEQ01. Au total, environ 27000 trames ont été étiquettées. La table Tab.1. présente la répartition des configurations formantiques pour les 20 phrases de [PEQ01,PEQ02].

Formants	Occurences	
1 2 3	abs	%
- - -	3138	30.4
- - x	56	0.5
- x -	198	1.9
- x x	93	0.9
x - -	219	2.1
x - x	70	0.7
x x -	1528	14.8
x x x	5010	48.6
	10312	

Tab.1: répartition des configurations formantiques sur l'ensemble des corpus [PEQ01,02]. (le tiret marque l'absence, la croix la présence)

La figure Fig.3 présente les histogrammes π des probabilités intrinsèques de présence pour les espaces F1,F2 et F3. On constate que le manque pour l'instant de certaines configurations formantiques à l'apprentissage du MMC conduit à une allure presque bimodale de $\pi(F1)$. Ce qui conduit à des transitions relativement abruptes sur les courbes de F1 détectés, le modèle passant d'un mode à l'autre par un phénomène de capture.

C-4. Résultats

La figure Fig.4 présente l'évolution de la détection automatique des formants de la même phrase cité plus haut suivant le nombre de trames apprises et pour un seuil δ_{min} de 0.5. Dans une première étape, les 10 phrases du corpus PEQ02 ont été utilisées pour l'apprentissage, puis les 10 phrases du corpus PEQ03 ont été ajoutées. Enfin la phrase elle-même a été introduite dans le corpus d'apprentissage. On remarque ainsi que le modèle à taille de l'espace V constant améliore sa détection au fur et à mesure de l'enrichissement de son apprentissage.

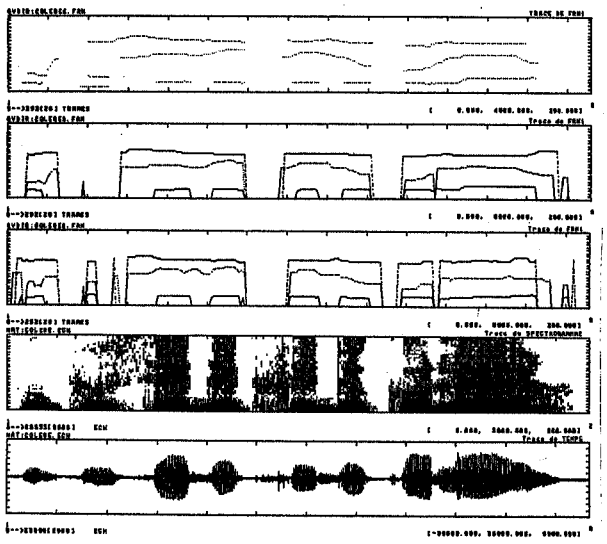


Fig.4. Exemple de détection automatique des 3 premiers formants sur la phrase: "vous poussez des cris de colère". Les 5 cadres figurent respectivement de bas en haut: le signal, son spectrogramme large bande, les 3 formants détectés sur la bande [0,4500Hz] successivement après l'apprentissage de PEQ02, puis PEQ03 et enfin après l'ajout de la phrase elle-même.

Afin de donner un apercu de la qualité de la détection, un seuil d'erreur grossière a été considéré égal a +/-20% de la valeur du formant attendue. D'autre part, ces erreurs regroupant les erreurs grossières, les formants manqués et les fausse alarmes, on a déterminé la répartition des erreurs suivant ces trois cas. Pour un apprentissage comportant 30 phrases PEQ01..3 et utilisant les 20 phrases PEQ04..5 comme corpus de test, les statistiques ont dégagé les valeurs suivantes pour un seuil δ_{min} égal à 0.5:

Type d'erreur (%)	Formants 1	2	3
Erreurs larges	2.5	2.1	1.5
DONT			
Fausses alarmes:	34	10	45
Formants manqués:	21	55	15
Erreurs:	45	35	40

C-5. Discussion

L'application de cette approche de modélisation du lien de congruence entre deux espaces phonétiques à la détection de formants n'a été entrepris que pour évaluer la fiabilité et les performances d'un outil sur une application connue et déjà étudiée. Il reste à mener une étude plus systématique sur le choix des paramètres de l'espace V de cette application et de la métrique associée qui devra privilégier en ce cas la structure formantique. Les premiers résultats obtenus sur un corpus de BDSON pourra nous l'espérons, provoquer une comparaison systématique des algorithmes de détection de formants comme ce fût le cas des algorithmes de détection de FO.

L'avantage de cette méthode est de pouvoir prédire des trajectoires indicielles même en l'absence d'une congruence marquée entre les deux

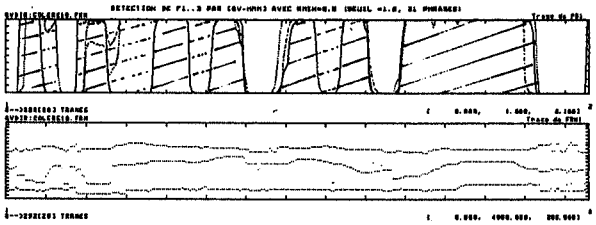


Fig.5. Détection des trois premiers formants avec un seuil δ_{min} égal à 0. Le cadre du haut donne l'allure des courbes $\delta(F_*)$ faisant apparaître les ilots de confiance.

espaces. Grâce à ses propres contraintes dynamiques, l'espace Q peut spontanément générer les indices appelés précédemment "fausses alarmes", qui présentent néanmoins souvent une nette cohérence (cf. Fig. 5) avec les ilôts de confiance dégagés par des segments possédant un δ très proche de 1. Cette propriété peut être utilisée avec à-propos dans un système d'étiquetage automatique pour synthèse à formants.

CONCLUSIONS

L'adéquation des modèles markoviens à suivre l'évolution dynamique de deux systèmes congruents a été suggérée. Le fonctionnement de ce système sur une analyse acoustique classique de signal de parole a été évaluée sur la base de donnée BDSON. Les résultats sont très encourageants et valident cette démarche. Il reste à ajouter à cet outil un module d'interprétation permettant de réévaluer les formants ainsi détectés dans le cadre d'un modèle de production.

La polémique essentielle de ce type de démarche globale reste le point de couplage entre l'analyse purement ascendante des faits et les connaissances purement descendantes dans un système de reconnaissance. Les systèmes de corrélats-indices se sont basés jusqu'à présent sur des traitements acoustiques et perceptifs du signal de parole. Cette option révèle d'une part le manque de connaissances sur le système de production, l'inadéquation des modèles de géométrie du conduit vocal à décrire la dynamique articulatoire et enfin le manque d'outils de modélisation des congruences articulatoires-acoustiques. L'outil présenté ici pourrait être un moyen de mettre en relation données physiologiques (et/ou psycho-acoustiques) et formes acoustiques.

BIBLIOGRAPHIE

[Abry,Boë 81] "Unités et niveaux de traitement du signal de parole: apports d'une analyse linguistique infrastratique", Processus d'encodage et de décodage phonétiques, GALF, Toulouse, 30-38.

[Atal 74] "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", Journ. of Acoust. Soc. of Am., 55, 6, 1304-1312.

[Atal, Chang, Mathews,Tukey 78] "Inversion of articulatory-to -acoustic transformation in the vocal tract by a computer-sorting technique", Journ. of Acoust. Soc. of Am., 63, 5,1535-1555.

[Bahl, Jelinek,Mercer 83] "A maximum likehood approach to continous speech recognition",IEEE Trans. on PAMI, 5.

[Baker 75] "Stochastic modelling for automatic speech understanding", Speech Recognition, Reddy, New-York Academic.

[Caelen, Caelen-Haumont 81] "Indices et propriétés dans le projet ARIAL II", Proceedings GALF-CNRS, Processus d'encodage et de décodage phonétique.

[Caelen, Vigouroux 85] "Une base acoustique et phonétique hiérarchisée: des faits aux connaissances", Actes du symposium Franco-suédois, Grenoble.

[Kopec 86] "Formant tracking using Hidden Markov Models and Vector Quantization", IEEE Trans. on Acoust., Speech & Sig. Proc., 34,4, 709-729.

[Linde, Buzo, Gray 80] "An algorithm for vector quantizer design", IEEE Trans. on Comm. Techn., 2,.

[McCandless 74] "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. on Acoust. Speech & Sig. Proc,2,135-141.

[Miclet,Dabouz 85] "Un vocodeur à classification", Séminaire GALF-GRECO, Paris, 53-90.

[Ney 83] "Dynamic programming algorithm for optimal estimation of speech paranmeter contours", IEEE Trans. on Syst., Man, Cybern., 13.

[Niederjohn, Lahat 85] "A zero-crossing consistency method for formant tracking of voiced speech in high noise levels", 33,2., 349-355.

[Olive 71] "Automatic formant tracking by a Newton-Raphson technique", J.A.S.A., 661-670.

[Rossi 83] "Indices acoustiques multilocuteurs et inédpendants du contexte pour la reconnaissance automatique de la parole", Speech Comm., 2, 215-217.

[Tou, Gonzalez 74] "Pattern recognition principles", Addison-Wesley, 90-92.

EXPERIENCES EN VUE DU DECODAGE ACOUSTICO PHONETIQUE A PARTIR D'UNE RECHERCHE STATISTIQUE D'EVENEMENTS ARTICULATOIRES ET D'UN CODAGE VECTORIEL

Régine André-Obrecht et Huan Yu Su

Campus de Beaulieu
35042 Rennes Cedex - France

ABSTRACT

We present here a new approach of the acoustic-phonetic decoding. The main idea is to use statistical methods to study the speech signal sample-by-sample and to extract acoustic units which may be a priori independent of phonetic units.

Then, the problem is to characterize the so obtained units and to propose an automatic recognition based on phonetic rules and vector quantization.

INTRODUCTION

Dans un processus de reconnaissance de parole continue, la reconnaissance doit faire appel au décodage acoustico-phonétique. La transcription d'un signal continu en une suite discrète d'éléments appartenant à un vocabulaire fini (par exemple, l'ensemble des phonèmes), est un problème difficile et crucial puisque les performances des niveaux supérieurs de reconnaissance et du système lui-même dépendent essentiellement de la qualité de ce codage.

Nous avons essayé de reprendre le problème de décodage acoustico-phonétique à partir du signal numérique lui-même et non des paramétrisations communément utilisées en parole (par exemple LPCC), afin d'en extraire par des techniques propres au traitement signal (E.E.G., signaux géophysiques...), le maximum d'informations. Cette étude statistique du signal n'a pour seule prétention que de signaler des événements effectivement présents au niveau acoustique, et d'en proposer une caractérisation a posteriori.

Dans ce papier nous présentons une segmentation automatique statistique, la description des unités ainsi obtenues (segments événementiels, transitoires et stables) et faisons quelques propositions en vue de leur classification.

II - PRETRAITEMENT : SEGMENTATION STATISTIQUE DU SIGNAL

Cette approche a pour but de détecter séquentiellement des changements dans les caractéristiques spectrales du signal numérique. La méthode consiste donc à modéliser le signal et à utiliser un test statistique pour affirmer ou non la présence d'une rupture. Plusieurs méthodes ont été expérimentées [2], seule, celle donnant les meilleurs résultats sera décrite ici.

II.1 - Modélisation et statistique : la méthode de divergence

Le signal est supposé décrit par une suite d'unités homogènes, chacune étant caractérisée par un modèle autoregressif (LPC); il s'agit de détecter séquentiellement un changement dans le vecteur $\theta^T = (a_1, \ldots, a_p, \sigma)$. La statistique utilisée est une mesure de distance entre deux modèles θ_0 et θ_1 (figure 1)

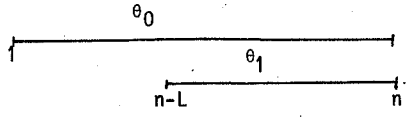


Figure 1: Position relative des deux modèles dans la recherche d'une rupture.

Cette distance dérive de l'entropie mutuelle entre lois conditionnelles [3], et elle est donnée dans le cas gaussien par :

$$W_n = \sum_{k=1}^{n} w_k \qquad (1)$$

avec

$$w_k = \frac{1}{2}\left\{ 2\,\frac{e_{0,k}\,e_{1,k}}{\sigma_1^2} - \left(1 + \frac{\sigma_0^2}{\sigma_1^2}\right)\frac{e_{0,k}^2}{e_{1,k}^2} + \left(1 - \frac{\sigma_0^2}{\sigma_1^2}\right)\right\}$$

où $e_{0,k}$ et $e_{1,k}$ sont les erreurs résiduelles correspondant à chaque modèle $\theta_0^T = (a_{01}, \ldots, a_{0p}, \sigma_0)$ et $\theta_1^T = (a_{11}, \ldots, a_{1p}, \sigma_1)$.

II.2 - Expérience

Le modèle θ_0, dit modèle long terme, est identifié séquentiellement par la méthode de Burg utilisée sur une fenêtre croissante, tandis que le modèle court-terme θ_1 est identifié sur une fenêtre glissante parla méthode d'autocorrélation. Une fenêtre de longueur L est nécessaire à l'initialisation (figure 1).

Une modification de la statistique :

$$\tilde{W}_n = \sum_{k=1}^{n} (w_k - \delta)$$

où δ est un biais positif fixé, conduit à déceler une rupture de modèles dès que :

$$\max_{1 \leq m \leq n} \tilde{W}_m - \tilde{W}_n > \lambda \qquad (\lambda \text{ seuil})$$

et l'instant de rupture r correspond à l'argument de ce maximum (figure 2)

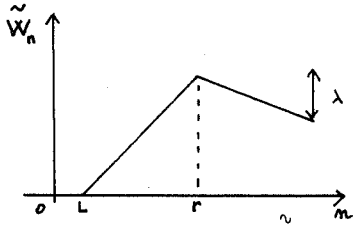


Figure 2: Allure de la statistique $\tilde{W}_n$ en présence d'une rupture à l'instant r

A la suite d'expériences sur une phrase extraite d'une liste de phrases phonétiquement équilibrées, la longueur du modèle court terme, L, a été fixée à 256 (fréquence d'échantillonnage 12.8 KHz), et l'ordre du modèle à 16.

Les paramètres les plus difficiles à règler sont le biais δ et le seuil λ: les premiers résultats ont montré que le comportement de la statistique est plus ou moins bruité suivant que le signal est voisé ou non ; à l'issue d'une décision grossière "voisé-non voisé" prise à partir de l'énergie, le test est calculé avec :

(δ_v, λ_v) = (0.2, 40) pour les zones voisées

(δ_b, λ_b) = (0.8, 80) sinon.

Il s'est également confirmé que cette statistique n'est pas symétrique : est détecté le passage /m e/ alors que la transition /ə m/ ne l'est pas. Cette remarque est à l'origine de la dernière amélioration de ce test.

II.3 - La méthode de divergence forward-backward

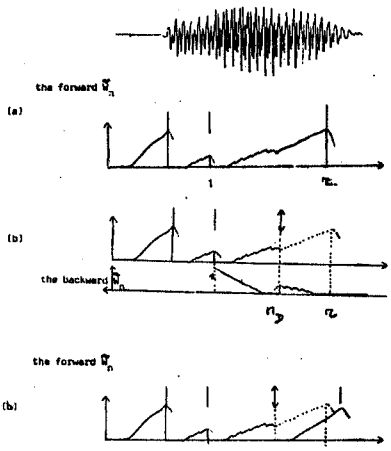


Figure 3: Principe de la méthode de divergence forward-backward

Dans cette nouvelle procédure, illustrée par la figure 3, nous introduisons Lmin, la longueur minimale permise pour deux phonèmes voisés juxtaposés. Lorsqu'un nouveau segment voisé est trouvé, de longueur r > Lmin, la méthode de divergence lui est appliquée dans le sens rétrograde. Deux cas se présentent alors :

1. aucune nouvelle détection n'est apparue et la segmentation se poursuit à partir de l'instant r

2. il existe une nouvelle détection à l'instant n_D (figure 3b), le segment $|1,\ldots,n_D|$ est validé et le processus reprend à partir de n_D.

II.4 - Résultats qualitatifs et quantitatifs

Après avoir fixé les différents paramètres, ce test statistique a été expérimenté sur 4 ensembles de données :

- 5 listes de phrases phonétiquement équilibrées, monolocuteur,
- des suites de nombres extraits de l'ensemble {0,1,...,99} prononcées par 10 locuteurs (4 hommes, 6 femmes)
- 50 suites de numéros téléphoniques (8 chiffres) monolocuteur, pour l'évaluation quantitative
- un ensemble de logatomes afin de préciser la nature de certaines frontières.

Les résultats de la segmentation ainsi obtenue ont été étudiés avec l'aide des sonagrammes correspondants. Les frontières correspondent toujours à des changements acoustiques "révélateurs de changements articulatoires"; une rupture peut être :

- un changement "brusque" comme le début ou la fin d'un voisement ou d'une friction (|1|)
- ou le début d'un changement spectral progressif.

La caractérisation des frontières nous a amené à classer les segments en trois catégories :

- des segments homogènes ou quasi-stationnaires; ils correspondent à la partie stable des phonèmes, lorsqu'elle existe,
- des segments transitoires lorsque se produit une variation spectrale lente
- des segments courts "événementiels" définis uniquement par leurs frontières (closion - début de silence, explosion - début de voisement) (figure 4).

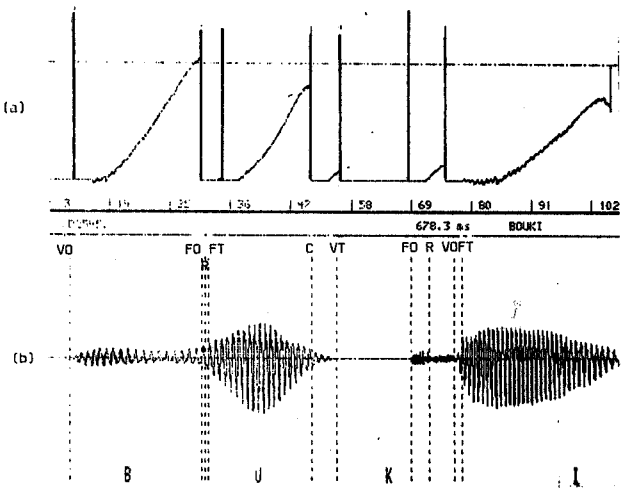


Figure 4: Comparaison entre segmentation manuelle et automatique sur le logatome /buki/

Quantitativement, nous avons trouvé 3379 segments pour 1534 phonèmes, soit 2,20 segments par phonème, ce qui correspond grossièrement à la succession d'un segment transitoire par un stable. Si l'on se réfère à la classification précédente, il est apparu les omissions suivantes :

- 12 omissions entre /ə v/ sur 59 /v/ présents
- 20 omissions concernant l'explosion des plosives sur les 281 : il n'y a pas eu isolement de cette zone surtout dans les contextes /t s/ et /t r/...

Signalons également 11 mauvaises localisations (retard à la détection).

III - ETIQUETAGE DES FRONTIERES RECHERCHE DES SEGMENTS TRANSITOIRES

Guidé par les résultats du test précédent, nous avons recherché à effectuer automatiquement la classification proposée.

III.1 - Etiquetage des frontières

Le but de cette étude est l'étiquetage des frontières correspondant à certains événements articulatoires |1| tels que

VO-VT : début et fin de voisement
FO-FT : début et fin de friction
C : closion
B : fin d'occlusion ou explosion d'une plosive

Pour réaliser cette phase ont été retenus un test de détection d'explosion de plosives, et le test de segmentation "haute-fréquence" :

- la méthode de divergence (sens direct seulement) montre que l'on obtient sur le signal filtré haute fréquence (> 4000 Hz) des segments pertinents relativement à la notion de bruit ; les segments obtenus sont longs et lorsqu'il s'agit d'une zone bruitée, les frontières correspondent exactement à l'apparition et à la fin d'une friction
- l'instant d'explosion des plosives est fourni par un test classique relatif à la détection de changements brusques d'une variance |3|:

> à l'instant n, un modèle autorégressif gaussien (LPC) $\theta = (a_1,\ldots,a_p,\sigma)$ est identifié sur la fenêtre |n-L,n| et la statistique est donnée par
>
> $$T_n = \sum_{m=n+1}^{n+\tau} \frac{(y_m - \sum_i a_i\, y_{m-i})^2}{\sigma^2} \qquad (\tau=16)$$

Ce test est activé en parallèle avec celui détectant un accroissement d'énergie Δe_n en haute fréquence

> $$\Delta_n = \alpha\, \Delta_{n-1} + \beta\, \Delta_{n-2} + \Delta e_n$$

et celui détectant les zones de faible énergie dans la zone formantique 600 < < 4000 Hz. La complémentarité de ces décisions permet d'obtenir les silences ou les barres de voisement couplés à l'instant d'explosion de la plosive.

L'étiquetage se déroule alors de la manière suivante : le test "voisé-non voisé-silence" est activé sur chaque segment fourni par le prétraitement; seules les étiquettes VO, VT, SO, ST (début et fin de silence) sont fixées puisqu'un segment de type "FO-VT" ou "VO-FT" est toujours classé voisé. Les étiquettes FO, FT sont posées après passage du test de voisement sur les segments obtenus à partir du signal filtré. Une coordination temporelle réalisée à partir de règles phonétiques, des résultats précédents, et des tests liés aux plosives, permet d'obtenir l'étiquetage final en corrigeant et ajoutant les événements C et B. Après cette analyse, sont classés les segments événementiels du type :

VO-FT, FO-VT, SO-ST, C-VT, B-FT ...

Nous avons cependant remarqué (dans les corpus des numéros téléphoniques) des erreurs : il arrive qu'un /v/ soit assimilé à une plosive voisée; la lecture des sonagrammes confirment que ces confusions sont inévitables à ce niveau.

III.2 – Recherche des segments transitoires

Cet étiquetage ne concerne, en l'état actuel du travail que les segments de durée inférieure à 40ms. Il s'agit de décider de leur caractère transitoire ou non.

La décision peut êtreprise immédiatement après examen de la statistique de divergence : si la valeur moyenne de la statistique sur le segment est négative le segment est classé comme transitoire.

Sinon est activé le test d'hypothèses suivant:

si le segment est modélisé par un modèle autoregressif gaussien

H_0 : les paramètres du modèle sont indépendants du temps (le segment est stable)

H_1 : les paramètres du modèle dépendent du temps (le segment est transitoire)

Le dépendance des paramètres en fonction du temps est réalisée en supposant que les 2 premiers pôles sont des fonctions linéaires du temps; la dérivation du modèle autorégressif correspondant montre que les coefficients (a_i) sont des fonctions développables en série de Fourier; l'identification de ce modèle repose sur les études faites par Yves Grenier (ENST) |4|. Une fois les deux modèles identifiés, le critère d'Akaike faisant intervenir la vraisemblance et le nombre de paramètres, np_i, de chaque modèle fournit la décision recherchée :

$$AIC = L \log \sigma_1^2 + 2\, np_1 - L \log \sigma_0^2 + 2\, np_0$$

Un résultat de cet étiquetage est donné par la figure 5.

IV – RECONNAISSANCE STATIQUE DES PARTIES DITES STABLES

La reconnaissance des unités stables est faite en utilisant une classification vectorielle. Une étude sérieuse a été réalisée |5| afin de définir quel algorithme de quantification vectorielle était le plus approprié à notre problème. En effet la qualité d'un tel système repose essentiellement sur le choix de la paramétrisation vectorielle et sur l'apprentissage c'est à dire la construction et la qualité du dictionnaire.

IV.1 – Paramétrisation et construction du dictionnaire

L'ensemble final d'apprentissage est composé de trois listes de 10 phases phonétiquement équilibrées, monolocuteur. Les vecteurs sont obtenus à partir d'une segmentation manuelle sur le signal (1175) et sont regroupés en 21 classes phonétiques :

/A, i, j, y, u w, E, ə, O, on, an, in, m, n, (bdgv), l, z, r, s, ʃ, f/.

Pour chaque classe phonétique, nous construisons un dictionnaire de références en utilisant l'algorithme de Lloyd couplé à un splitting dont la perturbation est aléatoire |5|. Après expérimentation, nous avons choisi d'arrêter l'algorithme par un seuil sur la distorsion des classes, quelle que soit la paramétrisation considérée.

Le dictionnaire global est tout simplement la réunion des sous dictionnaires. Remarquons qu'ainsi un même phonème sera représenté plusieurs fois, d'où la différence entre "plus proches voisins" et "plus proches voisins différents" lorsque l'on impose aux "voisins" de représenter ou non des classes phonétiques différentes.

L'évaluation du dictionnaire a été faite en considérant plusieurs sortes de paramétrisations :

- 24 canaux fréquentiels (échelle CNET)
- 24 canaux suivant le modèle d'oreille proposé par le CERFIA
- représentation du spectre lissé dans $\mathbb{R}^{256}$.

Après examen des taux de reconnaissance sur l'ensemble d'apprentissage, la première paramétrisation a été retenue.

Paramétrisation / Règle de la reconnaissance	Canaux		Spectre lissé
	CNET	CERFIA	
Plus proche voisin	93,4	86,1	91,3
3 plus proches voisins	99,2	97,0	98,4
3 plus proches voisins différents	99,5	98,1	99,1

Tableau 1: Taux de reconnaissance sur l'ensemble d'apprentissage (1175 vecteurs, segmentation manuelle fenêtre d'analyse : 40ms).

IV.2 – Reconnaissance

L'ensemble test est formé de 2 listes phonétiquement équilibrées (locuteur identique).

Sont testés les 569 segments, obtenus après élimination des segments événementiels et transitoires détectés. Une fenêtre de 512 ou 256 échantillons est centrée sur le segment stable à reconnaitre suivant sa taille. Le vecteur de paramètres ainsi obtenu est comparé aux éléments du dictionnaire

Règle de la reconnaissance	Taux de reconnaissance (%)
Plus proche voisin	74,5
2 plus proches voisins	86,8
2 plus proches voisins différents	89,8
3 plus proches voisins	92,3
3 plus proches voisins différents	93,5

Tableau 2 : Taux de reconnaissance sur l'ensemble-test (segmentation et pré-traitement automatique, 569 vecteurs, fenêtre d'analyse : 40 ou 20 ms).

Une adaptation du dictionnaire au locuteur en cours de reconnaissance est à l'étude.

CONCLUSION

Nous avons proposé une nouvelle approche du décodage acoustico-phonétique sous forme de recherche de trois types de segments et de trois classifications différentes.

Il reste pour valider cette approche à poursuivre l'étude dans le cadre "indépendance du locuteur" en particulier pour la quantification vectorielle et à proposer un modèle de décodage de la suite ainsi obtenue en une suite de "phonèmes".

REFERENCES

|1| C. Abry : *"Organisation segmentale et temporelle du signal de parole en fonction de sa production"*. Rapport IPG 1984.

|2| R. André-Obrecht : *"Segmentation automatique du signal de parole"*. Thèse de 3ème cycle, Rennes, Mai 1985.

|3| M. Basseville, A. Benveniste : *"Detection of abrupt changes in signals and dynamical systems"*. Lecture "Notes in Control and Information Sciences", Springer-Verlag, 1986.

|4| Y. Grenier : *"Modélisation de signaux non stationnaires"*. Thèse d'Etat, Orsay, Octobre 1984.

|5| H.Y. Su : *"Application de la Quantification Vectorielle à la Reconnaissance des Classes Phonétiques"*. Rapport IRISA, à paraître.

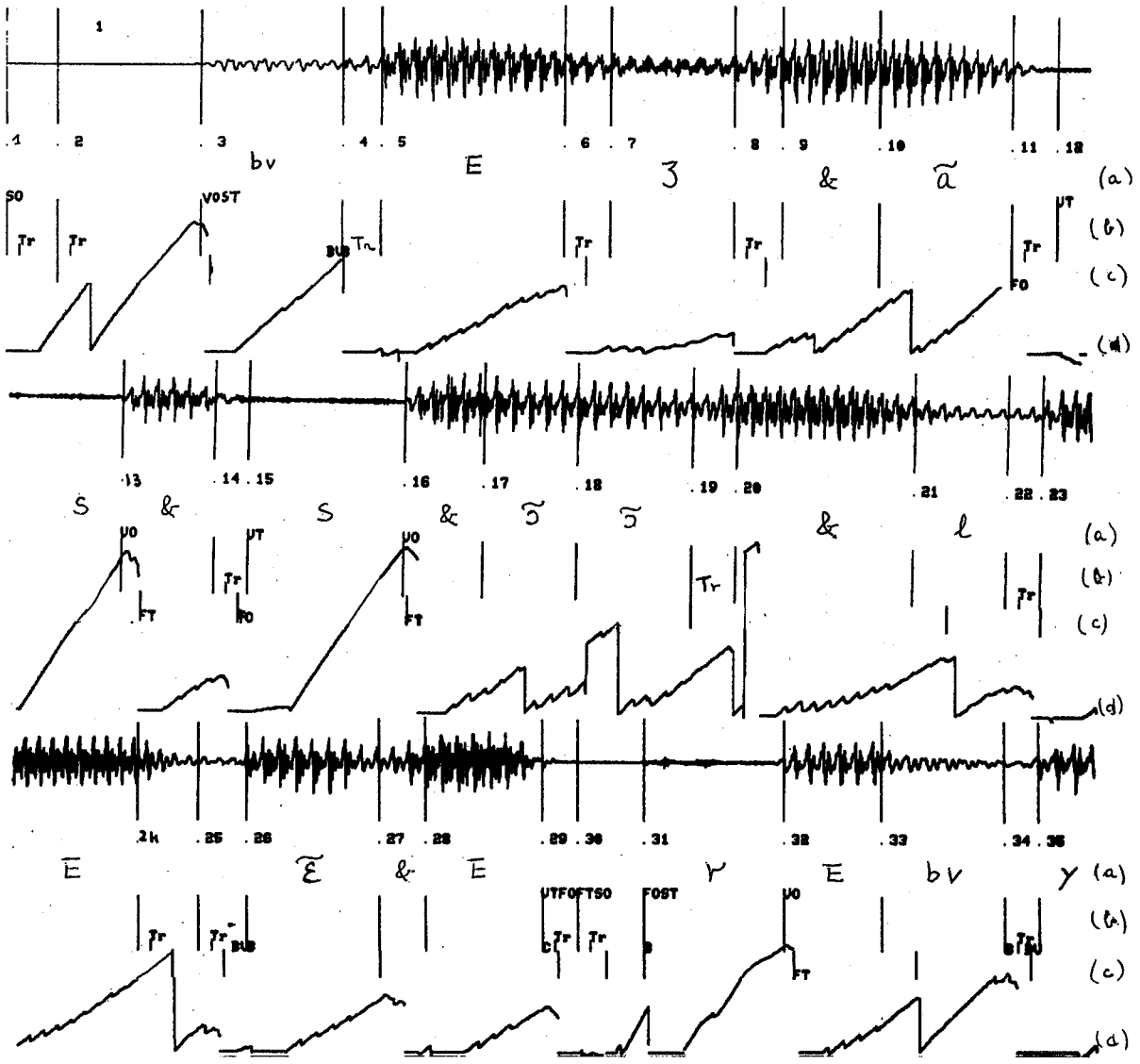


figure 5: Résultat final du décodage sur la phrase " des gens se sont levés dans les tribunes".
(a) resultat de la quantification vectorielle (plus proche voisin)
(b) segments obtenus par la methode de divergence. (c) apres filtrage
Les segments transitoires sont repérés par le symbole Tr.

RECONNAISSANCE DE PARTIES TRANSITOIRES DANS LE SIGNAL DE PAROLE CONTINUE

Dominique VICARD* - Laurent MICLET**

* : ENST Departement SYC 46, rue Barrault 75013 PARIS.
TELIC-ALCATEL 206 route de Colmar 67023 STRASBOURG Cedex.
**: CNET-LAA route de Perros BP 40 22300 LANNION Cedex.

ABSTRACT

This paper describes a recognition system for the transient parts of continuous speech. The algorithm uses DTW as distorsion measure, and applies it starting at each frame of the unknown signal. The location of the transient parts in the speech signal is made using a spotting algorithm. Experiments are made with a specific small transient part dictionary, as well as with a full diphone one. The output of this system is a lattice of hypothesis among which 65% of the transitions are correctly spotted and recognized. A complete system, using a steady part recognizer (previously reported) has been also designed. This work is supported by grants of ANRT and is part of the SPIN/ESPRIT Project.

RESUME

Cet article présente un système de reconnaissance de parties transitoires du signal de parole continue. Le signal inconnu est comparé par un algorithme de DTW au formes de réference, commencant une comparaison à chaque trame. La localisation des parties transitoires est assurée par un algorithme de détection au vol ("spotting"). Des expériences ont été menées avec un dictionnaire spécifique au vocabulaire à reconnaitre, et avec un dictionnaire de diphones. Ce système délivre un treillis d'hypothèses parmi lesquelles 65% des transitions sont correctement localisées et reconnues. Il forme, avec le système de reconnaissance de parties stables précedemment décrit un module complêt pour le décodage acoustico-phonétique. Ce travail est financé par l'ANRT et fait partie du projet SPIN/ESPRIT.

INTRODUCTION

Le décodage Acoustico-Phonétique est une étape clef dans le processus de la reconnaissance de parole continue. Le système décrit dans cet article vise à obtenir un treillis d'hypothèses pseudo-phonétiques le plus riche et le plus "propre" possible, destiné à un étage de traitement "intelligent" pour les vérifications de syntaxe, lexique et de cohérence. Un système de reconnaissance (/VIC.86/) permet d'obtenir les positions et identitées des parties stables du signal de parole. Cependant, une grande partie de l'information utile est contenue dans les parties transitoires.

APPROCHE STATISTIQUE

Deux principales classes de méthodes de reconnaissance sont actuellement utilisées pour le décodage acoustico-phonétique (en dehors des méthodes par règles): les méthodes de type Modèle de Markov Cachè (HMM), où l'on essaie d'identifier un modèle, en évaluant la probabilité que la séquence d'évènements observée en soit produite; les méthodes directes, où une portion du signal inconnu est directement comparé, grâce à une mesure de distorsion, à un dictionnaire de formes. C'est cette deuxième approche, avec pour mesure le DTW et une métrique L1, que nous avons utilisée.

Les raisons de ce choix sont les suivantes:

-apprentissage plus rapide et plus souple (l'inconvénient principal des méthodes de type HMM étant cette phase d'apprentissage des probabilités). Il est possible d'utiliser comme forme de référence des portions de signal, tout comme des formes issues de divers traitements.

-meilleures performances (le DTW reste le type de méthode le plus utilisé dans les systèmes commerciaux).

-simplicité.

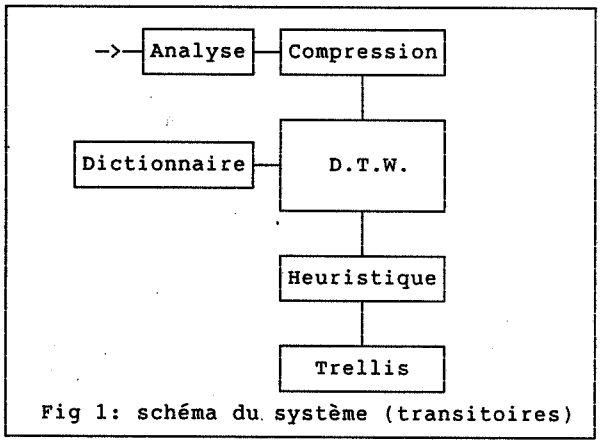


Fig 1: schéma du système (transitoires)

PRETRAITEMENT

Le signal (12 bits/8kHz) est analysé par prédiction linéaire d'ordre 10, puis transformé en 10 coéfficients cepstraux normalisés. Les fenêtres d'analyses sont de 128 points et sont recouvrantes à 50%. Les séquences de vecteurs ainsi obtenues sont

compréssées par un algorithme de type "Frame Selection" (/GAU.83/). Cet algorithme séquenciel élimine du flôt de vecteurs certains vecteurs intermédiaires, assurant ainsi que la distance entre deux vecteurs consécutifs est inférieure à un seuil ε. On obtient ainsi une nouvelle séquence dans laquelle l'échelle du temps a été non-linéairement modifiée.

Un tel prétraitement élimine dans le signal résultant l'importance des parties stables (à l'origine longues, mais dont ne subsistent après traitement que quelques vecteurs représentants). L'algorithme employé est assez rudimentaire, mais présente l'avantage de ne nécessiter aucune information globale (si ce n'est la valeur de la dernière trame retenue) et peut de ce fait être employé en ligne séquentiellement. De plus, sa simplicité permet d'en envisager une implantation matérielle simple.

DICTIONNAIRE

Deux dictionnaires ont été construits et expérimentés. Le premier (DIC0) est constitué de formes représentant toutes les transitions contenues dans une phrase phonétiquement équilibrée (/COM.79/) « Il se garantira du froid avec un bon capuchon ». A ces formes sont ajoutées quelques formes proches (dans l'optique des séries minimales) et qui sont de ce fait facilement confondables. Les formes sont extraites du corpus des Polysons (/MIC.84/) et ne sont pas forcement dans le même contexte. La taille du dictionnaire est alors d'une trentaine d'élements.

Le deuxième dictionnaire (DIC1) est formé de diphones. Toutes les alternances CV, VC et CC contenues dans le corpus des Polysons ont été extraites semi-manuellement, et constituent ainsi un ensemble de plus de 800 élèments. Dans ce cas, les probabilités de confusion entre deux éléments sont assez hautes, car il existe de nombreuses séries de diphones ne différant que par un phonème.

Dans les deux cas, un très grand soin a été apporté lors de la segmentation du corpus. Malgré ces efforts, des irrégularités, particulièrement sur la localisation des frontières à la limite de la zone stable adjacente, ont été notées. Afin de réduire cet inconvénient, et pour harmoniser le dictionnaire avec la nature du signal inconnu, DIC1 a été compressé avec un algorithme similaire à celui employé dans le pré-traitement. La différence réside dans le fait que les seuils de compression ont été dynamiquement ajustés afin de normaliser les durées des transitions à trois tailles: 4, 8 et 16 trames. Le taux moyen de compression sur tout le dictionnaire reste cependant égal à ε. Les expériences menées avec DIC0 l'ont été sans utiliser de compression, ni sur le signal inconnu, ni sur le dictionnaire.

COMPARAISON

L'étage de comparaison est composé de deux parties: la première reçoit les trames (après compression) et les stocke. Le rythme d'arrivée des vecteurs étant perturbé par la compression non linéaire, cet étage sert également de "régulateur de débit". Il est ainsi possible d'obtenir une "fenêtre glissante" de 16 vecteurs dans le flôt de vecteurs inconnus. La forme ainsi constituée est alors comparée à l'ensemble des formes contenues dans le dictionnaire.

L'algorithme de programmation dynamique retenu est un DTW à contraintes d'extrémités fixes, avec un débattement autorisé de ± 5 vecteurs par rapport à la diagonale, et des contraintes locales SAKOE-CHIBA type I.

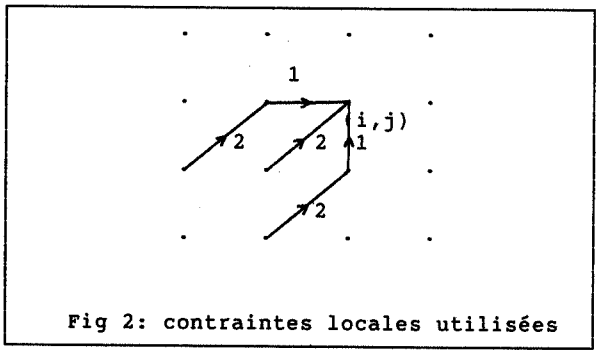


Fig 2: contraintes locales utilisées

Il correspond aux équations:

$$G_{1,1} = d_{1,1}$$

$$G_{i,j} = \min \begin{cases} G_{i-2,j-1} + 2d_{i-1,j} + d_{i,j} \\ G_{i-1,i-1} + 2d_{i,j} \\ G_{i-1,j-2} + 2d_{i,j-1} + d_{i,j} \end{cases}$$

avec $|i-j| < R$ $(R=5)$

$d_{i,j}$ est la distance entre le vecteur i de la forme test (i de 1 à N1) et le vecteur j de la forme de référence (j de 1 à N2).

$$d_{i,j} = \sum_{k=1}^{M} | V_i(k) - V_j(k)| \quad \text{(norme L1)}$$

La distance D(U,T) entre les deux formes U et T est donnée par $G_{N1,N2}$.

"SPOTTING"

La méthode utilisée pour la localisation et la reconnaissance des parties transitoires est une méthode de détection au vol. Si U(t) est la forme inconnue présente dans la fenêtre glissante d'analyse, soit S(t) la fonction:

$$S(t) = \min \{ D(U,T) / T \in \text{Dictionnaire} \}$$

Cette fonction S(t) sera appelée fonction de spotting. Elle présente des minima en fonction du temps, localisés aux endroits de bonne coïncidence entre un élément du dictionnaire et U(t).

La taille des formes issues de U et qui sont comparées aux éléments du dictionnaire dépend de la taille de ceux-ci. L'algorithme de DTW donnant de meilleurs résultats sur des formes de tailles égales, c'est la portion de U commencant à la première trame de la fenêtre glissante, et pour une

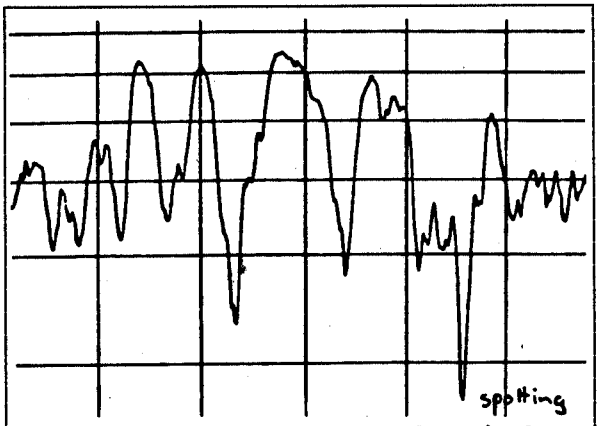


Fig 3: fonction de spotting de /yʃ/ dans DIC0 pour "Il se garantira ... capuchon"

longueur égale à celle de T qui est utilisée.

Pour la première expérience menée (avec le dictionnaire spécifique DIC0), les résultats sont excellents et il est possible de lire sur le listing directement la transcription phonétique de la phrase. L'étude de ce listing fait ressortir plusieurs propriétés:

-la distance de forme à reconnaitre à son homologue contenu dans le dictionnaire présente un minimum au voisinage de la position correcte. Autour de cette position, elle reste suffisament faible pour être la distance de spotting (distance au dictionnaire). On assiste donc à une répétition de l'étiquette correcte durant plusieurs trames.

-les détections parasites sont facilement éliminables par simple vérification de cohérence. Un algorithme de type Viterbi prenant simplement en compte cette question de cohérence trouve la chaine complête sans problème.

HEURISTIQUE

Ces constatations nous ont permis de mettre au point une heuristique qui est appliquée au test avec le dictionnaire de diphones. Dans ce cas, en effet, les minima de la courbe de spotting sont plus "noyés" dans un bruit. Nous avons utilisé le fait que la distance de l'hypothèse correcte à la forme inconnue restait faible au voisinage de celle-ci. Une hypothèse hybride est donc fabriquée en fonction des résultats de la trame sur laquelle se produit le "spotting", ainsi que ceux de la trame suivante et de la précedente. Pour chacune de ces trames, les 5 meilleurs candidats sont retenus, et participent à un vote. Ce vote concerne d'une part l'hypothèse du phonème amont (début de transition) et d'autre part le phonème aval (fin). On obtient ainsi une hypothèse hybride se présentant comme dans la figure 4.

L'étude des propriétes de cette heuristique montre que le choix par vote donne de meilleurs résultats que la simple hypothèse formée par l'étiquette au moment du spotting, bien que les décisions ainsi prises soient souvent les mêmes.

```
temps t
     t+1
      .
      .     F1$F2 F11 F12 F13     ∧ durée
      .     F1$F2 F11 F12 F13     |estimée
      .     F1$F2 F11 F12 F13     | de la
      .     F1$F2 F21 F22 F23     |transi-
      .     F1$F2 F21 F22 F23     | tion
      .     F1$F2 F21 F22 F23     ∨
      .

F1$F2 : transition détectée au minimum de
        la fonction de "spotting".
F11,F12,F13 : trois meilleures hypothèse
              pour le phonème amont.
F21,F22,F23 : trois meilleures hypothèse
              pour le phonème aval.
```

Fig 4: hypothèse hybride

PRE-SELECTION

Afin d'augmenter le taux de fiabilité de l'algorithme, et pour compenser l'absence d'information sur l'énergie du signal (due à la modélisation choisie), une procédure de pré-sélection simple permet, en fonction du taux de passage par zéro (PPZ) et de l'énergie du signal, de n'explorer qu'une partie du dictionnaire de transition.

Cette procédure examine simplement la dynamique contenue dans la transition (différence en log entre l'énergie maximale et minimale) pour détecter les sons plosifs, ainsi que le nombre de trames de signal présentant un PPZ supérieur à un seuil donné pour conclure à la présence d'une fricative.

Dans les deux cas, les conclusions de ce module sont du type oui/non/indécis. Si les valeurs sont trop proches des seuils de décisions, l'indécision est choisie, et le dictionnaire est intégralement exploré. Les résultats de cette procédure appliquée au dictionnaire donnent les résultats suivants:

	Observé	Attendu
présence d'explosion	208	208
absence d'explosion	538	668
indécision	130	
présence de friction	322	290
absence de friction	400	586
indécision	154	

Fig 5: résultats de pré-sélection sur DIC1

Les erreurs subsistantes proviennent d'une prononciation dévoisée de certaines voyelles finales ou même médianes.

RESULTATS

Un test a été effectué sur 50 phrases phonétiquement équilibrées. Le taux de compression adopté est d'environ 2/3. Le dictionnaire DIC1 comporte 876 transitions, et la procédure de pré-sélection est appliquée. DIC1 a été compressé à des formes de taille fixe de 4, 8 ou 16 vecteurs.

Nombre de transitions à trouver	769	
Nombre de transitions correctement localisées et reconnues	504	65%
Nombre de transitions supplémentaires détectées	1292	
Taux de fausse alarme	2.5	

Fig 6: résultats de reconnaissance (DIC1)

Ces chiffres sont bien entendu à rapprocher du nombre de transitions contenues dans le dictionnaire (876). Cependant, une telle expérience donne de bons espoirs sur les possibilités de cette méthode. Le taux de fausse alarme est très important, mais aucune correction (répétition d'étiquettes, incohérence) n'a été appliquée.

CONCLUSIONS

Cette méthode de spotting de transitions donne de bons résultats. De nettes améliorations sont possibles, par exemple en formant un dictionnaire de transition plus adapté à la langue (utilisant les fréquences d'apparitions (/TUB.85/)), et surtout en utilisant plus d'une élocution par forme pour la phase d'apprentissage. Compte tenu de ces conditions expérimentales, les résultats obtenus sont honorables et encourageants. D'autres études sur le même type de méthode donnent également de bons résultats (/ROS 87/). Parallèlement à ce travail, des cellules de VLSI sont étudiées et réalisées, permettant d'assurer la masse de calculs nécessaire en temps réels.

BIBLIOGRAPHIE

/COM 79/ **P. COMBESCURE**
Phrases phonétiquement équilibrées.
Recherches Acoustiques vol VI 1979

/GAU 83/ **JL. GAUVAIN JJ. MARIANI JS. LIENARD**
On the use of time compression for word based recognition.
Proc. of ICASSP 83 pp 22.3.1-4

/MIC 84/ **L. MICLET**
Enregistrement d'une base de données vocale.
Rapport de mission CNET/LLL/TSS/RCP Juillet 84.

/ROS 87/ **A.E. ROSENBERG A.M. COLLA**
A connected speech recognition system based on spotting diphone-like segments - preliminary results.
Proc. of ICASSP 87 pp 3.6.1-4

/TUB 85/ **J.P. TUBACH J.L. BOË**
Un corpus de transcriptions phonétiques: constitution et exploitation statistique.
Rapport ENST 85D001 Avril 85

/VIC 86/ **D. VICARD L. MICLET**
Reconnaissance de parties stables de parole continue pour le décodage Acoustico-Phonétique.
Actes des 15ème JEP Mai 1986
pp 239-242

DECODAGE ACOUSTIQUE-PHONETIQUE PAR RECONNAISSANCE DE DIPHONES

Chen-Guang Wang, Jean-Pierre Tubach

département SYC (CNRS, UA 820)
E.N.S.T., 46, Rue Barrault, 75634, Paris cedex 13

abstract : *This paper describes a study we conducted at ENST about the acoustic-phonetic decoding of continous speech, based on diphone recognition. The diphone dictionary is built automatically by segmentation and labeling of a special corpus. For recognition, we segment the continous speech and determine the "centers" and "gross classes" of diphones to be searched for. Selection of dictionary diphones is performed using a DTW or linear distance, yielding a diphone lattice. Then a majority criterion is used to produce a phoneme lattice.*

1.INTRODUCTION

Nous présentons une étude sur le décodage acoustique-phonétique de la parole continue par reconnaissance de diphones, s'appuyant sur un corpus d'apprentissage dont nous disposons ("polysons"), tout en utilisant un apprentissage automatique.

Le caractère principal qui distingue notre travail des autres est le caractère entièrement automatique de l'apprentissage. Le dictionnaire de diphones est construit automatiquement, en utilisant le programme de segmentation et d'étiquetage. Pour la reconnaissance On se procède de la façon suivante : on a segmenté la phrase à reconnaître pour localiser les endroits susceptibles d'être des centres de diphones, et cherché les candidats de diphones uniquement au voisinage des endroits déterminés par la segmentation.

On s'est basé d'abord sur la valeur des distances, ensuite on a appliqué une décision majoritaire, respectivement sur les consonnes et les voyelles, afin d'extraire du treillis de diphones un treillis de phonèmes de profondeur bien moindre.

2. CONSTITUTION DU DICTIONNAIRE DE DIPHONES

Nous avons utilisé l'analyse LPC (ordre 10), Les 11 coefficients cepstraux (LPCC) sont les paramètres utilisés. La longueur de fenêtre est 10 ms, la fenêtre de Hamming et une préaccentuation sont utilisées.

Le corpus "polysons" [1] est utilisé pour extraire les diphones consonne+voyelle (CV), et voyelle+consonne (VC).

On extrait les CV de mots " a consonne + voyelle t a " ; et les VC de " a t voyelle + consonne a t a". par exemple : le diphone /pa/ est extrait de /a p a t a/, et /a p/ est de / a t a p a t a/.

Les diphones sont extraits automatiquement, par un programme de la segmentation développé à partir de [2], les paramètres utilisés sont énergie totale, ZCR, et énergies partielles dans les bandes (0-1000 Hz, 1000-3500Hz, 500-1000Hz).

On détermine finalement les classes de segments suivantes :

"voyelle", "sonorante", "occlusion", "fricative", "burst", "silence", que nous notons :
VOY, SON, OCC, FRI, BUR, SIL.

Ces désignation mnémoniques ne doivent pas être prises au pied de la lettre par rapport à leur signification habituelle en phonétique : on est satisfait lorsque l'on obtient les correspondances suivantes :

VOY	phonèmes voyelles et semi-voyelles
SON	l, R, m, n et semi-voyelles
OCC	b,d, g,m,n,v
FRI	s, ʃ, f, z, 3, v
BUR	burst de p,t,k
SIL	silence, tenue de p,t,k

L'étiquetage est réalisé en mettant en correspondance le résultat de la segmentation et la transcription phonétique de mots.

Les frontières de diphones sont déterminées de manière automatique. On vérifie d'abord le nombre de voyelles. Si le nombre de voyelles n'est pas correct, on juge qu'il y a une erreur grave de la segmentation, on saute au mot suivant. Dans le cas où la segmentation du mot contenant le diphone considéré semble correcte, on sait localiser ce diphone dans la suite des segments, par exemple dans le cas CV du mot "a CV t a", on aura la configuration suivante :

a	C	V	t	a
<-------->	<---------------------->	<------->	<---------->	<------>
VOY	FRI ou OCC ou SON ou SIL [BUR]	VOY	SIL [BUR]	VOY

Le début de diphone est pris au début de segment FRI, ou OCC ou SON, ou au milieu du segment SIL. La fin est prise au milieu de la voyelle. La situation est exactement inverse pour le cas de VC (à part qu'il n'y pas de burst éventuel).

Le dictionnaire contient uniquement les diphones de type CV, et VC. Il y a 14 voyelles (a o e i y oe u ɛ ɸ ɔ ɔ̃ œ̃ ɛ̃ ɑ̃ , et on n'a pas distingué les 2 /a/), et 18 consonnes (p t k b d g v z 3 f s ʃ l R m n w j, on a omis ɥ et ʁ). Le nombre de diphones CV est donc 14x18, de même pour les diphones VC (14x18), plus les 14 diphones $V ($ =silence), on en a au total 518. Dans le dictionnaire dont nous disposons, certains diphones sont absents. Ils ont été éliminés soit à cause d'erreur de la segmentation, soit à cause de leur courte

longueur (inférieure à 50 ms). Il manque environ 20% des diphones dans le dictionnaire.

La longueur des diphones est très variée, de 5 trames à 23 trames (1 trame = 10 ms).

3. EXPERIENCES DE RECONNAISSANCE

La méthode consiste à utiliser sur la parole continue la même programme de segmentation qui est utilisé pour la détermination des diphones à l'apprentissage : on utilise les frontières de segments et les étiquettes de classes phonétiques grossières qu'il fournit pour réduire la combinatoire de la recherche de diphones systématique selon la méthode précédente, en se restreignant à des sous-ensemble pertinents du dictionnaire et des instants de comparaison.

On divise le dictionnaire de diphones en sous ensembles, en fonction des étiquettes grossières susceptibles d'être fournies par le programmme de segmentation : VOY, SON, OCC, BUR, FRI, SIL.

Ces sous-ensembles sont :

1 SIL + VOY ou SIL [BUR]+VOY

2 FRI + VOY

3 OCC ou SON + VOY

4 VOY + SIL

5 VOY + FRI

6 VOY + OCC ou SON

En reconnaissance, plaçons-nous à une frontière de segmentation : elle est susceptible de constituer le "centre" d'un diphone. On en détermine les limites de ce diphone potentiel exactement comme à l'apprentissage (segment FRI, SON, OCC entier, moitié de segment VOY ou SIL). Entre ces bornes, on va faire une comparaison (linéaire ou dynamique) avec les diphones qui possèdent une configuration de classes grossières comparable, c'est-à-dire qui appartiennent à un seul des 6 sous-ensembles du dictionnaire définis ci-dessus.

Par exemple à une frontière OCC VOY, on restreindra la recherche aux diphones OCC VOY ou SON VOY, c'est-à-dire à (b,d,g,l,R,m,n,w,j,v) (voyelle).

Pour tous les cas, après la comparaison on trie les diphones suivant les distances dans l'ordre décroissant, les 5 premiers diphones qui ont les plus petites distances sont choisis comme réponse, ce qui nous donne un treillis de diphones.

La comparaison dynamique est utilisée dans l'expérience : on compare chaque diphone du sous ensemble considéré du dictionnaire avec la zone "candidat diphone" déterminée par la segmentation (voir ci dessus, la longueur de cette zone est *l*), qui va des fenêtres *i* à *i+l*, et également avec les zones *i-1* à *i+l-1*, et *i+1* à *i+l+1* ; on garde le meilleur résultat comme distance (ceci est destiné à ne pas être trop tributaire de la précision de la segmentation)

Pour l'exemple suivant, la phrase à reconnaître était "ce petit canard apprend à nager".

Le résultat de la segmentation :

fichier no 9 Nb de segments = 25

	type	debut	fin	
s	SILENCE	1	24	
ə	FRICATIVE	25	30	
p	VOYELLE	34	37	
ə	SILENCE	41	44	
t	VOYELLE	46	50	
.	SILENCE	53	56	
i	FRICATIVE	57	59	
k	VOYELLE	61	63	
a	SILENCE	66	70	
n	BURST	71	72	
a	VOYELLE	73	77	
R	SONORANTE	78	85	
a	VOYELLE	86	105	
pR	OCCLUSION	106	119	
a	VOYELLE	120	124	
ʒa	SILENCE	128	132	
n	OCCLUSION	133	141	
æ	VOYELLE	142	160	
	SONORANTE	161	161	*
ʒ	VOYELLE	162	164	
e	OCCLUSION	165	167	
	VOYELLE	168	174	
	SILENCE	175	186	
		187	192	
		200	282	

* erreur de segmentation

Le treillis de diphones :

s	25	36	SOE	SEU	SAU	SON	SOU	
ə	36	42	EU P	ON P	EU T	EU K	OE T	
p	43	48	PEI	KUN	K A	KIN	KOE	
ə	48	54	EI P	EI T	OE K	AI T	U K	
t	57	62	S I	Z I	F I	SAN	SEI	*
i	62	72	I K	OU T	OU K	I T	EI K	
k	69	75	K A	KEI	$EI	KAI	TEI	
a	75	82	AI M	A N	OE M	AI L	AI R	
n	82	96	L A	M A	N A	N O	MAI	
a	96	113	A R	IN R	UN R	OE R	A B	
R	113	122	B A	B O	M O	M A	RUN	
a	122	129	OE P	A P	A T	O T	O P	
pR	137	151	R O	B A	MAU	B O	M A	
ʒa	151	161	A N	A M	IN R	O N	UN R	
n	161	163	NEI	NEU	NAI	NON	N U	* *
	163	166	EI N	EU N	AI N	OE N	AU N	
æ	166	171	L A	NAI	MAI	NOU	LAI	
	171	181	OU L	I N	A G	UN Y	AI L	
ʒ	181	190	N U	G A	GIN	BEI	G U	
ə	190	201	U T	EI P	AI T	AI P	EI T	

* burst long de /t/ reconnu comme fricative /s/.
** erreur de segmentation

On a, comme dans la suite, souligné les phonèmes exactes, en admettant les confusion voyelle ouverte -voyelle fermée, de type ϕ -∞(EU-OE) ainsi que $\tilde{\varepsilon}$ - $\tilde{\infty}$ (IN-UN).

N.B. la denière ligne s'explique par l'absence de références voyelle+silence dans le dictionnaire.

On peut utiliser un critère majoritaire pour obtenir un treillis de diphones réduit, puis un treillis phonémique. On procède de la façon suivante :

A partir du treillis de diphones à 5 candidats, on obtiendra un treillis réduit par procédure majoritaire. Chaque ligne est traitée indépendamment. Les moitiés gauches et droites (consonnes, voyelles) sont considérées séparément. Pour chacune on utilise les règles suivantes :

a) si on a au moins 3 fois le même phonème, on le garde.

b) si on a deux fois le même phonème, on le garde (il peut y en avoir deux)

c) si aucun phonème ne paraît plus d'une fois, on ne change rien.

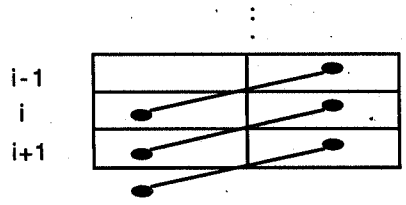
le treillis "réduit" à partir du treillis de diphones ci-dessus :

	partie gauche				partie droite				
s	S				* | *	EU		
ə	EU				* | *	P	T	
p	K				* | *	IN		
ə	EI				* | *	T	K	
t	S				* | *	I		
l	OU	I			* | *	K		
k	K				* | *	EI		
a	EI				* | *	M		
n	M	N			* | *	A		
a	A	IN			* | *	R		
k	B	M			* | *	A	O	
a	A	O			* | *	P		
pR	B	M			* | *	O		
ãa	A	IN			* | *	N	R	
	N				* | *	EI		+
n	EI	EU			* | *	N		| *
ɥ	N	L			* | *	EI		+
	OU	I	A	UN AI	* | *	L		
ʒ	G				* | *	U		
e	EI				* | *	T		

*erreur de segmentation.

Une fois que l'on a le treillis de diphones "réduit", on le transformera en treillis de phonèmes.

Plaçons nous à la i ème ligne de ce treillis réduit: on prendra pour partie gauche l'intersection de la partie droite de la i-1 ème ligne et de la partie gauche de la i ème. De même, on prendra pour partie droite l'intersection de la partie droite de la i ème ligne et de la partie gauche de la i+1 ème (si une telle intersection est vide, on prend l'union des deux parties correspodantes).



Exemple :

EI, EU | EI, AU ----> EI /
EI,EU | OU ----> EI EU
OU /

le treillis phonémique obtenu pour les données précédentes :

S /

EU /

P T
K /

IN /
EI /

T K
S /

I /

K /

EI /

M /

A /

R /
B M

A O

P /
B M

O /
A IN

N /

EI /

N /

EI /
?

L /
G /

U /
EI /

T /

4. CONCLUSIONS

Il est vrai que la segmentation peut introduire certaines erreurs dans le résultat final ; néanmoins d'après notre expérience la reconnaissance avec segmentation et pré-classement nous donne le meilleur résultat, en plus le temps de calcul est beaucoup réduit. Les erreurs du résultat final sont principalement causées par les erreurs de la segmentation : premièrement les erreurs de la segmentation en construction de dictionnaire, ceci fait que dans le dictionnaire, il manque des diphones(20%) ; deuxièmement les erreurs de segmentation en reconnaissance, comme on a divisé le dictionnaire en plusieurs sous-ensembles, une petite erreur pénalise le résultat final. En travaillant plus sur la segmentation, on pourrait avoir de meilleurs résultats. Pour l'ensemble de cette expérience, on s'est aperçu que les résultats pour les consonnes sont meilleurs que pour les voyelles.

Un développement intéressant serait donc de combiner notre reconnaissance de consonnes au moyen de diphones avec des résultats de reconnaissance des noyaux stables par une autre méthode, celle de Vicard par exemple [3]. On pourrait

ainsi, soit corriger nos résultat sur les voyelles, soit limiter les diphones candidats à ceux qui comportent les voyelles le plus probablement reconnues par ailleurs. C'est à notre avis dans ce sens que doit s'orienter la poursuite de ce travail.

Par ailleurs, nos expériences ont été menées en monolocuteur, le corpus "polysons" d'apprentissage et les phrases phonétiquement équilibrées (le corpus de reconnaissance) [4] ayant été enregistrés par la même personne. Nous pensons que la méthode d'adaptation au locuteur, développée par Choukri [5] à partir des idées de Grenier pourrait être adoptée pour parvenir à une solution plus générale du décodage acoustique-phonétique.

5. BIBLIOGRAPHIE

[1] **L. Miclet** : *Enregistrement d'une Base de Données Vocale.* Rapport de Mission au CNET LAA-RCP, ENST, Paris, Juillet 1984

[2] **M. D. Di Benetto, J. P. Tubach** : *Two Cooperative Methods for the Segmentation of Runing Speech.* Congrès FASE/DAGA, Goettingen, RFA, pp. 907-910, Septembre1982

[3] **D. Vicard, L. Miclet** : *Steady Part Recognition of Continuous Speech for Acoustic - Phonetic Decoding.* Proc. of ICASSP 1986. Avril, 1986

[4] **P. Combescure** : *Vingt Listes de Dix Phrases Phonétiquement Equilibrées.* Revue d'Acoustique, Vol. 14, No.56, 1981

[5] **K. Choukri, G. Chollet, Y. Grenier** : *Spectral Transformations Through Canonical Correlation Analysis for Speaker Adaptation.* Proc. ICASSP. pp.2659-2662, 1986

SYSTEMES VOCALIQUES

LES CONSEQUENCES LINGUISTIQUES D'UN POSSIBLE CONTROLE LINGUISTIQUE DU PHARYNX
Le cas des voyelles [± ATR] du lamba (Togo) et leur variabilité dans la zone de focalisation des [i]

A.S. ARITIBA[1] C. ABRY[2] & L.-J. BOE[2]

[1] Centre de Dialectologie Africaine, GRENOBLE III
[2] Institut de Phonétique de Grenoble
Institut de la Communication Parlée
UA CNRS n° 368

ABSTRACT

Acoustic data concerning high vowels in a Voltaic language (Lamba) with an [Advanced Tongue Root] contrast, are used to argue that changes, observed mainly on F1, can be obtained by a simple widening of the pharynx (with or without larynx lowering).

Previous simulations of [ATR] starting from a constriction hypothesis [-ATR] are shown (with a simplified six-tube modelling in FANT's vein) to be inadequate to characterize this maneuver properly, since they obtained for [+ATR] control the same acoustic effects as for tongue raising (i.e. F1 lowering and F2 raising), especially for [i] vowels.

Hence we will cope with the main difficulties in separating the two acoustic effects in this anterior region that we named [i] "focalization" zone.

INTRODUCTION

Parmi les traits des voyelles encore mal connus, le trait [Advanced Tongue Root] occupe une "place de choix". Pour ces voyelles dites [Tendues], [Hautes Elevées], [Couvertes], [ATR] (depuis [1]), [Expansées], etc., les données radiologiques (revue in [2]; en France [3]) ont mis en évidence une expansion du pharynx, accompagnée d'un abaissement du larynx, comme manoeuvre essentielle pour harmoniser les voyelles de base [-ATR] avec les voyelles gouvernantes [+ATR].

Ce trait a en outre - hormis la spécification de "l'inconnu" - un autre attrait heuristique : le rôle du pharynx étant encore généralement mal compris, dans la production courante des voyelles [4], nous espérons qu'une commande linguistique nettement différenciée pourra avancer nos connaissances sur les contrôles possibles de cette partie du conduit vocal.

Cette première description acoustique du lamba offre aussi, à notre connaissance, un lieu d'observation particulièrement heureux : les conséquences acoustiques d'[ATR] (limitées ici à un abaissement de F1) y semblent nettement distinctes de l'effet habituel d'une élévation de la langue (avec F1 variant non orthogonalement à F2); effet duquel les données actuellement disponibles sur [ATR] ne sont pas bien parvenues à le différencier ([2]; d'où la représentation de ces voyelles, dans une base de données phonologiques récente comme UPSID [5], sur les seuls critères de qualité : [Lowered High Vowel] pour [-ATR], comme pour [-Tense]).

[ATR] DANS LE SYSTEME VOCALIQUE DU LAMBA

Le lamba est une langue gurunsi (langues voltaïques) parlée au nord du Togo. Le parler décrit ici est celui de Défalé. Son système vocalique comprend, selon l'analyse phonologique que nous avons effectuée [6], neuf voyelles :

- 6 hautes, dont :
  - . 3 [+ATR] : [i, ɪ̟, u]
  - . 3 [-ATR] : [I, ɪ̠, ʊ]
  - . 2 moyennes : [ə, ɔ]
  - . 1 basse : [a]

L'abondance de voyelles centrales place notre langue parmi les 5,5 % de systèmes qui ont, selon Hagège [7], un rapport voyelles/voyelles non périphériques de 9/2 à 9/3. Si la présence de la voyelle haute centrale non arrondie [ɪ̟] (dans 10 % des quelques 200 langues examinées par Crothers [8]) avec [ə] ne surprend guère (47 % des cas selon Hagège [7]) on remarque une voyelle plutôt rare [ɪ̠], voyelle dite haute centrale non arrondie **abaissée** (ou [-ATR]); ce qui forme un contraste extrêmement peu fréquent (pour un autre exemple dans le grand ensemble Niger-Congo, cf. le dan, langue mendé [9]). A remarquer aussi un espace antérieur très lacunaire ([e] ou [ɛ] absents; comme dans seulement 9 des 317 langues d'UPSID; commentaire général sur ces "systèmes à trous" par Disner in [5]);

Le corpus que nous avons utilisé pour les mesures acoustiques comprend deux contextes :

- [hV́] (ton haut), pour toutes les voyelles sauf [ɪ̠], qui n'apparaît pas dans ce contexte.
- [wV́l], pour [ɪ̠] et pour les voyelles adjacentes [I, ɪ̟, ə, ʊ] ; ce qui nous a permis de la replacer dans l'ensemble.

Ces items, insérés dans les phrases porteuses [Í — lɔ́] (litt. "Vous -verbe -où ?") ont été prononcés, en ordre aléatoire, par un des auteurs A.S. A., âgé de 29 ans, natif de Défalé, trilingue (français et kabyé), niveau académique : DEA.

Les signaux enregistrés en chambre sourde (Nagra 4.2, micro Beyer Dynamic M69N) ont été numérisés à 10 kHz sur 12 bits. Un éditeur de signal [10] a permis d'extraire les réalisations de la phrase porteuse et de les contrôler; Fo et les 5 premiers formants ont été mesurés, après obtention d'un sonagramme (avec suivi de formants, à partir du cepstre) [11], sur trois prélèvements autour du centre acoustique de la voyelle.

Dans le plan F1/F2 (fig. 1), les dispersions des voyelles (10 par ellipses, à 90%) nous permettent de retrouver aisément les classes phonologiques. Les six voyelles hautes sont quasiment équiréparties : on constate simplement une tendance, courante pour les centrales, à se centraliser encore davantage. Noter qu'il nous est difficile d'évaluer ce qui, dans la situation de [ɪ̠] revient au contexte [wV́l], l'effet centralisant s'étant révélé net sur les voyelles [I, ə, ʊ], mais nul sur [ɪ̟].

En fait la seule"irrégularité" du système par rapport aux prédictions phonologiques vient de la position de la voyelle centrale [ə], pour laquelle les valeurs du premier formant ne sont pas aussi élevées que celles de sa partenaire moyenne [ɔ], et le second formant un peu plus bas que celui des

autres voyelles centrales.

En vue de la discussion qui va suivre, nous nous contenterons d'insister ici sur la possibilité de passer des 3 voyelles [-ATR] aux 3 [+ATR] en ne changeant pratiquement que F1 (les variations sur F2 n'étant pas statistiquement significatives), avec des valeurs de F1 particulièrement basses (jusqu'en dessous de 200 Hz).

MACRO-SENSIBILITES [ATR] DANS LA ZONE DES [i]

Nous n'utiliserons pas dans cette discussion les travaux sur les fonctions de sensibilité [12], valides seulement pour des micro-variations ; nous parlerons de **macro-sensibilités** pour des ordres de variation au delà de 10% [13].

Dans ce sens plusieurs simulations ont été proposées, non spécifiquement pour le trait [ATR] mais en relation avec [Rhotacisé] [14], et [pharyngalisé] [15], traits qui comprendraient tous deux une rétraction de la langue [-ATR]. Les résultats obtenus en manipulant les coupes sagittales pour rétrécir le pharynx au niveau de l'épiglotte, sur les versions d'un modèle issu des travaux de Ladefoged et al. [16], donnent une tendance générale à l'abaissement de F1, accompagné d'une élévation de F2. Les sensibilités sont évidemment différentes selon les voyelles : F2 étant très sensible pour [i] et peu pour [u]. L'effet crucial, pour reproduire nos données, est donc à rechercher du côté de [i], pour lequel les variations de F2 doivent être au contraire minimisées (cf. fig. 1). Notons que les manoeuvres simulées ajoutent toutes une **constriction** supplémentaire dans le conduit vocal : ce qui empêche de rendre compte des effets d'un simple élargissement pharyngal pour [+ATR], à partir d'une configuration courante, "plain" (ou [0 ATR]; ici "plain" [-ATR]). Or rappelons que, contrairement à la pharyngalisation, p. ex., il semble bien que la manoeuvre linguistique, pour l'harmonie vocalique, à partir des voyelles de base [-ATR], soit bien un élargissement. L'idéal serait bien entendu de simuler cette expansion sur un modèle qui possède un contrôle du pharynx intégré aux autres manoeuvres (en particulier l'élévation-avancement de la langue, comme c'est le cas pour **front-raising** et **back-raising** in [16], ou pour le **dos** et le **corps** in [17]). A notre connaissance un tel modèle n'existe pas encore.

Nous avons donc repris le problème "à la base", sur le modèle des simulations simplifiées de Fant [18]. Seulement celles-ci n'étant qu'à 4 tuyaux, nous avons ajouté la possibilité d'une constriction de 5 cm de long dans le pharynx (soit 6 tuyaux), et d'un allongement de 1 cm du côté du larynx (conduit de 16 à 17 cm). Les calculs ont été faits en utilisant un logiciel de Charpentier [19], simulant les pertes par viscosité, chaleur, vibration des parois et l'effet de rayonnement aux lèvres. Nous donnons les résultats pour un déplacement (entre 10 et 12.5 cm de la glotte) de la constriction buccale dans la zone de focalisation [20] du [i], intégrant du même coup, pour le mouvement supposé de la racine, de possibles sensibilités dues à un éventuel déplacement concomitant de la constriction linguo - palatale [21].

1. Manoeuvres simples.

A. Elargissement du pharynx :
de "plain" [-ATR] à [+ATR]

Nous distinguerons cet élargissement (dans notre modélisation simplifiée : un simple accroissement de la section du tuyau dit pharyngal) d'une **déconstriction** du pharynx (cf.infra). L'agrandissement d'une cavité de 2 cm² de section (fig. 3) à 8 cm² (fig. 4) ne change pratiquement pas F2 et F3; mais essentiellement F1 qui décroît (noter, aussi, l'élargissement des bandes passantes).

B. Déconstriction du pharynx :
de [-ATR] à "plain" [+ATR].

C'est le type de manoeuvre le plus proche des simulations précitées ([14] et [15] : de "plain" à [-ATR]). On obtient bien, en supprimant une constriction de 2 cm² de section (fig. 5), pour restaurer un pharynx à 8 cm² (fig. 4), une petite baisse de F1, et des élévations importantes pour pour F2 et F3 (ce dernier stable chez [15]).

C. Abaissement du larynx :
de "plain" [-ATR] à [+ATR]

L'accroissement de 1 cm de la longueur du conduit vocal, du côté du pharynx (figs. 4 et 6), produit un abaissement du deuxième et surtout du troisième formant, laissant le premier pratiquement inchangé.

2. Manoeuvres composées

Pour plus de simplicité nous considérerons les manoeuvres **Elargissement** (A) et **Déconstriction** (B) comme concurrentes. Restent donc deux manoeuvres composées possibles.

D. Elargissement du pharynx et abaissement du larynx : de "plain" [-ATR] à [+ATR].

Un tel élargissement, à partir d'une section de 2 cm² (fig. 3), à une section de 8 cm², avec abaissement de 1 cm du larynx (fig. 6), donne une diminution nette de F1 et de F3 (F2 n'étant pas touché).

E. Déconstriction du pharynx et abaissement du larynx : de [-ATR] à [+ATR].

En passant d'une constriction de 2 cm² (fig. 5) aux 8 cm² du pharynx, allongé de 1 cm (fig. 6), nous baissons F1 et élevons de beaucoup F2.

3. Discussion.

En comparant ces résultats entre eux, on s'aperçoit que les seules tendances qui correspondent nettement à nos données lamba sont l'élargissement du pharynx seul (cas A) ou accompagné de l'abaissement du larynx (cas D) : nos voyelles [+ATR] ont en effet essentiellement un F1 plus bas pour un F2 peu changé. Leur comportement ne semble donc pas pouvoir induire : une manoeuvre de déconstriction (cas B) qui, accompagnée d'un abaissement du larynx (cas E) tend à renforcer l'effet d'élévation sur F2; ni, non plus, un abaissement du larynx seul (cas C) qui correspond essentiellement à un abaissement de F2 seul.

Notons que les manoeuvres d'abaissement du larynx et de déconstriction sont contradictoires pour F2 (re : abaissement contre élévation), alors qu'elles sont "additives" sur F1 (abaissement).

CONCLUSION

Nous avons ainsi présenté, à notre connaissance le seul cas d' [ATR] qui nous permette d'induire la ou les manoeuvre(s) pharyngale(s) changeant essentiellement F1 et non F1 et F2. Ce qui a pour avantage de ne pas aboutir à la confusion des produits de l'élévation du point le plus haut de la langue avec les conséquences d'une expansion du pharynx. Nous montrons par là qu'il est possible, dans une certaine mesure, d'**orthogonaliser les deux gestes phonologiques.** Dans cette perspective, d'autres données articulatoires et acoustiques, sur d'autres langues possédant cette harmonie, seront à collecter, avant la nécessaire intégration de ces connaissances dans les modèles articulatoires existant, pour améliorer leurs contrôles pharyngaux.

Remerciements

A Denis Creissels, pour l'inlassable effort de description des langues de l'Afrique de l'Ouest qu'il poursuit et auquel il a su nous intéresser de longue date.

REFERENCES

[1] STEWART J.M. (1967) Tongue Root Position in Akan Vowel Harmony. Phonetica 16, 185-204.

[2] LINDAU M. (1979) The Feature Expanded. J. of Phonetics 7, 163-176.

[3] RETORD G. (1977) Etude radiocinématographique des articulations de l'agni-sanvi. Thèse de Doct. d'Etat, Paris III

[4] JOHANSSON C. SUNDBERG J. WILBRAND H. & YTTERBERGH C. (1987) From Sagittal distance to area. A Study of transverse, cross-sectionnal area in the pharynx by means of computed tomography. ms.

[5] MADDIESON I. (1984) Patterns of Sounds. Cambridge University Press.

[6] ARITIBA A.S. (1987) Le lamba de Défalé (Togo). Description phonologique et morphologique. Thèse de 3° Cycle, Université Grenoble III (à paraître).

[7] HAGEGE C. (1982) La structure des langues. P.U.F., Paris.

[8] CROTHERS J. (1978) Typology and Universals of Vowel Systems. in GREENBERG J.H. FERGUSON C.A. & MORAVCSIK E.A. (eds.), Universals of Human Language 2, 93-152. Stanford University Press.

[9] WELMERS W. (1973) African Languages Structures. Berkeley & Los Angeles Univ. Press.

[10] BENOIT C. (1984) EDISIG : Encore un éditeur de signal ?! 13° JEP du GCP du GALF, 211 -213.

[11] FENG G. (1983) Analyse cepstrale, visualisation sonagraphique et détection de formants. Séminaire GALF - GRECO "Analyse du Signal de Parole", Paris, 206 - 217.

[12] FANT G. & PAULI S. (1975) Spatial Characteristics of Vocal Tract Resonance Modes. Speech Comm. Seminar 2, 121-132. G. FANT (ed.) Almquist & Wiksells, Stockholm.

[13] MAJID R. (1986) Modélisation articulatoire du conduit vocal. Exploration et exploitation. Fonctions de macrosensibilité paramétriques et voyelles du français. Thèse Doct. Ingénieur, INP Grenoble.

[14] LINDAU M. (1975) Vowel Features. Language 54, 541-563.

[15] NORLIN K. (1987) A Phonetic Study of Emphasis and Vowels in Egyptian Arabic. Lund Univ. Depart. of Linguistics Working Papers 30.

[16] LADEFOGED P. (1985) Macintosh Models and Plots for Phoneticians. UCLA Working Papers in Phonetics 61, 65-71.

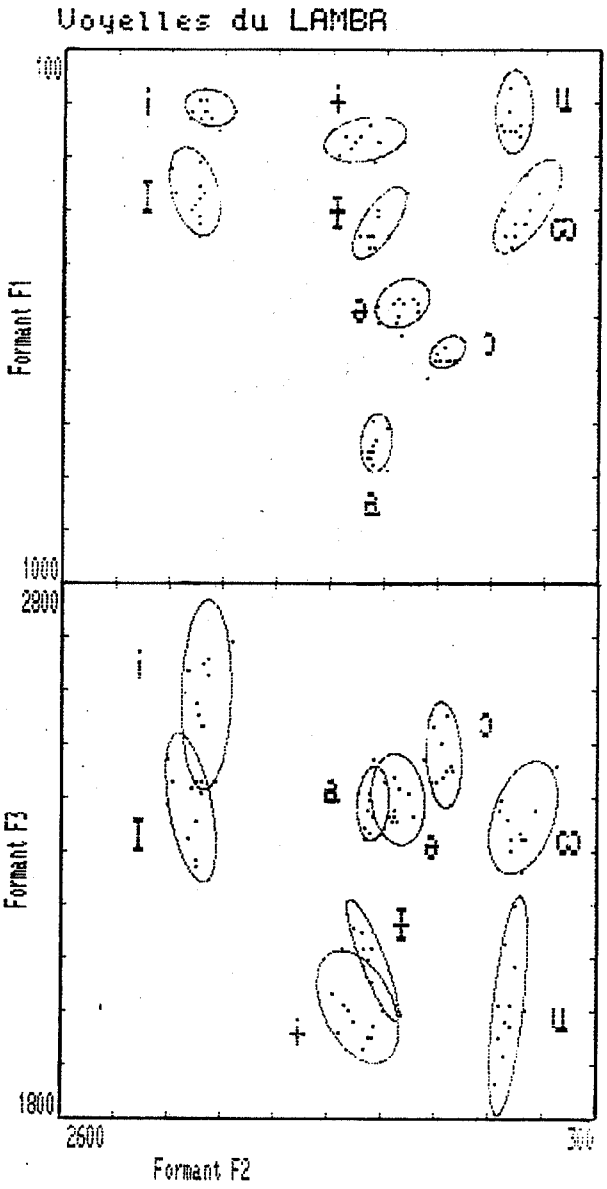
[17] MAEDA S. (1979) Un modèle acoustique basé sur une étude acoustique. Bull. Inst. Phonétique de Grenoble 8, 35-55.

[18] FANT G. (1960) Acoustic Theory of Speech Production. Mouton, The Hague.

[19] CHARPENTIER F. (1982) Un logiciel de simulation du conduit vocal. CNET (comm. pers.).

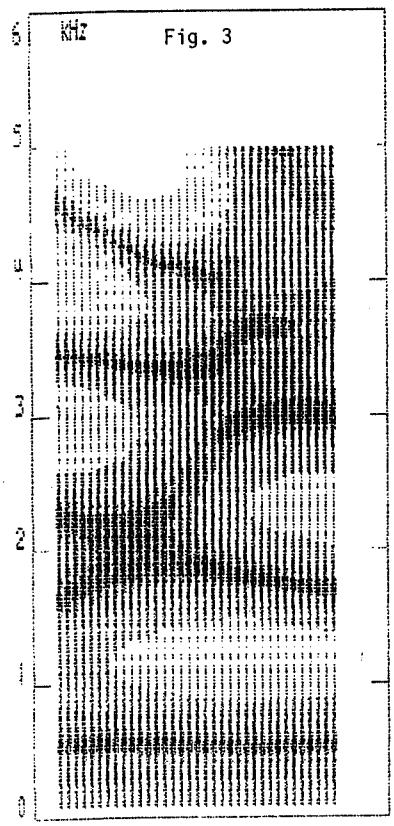
[20] BOE L.J. & BOE L.J. (1986) Nomogrammes et systèmes vocaliques. 15° JEP du GCP du GALF, 303-306.

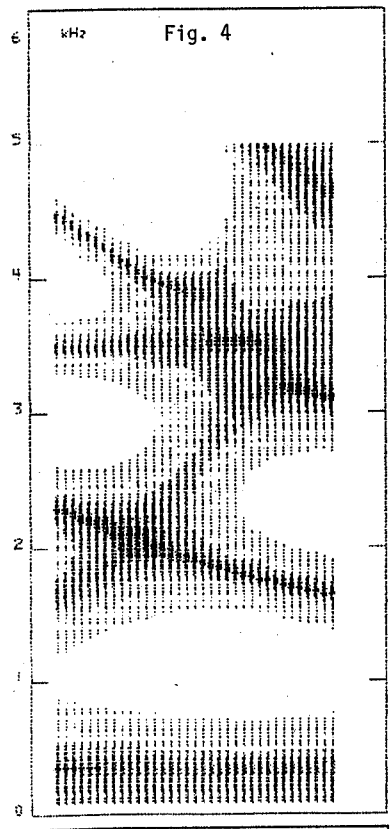
[21] WOOD S. (1986) The Acoustical Significance of tongue, lip, and larynx maneuvers in rounded palatal vowels. J. Acoust. Soc. Am. 80, 391-401.

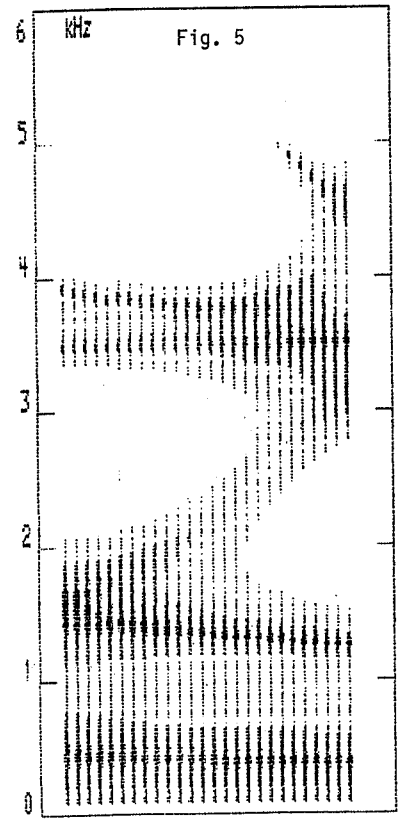


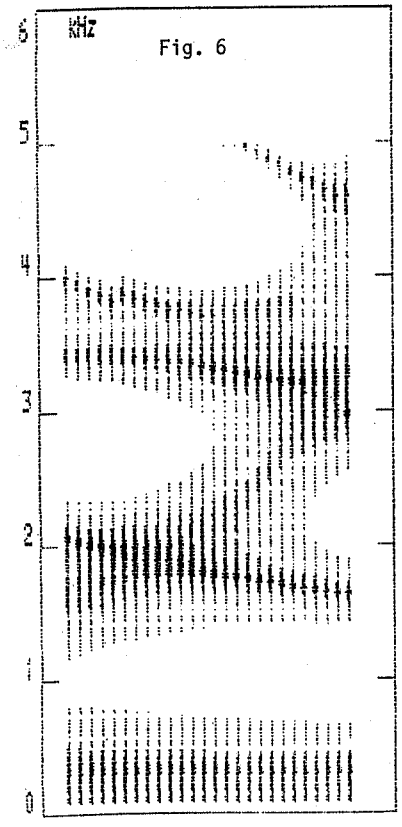
Figures 1 et 2. Ellipses de dispersion (à 90%) des voyelles du lamba dans le plan F1/F2 (en haut) et F2/F3 (en bas), en Hz.
[i, ɨ, u] sont [+ATR]
[I, ɪ, ʊ] étant [-ATR]

N.B. Contexte [hV́], sauf pour [ɪ] (contexte [w-́l]).









Nomogrammes autour du point focal [i]. Modèle à 6 tuyaux simulant :

* les lèvres: A_1 = 4 cm², l_1 = 1 cm
* la cavité d'avant : A_2 = 8 cm²
* la constriction buccale : A_3 = 0.65 cm², l_3 = 5 cm, avec X_c (abscisse du milieu de la constriction par rapport à la glôtte) variant de 10 à 12.5 cm (de 11 à 13.5 cm pour l_6 = 2 cm)
* la cavité haut-pharyngale : A_4 = 8 cm² à 2 cm²
* la constriction pharyngale : A_5 = 8 cm² à 2 cm², l_5 = 5 cm
* la cavité bas-pharyngale : A_6 = 8 cm² à 2 cm², l_6 = 1 ou 2 cm

N.B. $l_2 + l_3 + l_4$ = 9 cm

Fig. 3 : l_6 = 1 cm, $A_4 = A_5 = A_6$ = 2 cm²
Fig. 4 : l_6 = 1 cm, $A_4 = A_5 = A_6$ = 8 cm²
Fig. 5 : l_6 = 1 cm, A_4 = 8 cm², A_5 = 2 cm₂, A_6 = 8 cm₂
Fig. 6 : l_6 = 2 cm, $A_4 = A_5 = A_6$ = 8 cm²

RECONNAISSANCE DES VOYELLES ET DES TRAITS VOCALIQUES EN FRANCAIS

ANNE BONNEAU

Cnet LAA/TSS/RCP
22300 Lannion FRANCE

MARIO ROSSI

Institut de phonetique
13100 Aix-en-Provence FRANCE

ABSTRACT

This paper concerns two methods aiming to the automatic recognition of French vowels in continuous speech. The first part presents the results obtained by an algorithm based on the detection of context- and speaker-independent acoustic cues for the detailed identification of the vowels. The second part concerns the preliminary results obtained for the detection of the features open/close and front/back, by context-independent cues and partially speaker- independent cues (the frequency ranges on which certain rules operate are adapted to the sex of the speaker). The limits of the two methods are discussed. It is suggested that the recognition of the vowels should be performed using a mixed strategy: an "invariant feature" recognition module, to classify the vowels, followed, for each vowel class, by a specific module which would partially be speaker- and context- dependent.

INTRODUCTION

La reconnaissance fine des voyelles indépendamment du contexte et du locuteur est particulièrement difficile dans des langues comme le français qui possèdent un système vocalique très riche (cf figure 1). En revanche des connaissances fiables et relativement invariantes semblent pouvoir être dégagées pour la reconnaissance des grands traits; celles-ci sont actuellement utilisées par beaucoup de systèmes [1] [2] pour permettre un premier accès au lexique suivi d 'une stratégie d' "hypothèse-vérification" pour la reconnaissance des mots. Pour ce qui concerne l' identification directe des voyelles, sans accès préalable au lexique, on peut envisager une stratégie mixte: élaborer un module de reconnaissance des traits, relativement indépendant du locuteur et du contexte, suivi pour chaque classe vocalique identifiée, d' un module spécifique, monolocuteur et dépendant du contexte. Cette stratégie nous a été suggérée par les résultats actuels de la reconnaissance des voyelles du français et particulièrement par ceux de l' algorithme multi-locuteur et polycontextuel de Rossi [3]. La reconnaissance de ce module et son évaluation constituent la première partie de cet article, la seconde traite du module de reconnaissance des traits d' ouverture et d' antériorité.

I RECONNAISSANCE DES VOYELLES PAR L' ALGORITHME DE ROSSI

I.1 Algorithme et evaluation

Les règles s' appliquent sur la région centrale de la voyelle. La reconnaissance est effectuée selon un mode binaire, dans une arborescence où les indices sont hierarchisés. L' analyseur acoustique utilisé est un vocoder à 14 canaux. Le paramètre acoustique principal est le vecteur des énergies dans chaque canal du vocoder. Les indices testent généralement la forme spectrale: c.a.d effectuent des comparaisons entre les niveaux d' énergie dans deux ou plusieurs zones fréquentielles du spectre.

Par exemple, une des règles est:
si EK1 >= (EK3 + EK4) alors l' indice ouvert1 est vrai
où EK1, EK3, EK4 représentent le niveau d' énergie dans le 1er, 3ème et 4ème canal.
Les règles ont été conçues sur un corpus d' apprentissage constitué de 320 logatomes de type CVCV puis implémentées sous le système expert SERAC développé au CNET. Les corpus d' évaluation sont constitués de 20 phrases, prononcées par deux locuteurs masculins, de 50 nombres connectés, prononcés par deux autres locuteurs masculins, et de 300 nombres, prononcés par 6 locuteurs masculins. Le nombre de candidats proposés pour chaque voyelle varie de 1 à 3 selon la règle déclenchée. Tous corpus confondus, la liste des candidats contient la bonne solution dans 75% des cas en moyenne, dans 52% des cas, celle-ci est donnée par le premier candidat (cf table 1).

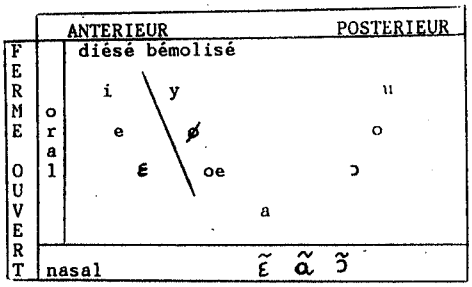


Figure 1: Triangle vocalique du français

I.2 Discussions

Les confusions apparaissent principalement entre:

-les deux voyelles fermées d' arrière /u/ et /o/;
-/oe/ et son correspondant nasal /œ̃/;
-les trois voyelles nasales /ɑ̃/, /ɔ̃/, /ɛ̃/.

Certaines erreurs sont imputables au fait que la région centrale de la voyelle n' est pas stable et inclut des transitions et aussi à un manque d' information dans les basses fréquences. Nous pensons également que les indices binaires ne sont pas très bien adaptés à cette étape de la reconnaissance; à ce sujet Klatt [4] a déjà parlé de "the undesirability of forcing an early decision".

Bien que l 'algorithme passe directement des indices aux traits sans une étape intermédiaire bien définie qui serait un module de reconnaissance des traits, il nous semble intéressant d' évaluer les confusions apparaissant entre des voyelles de classe opposée: confusions entre des voyelles ouvertes et fermées, orales et nasales...

Faisons d' abord quelques remarques sur la classification de certaines voyelles:
1) Le degré d'ouverture des voyelles moyennes -/e,ɛ,o,ɔ ,ø,oe/- n' est pas toujours facile à déterminer et ce ni à l' écoute (où la classification est souvent arbitraire) ni par des règles théoriques du type: voyelle ouverte en syllabe fermée et vice-versa, pas toujours suivies dans la réalité. Ces considérations nous ont conduits à ne pas prendre en compte les confusions survenant entre les voyelles moyennes. Deux degrés d' ouverture seulement sont donc distingués.
2) Nous adoptons les mêmes restrictions pour les voyelles /a/,/ ɛ̃ /,/oe/,/ ø /, qui ne sont pas classifiées à-priori sur l'axe antéro-postérieur.

CORPUS DE TEST	LOCUTEURS (masculins)	LISTE DE CANDIDATS	PREMIER CANDIDAT
.PHRASES (20)	2	79	55
.NOMBRES CONNECTES (50)	2	76	53
.NOMBRES (300)	6	72	50

Table 1: % de voyelles correctement reconnues

	A	P
A	95	3
P	5	97

	O	F
O	99	4
F	1	96

	OR	N
OR	82	36
N	18	64

	B	D
B	43	11
D	57	89

Table 2: matrices de confusions entre classes, seul le premier candidat est pris en compte.
axe horizontal: classe reconnue
A: antérieur, P: postérieur, O: ouvert, F: fermé, OR: oral, N: nasal, B: bémolisé, D: diésé

Les résultats (cf table 2) montrent qu' il apparaît un petit nombre de confusions entre les voyelles ouvertes et les voyelles fermées et entre les voyelles antérieures et postérieures (avec les restrictions énumérées ci-dessus).

Pour résumer, le module de reconnaissance des voyelles développé par Rossi a montré son aptitude à classifier les voyelles en grandes classes. Des règles dépendantes du contexte et une adaptation au locuteur semblent nécessaires pour une reconnaissance plus fine des voyelles.

D' après ces résultats, nous suggérons un processus de reconnaissance qui serait:
_ relativement indépendant du locuteur et du contexte pour la reconnaissance des traits les plus robustes. Par relativement, nous voulons dire que nous n' excluons pas à-priori une séparation des locuteurs en grandes classes en fonction de leur sexe ou de leur âge, ou d' introduire dans certains cas des règles contextuelles.
- dépendant du locuteur et du contexte pour une reconnaissance fine des voyelles.

Nous travaillons actuellement sur le module de reconnaissance des traits. Nous présentons ci-dessous notre méthodologie ainsi que nos premiers résultats.

II RECONNAISSANCE DES TRAITS

La méthodologie est la même que celle adoptée précédemment à deux exceptions près qui prennent en considération nos remarques précédentes (cf I.2):
- les indices sont uniquement évalués sur un simple échantillon au centre de la voyelle, ceci afin de minimiser l' influence du contexte et celle d' une mauvaise délimitation des frontières de la voyelle.
- les indices ne sont plus binaires.

Le corpus d' apprentissage est composé de 160 logatomes de type CVCVCVC, où les trois voyelles et les trois consonnes sont identiques et où V représente une des treize voyelles françaises et C une des 16 consonnes françaises. La segmentation automatique a été manuellement vérifiée. Deux locuteurs ont enregistré le corpus, un homme et une femme.

Nous avons testé une douzaine d' indices comme corrélats acoustiques de chacun des traits ouvert/fermé et aigu/grave. Des tests statistiques ont été utilisés pour sélectionner les plus discriminants. Chaque candidat est donné avec un taux de confiance qui tombe entre 0 et 1, selon la valeur de l' indice.

Comme nous l 'avons indiqué précédemment, le degré d' ouverture des voyelles moyennes n' est pas déterminé à-priori. Par conséquent ces voyelles ne sont pas prises en compte lors des tests sur le pouvoir discriminant des indices d' ouverture. Pendant la phase de reconnaissance, ces voyelles moyennes seront classées automatiquement par le programme comme ouvertes ou fermées selon que les valeurs de l' indice d' ouverture recoupent celles des voyelles ouvertes ou fermées, respectivement. On dispose par cette méthode d' un critère objectif pour distinguer entre les allophones. La même stratégie est adoptée pour /a,ɛ̃,oe,ø/ non classées a-priori sur l' axe antéro-postérieur. Nous allons présenter successivement les résultats obtenus sur le corpus d' apprentissage puis sur un corpus de test.

I.2 Resultats sur un corpus d' apprentissage:

La figure 2 montre les histogrammes des corrélats acoustiques des deux traits étudiés: les indices "ouvert1" et "aigu1". Chaque trait peut être identifié au moyen d' un unique indice avec un taux d'erreur inférieur ou égal à 3%. Ce taux peut être ramené à 1% en affinant la reconnaissance du trait antérieur/postérieur au moyen des deux critères suivants:
- adapter certaines zones fréquentielles au sexe du locuteur: pratiquement élever la bande de fréquence testée par l' indice "aigu1" pour les locutrices.
- ajouter une règle pour une meilleure distinction des voyelles /i/ et /u/. Ces voyelles possèdent parfois un deuxième formant de faible intensité (/i/ et /u/) ou très bas (/u/) et l' indice "aigu1" n' est alors plus bien adapté à la reconnaissance de leur lieu d' articulation. Un deuxième indice, tenant compte de la spécificité de ces voyelles, permet l' élimination d ' un grand nombre d' incertitudes ou d' erreurs les concernant .Afin de ne pas dégrader le score obtenu grace au premier indice "aigu1", seules les valeurs de ce second indice permettant une identificaion certaine du trait sont utilisées.

Pour résumer, trois indices sont suffisants pour identifier les traits ouvert/fermé et antérieur/postérieur avec un taux d' erreur inférieur ou égal à 1% sur le corpus de test: un indice d' ouverture, un indice d' antériorité, adapté au sexe du locuteur, et un autre indice d' antériorité utilisé dans les cas de certitude.

II.2 Premiers resultats sur le corpus de test

C' est un corpus de nombres, les tests ont porté sur sept locuteurs masculins et sept locutrices. Nous proposons d' évaluer les performances de notre ensemble de règles par deux critères:
-le taux d' erreur obtenu quand le candidat le plus probable est pris en compte (taux de confiance supérieur à 0.5).
- le taux d' erreur obtenu quand le taux de confiance est maximal (c.a.d 1). Nous indiquons avec ce taux d' erreur le % de voyelles pour lesquelles ce taux est obtenu. Il est, évidemment, important d' obtenir le plus grand nombre possible de voyelles avec un taux de confiance maximal.

Les résultats confirment ceux obtenus sur le corpus d' apprentissage (cf table 3). Les erreurs obtenues sur le trait d' ouverture concernent la voyelle /u/ dans le nombre 12 (/duz/). La même voyelle est également responsable du nombre plutôt faible de candidats donnés avec un taux de confiance maximal. Nous cherchons actuellement des solutions simples pour résoudre le problème posé par /u/. Les grandes variations articulatoires de cette voyelle dans le contexte dental ont déjà été signalées en français et dans d' autres langues [5]. Les résultats satisfaisants obtenus pour le trait antérieur/postérieur nous laissent espérer qu 'il se révèlera encore efficace sur d' autres corpus plus vastes et pour un plus grand nombre de locuteurs.

TRAIT	% D' ERREUR		NOMBRES DE CANDIDATS AVEC T=1
	T>.5	T=1	
ANTERIEUR-POSTERIEUR	1	0	80
OUVERT-FERME	1	0	40

Table 3: % d' erreurs pour les traits
T: taux de confiance

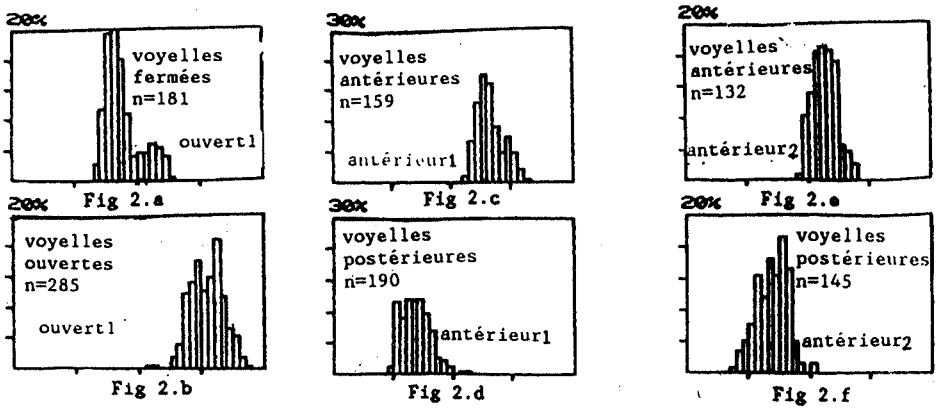


Figure 2: Histogrammes des valeurs des indices acoustiques pour chaque trait, calculés d' après le corpus d' apprentissage.
n: nombre de voyelles
ouvert1: corrélé acoustique du trait ouvert/fermé
antérieur1:corrélé acoustique du trait antérieur/postérieur
antérieur2:corrélé acoustique du trait antérieur/postérieur, l' équivalent de antérieur1 mais portant sur région fréquentielle légérement supérieure, mieux adaptée aux voix féminines.
ouvert1 calcule la différence entre EK1 et (EK3+EK4), EKi représente le niveau d' énergie dans le ième canal. antérieur1 calcule la difference entre (EK6 + EK7) et (EK4 + EK5).
Fig 2.a, Fig 2.b: deux locuteurs (h et f)
Fig 2.c, Fig 2.d: 1 locuteur masculin
Fig 2.e, Fig 2.f: 1 locutrice

CONCLUSION

Nous avons proposé dans ce module une stratégie mixte pour la reconnaissance des voyelles françaises: relativement indépendante du locuteur et du contexte pour la reconnaissance des traits ouvert/fermé et antérieur/postérieur, dépendante du locuteur et du contexte pour la reconnaissance fine des voyelles. Le but du module de reconnaissance des traits est de réaliser une première classification fiable des voyelles, pour l' accès au lexique, par exemple, ou d ' être connecté à un module de reconnaissance des voyelles. Sur un corpus de nombres, prononcés par 14 locuteurs, nous avons obtenu un taux d' erreur de 1% pour l' identification des traits ouvert/fermé et antérieur/postérieur et aucune erreur n' est faite sur les réponses données avec un taux de confiance maximal.

REFERENCES

[1] D.W. Shipman, V.W. Zue, "Properties of large lexicons. Implications for advanced isolated word recognition systems", Proc. IEEE ICASSP, Paris, 1982.

[2] G. Adda, M. Eskenazi, P.E Stern, "Reconnaissance de grands vocabulaires: utilisation et évaluation de traits grossiers" Journées d' étude sur la Parole, Aix-en-Provence, 1986.

[3] A. Bonneau, M. Mercier, M. Gerard, M. Rossi, "Décodage acoustico-phonétique à l' aide du système expert Serac-Iroise" Journées d' Etude sur la Parole, Aix-en-Provence, 1986.

[4] D.H. Klatt, "Models of phonetic recognition I: issues that arise in attempting to specify a feature-based strategy for speech-recognition", Proc. Montreal Symposium on Speech Recognition, 1986.

[5] K. Shirai, T. Kobayashi, J. Yazawa, "Estimation of articulatory parameters by table-look method and its application for speaker independent phoneme recognition". Proc. Icassp, 1986.

A PROPOS DES "UNIVERSAUX" DE LA NASALISATION

JEAN-MARIE HOMBERT

LAPHOLIA (Université Lyon 2) et LACITO (CNRS)

RESUME

L'étude des changements phonétiques peut contribuer à améliorer notre compréhension des mécanismes de production et de perception de la parole. Nous nous intéressons ici au phénomène de la nasalisation. Après avoir passé en revue les données généralement citées pour illustrer ce phénomène (chinois, français), nous insistons sur le fait que les "universaux" de la nasalisation reposent sur des données restreintes et discutables. Il nous paraît donc important d'étendre notre investigation à d'autre cas - en particulier à des zones linguistiques où le processus de nasalisation est en train de se développer. Nous présentons ici le cas des langues teke. Cette étude confirme le rôle de certains facteurs (e.g. aperture vocalique), remet en cause l'universalité d'autres facteurs (e.g. lieu d'articulation de la consonne nasale) et enfin propose d'ajouter la quantité vocalique aux facteurs pouvant jouer un rôle dans la chronologie du développement des voyelles nasales.

1. Changements phonétiques et modèles de production/ perception de la parole

L'étude systématique des changements phonétiques peut être un moyen d'améliorer notre compréhension des mécanismes d'encodage et de décodage de la parole. Nous ne nous intéressons ici qu'aux types d'évolutions phonétiques attestées dans des langues génétiquement distinctes et géographiquement distantes. Nous pensons en effet que lorsque des changements similaires ne peuvent êtres attribués ni à un héritage commun (appartenance à une même famille linguistique) ni à des phénomènes d'influence (contiguité géographique), leur explication est à chercher dans les mécanismes d'encodage et de décodage, seuls points communs entre des langues non apparentées et n'ayant pas été en contact[1].

L'origine de tels changements phonétiques réside soit dans les limitations de notre système articulatoire (e.g. problème de coarticulation résultant de l'inertie des articulateurs mobiles) soit dans les limitations de notre système perceptuel (e.g. problème du décodage des effets acoustiques produits par ces coarticulations). En particulier, le locuteur recevant un message doit, pour pouvoir le décoder, être en mesure de déterminer si les paramètres acoustiques qu'il détecte appartiennent à un segment donné ou sont dus à l'influence d'un segment pré ou postposé[2].

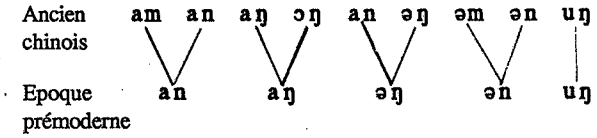
Pour étudier un changement phonétique, deux approches sont possibles. La première consiste à examiner dans des textes d'une même langue, mais datant de différentes époques, l'évolution des représentations graphiques des mots. La seconde approche prend en compte un ensemble de parlers actuels issus d'une même langue originelle et linguistiquement aussi proches que possible, et tente de reconstituer les différentes étapes de chacun des changements phonétiques les ayant affectés depuis le début de leur diversification. Naturellement, certains parlers pourront être conservateurs , i.e. conserver les archaïsmes, pour certaines évolutions et pas pour d'autres. La première approche est généralement utilisée pour les langues ayant un système d'écriture ancien et la seconde pour les langues dites "à tradition orale".

2. Tendances universelles de la nasalisation

Nous allons examiner ici le processus de formation des voyelles nasales. L'étude de la chronologie des différentes étapes de ce processus a fait l'objet de nombreux travaux[3]. La quasi-totalité de ces études reposent essentiellement sur deux groupes linguistiques : les parlers chinois et les langues romanes, en particulier le français. Résumons ces deux cas.

2.1. Nasalisation en chinois

Selon Chen(1972,1975), la nasalisation s'est développée dans les parlers chinois par un processus d'assimilation régressive[4], i.e. une consonne nasale finale a progressivement nasalisé la voyelle qui la précédait puis s'est amuïe (VN > V̲N > V̲). Cette nasalisation n'a commencé qu'à l'époque prémoderne alors que la fusion d'un certain nombre de finales nasales de l'Ancien chinois avait déjà eu lieu.



La nasalisation de ces cinq séquences VN a eu lieu dans l'ordre suivant : an, aŋ, ən, əŋ, uŋ. De cette chronologie, Chen conclut que deux facteurs phonétiques ont conditionné la nasalisation :

a. l'aperture de la voyelle : la voyelle ouverte [a] se nasalise avant la voyelle d'aperture moyenne [ə], qui elle-même se nasalise avant la voyelle fermée [u].

b. le lieu d'articulation de la consonne nasale : la consonne dentale nasalise avant la consonne vélaire. A noter que l'on ne peut rien conclure à propos de la nasale labiale puisque celle-ci avait déjà disparue en fusionnant avec la dentale (**am > an** et **əm > ən**).

Remarquons également que le premier facteur domine le second puisque **aŋ** est nasalisé avant **ən**. En ce qui concerne l'évolution du timbre des voyelles nasalisées, Chen note que la tendance prédominante est la fermeture du timbre bien qu'il reconnaisse que le mouvement inverse (ouverture du timbre) est aussi attesté.

2.2. Nasalisation en français

Comme pour le chinois, la nasalisation s'est développée à partir de l'influence d'une consonne nasale sur la voyelle qui la précédait (i.e. assimilation régressive). La chronologie généralement admise[5] montre que les voyelles ont été successivement touchées par cette nasalisation[6] :

a	au Xème siècle
e	à la fin du Xème siècle, début du XIème
o	au XIIème siècle
i	au XIIIème siècle
y	à la fin du XIIIème et peut-être au XIVème siècle.

Les voyelles[7] semblent avoir été affectées plus ou moins tôt en fonction de leur degré d'aperture (les plus ouvertes d'abord) ; en outre, à degré d'aperture égal, les voyelles les plus antérieures ont été nasalisées plus tôt (**e** avant **o** et **i** avant **y**). Contrairement au chinois, le lieu d'articulation de la consonne nasale postvocalique n'a pas joué un rôle déterminant dans la chronologie de la nasalisation des voyelles du français. Enfin, à la suite de la nasalisation, le timbre des voyelles s'est généralement ouvert : **e̯ > a̯**, **o̯ > ɔ̯**, **i̯ > ɛ̯** et **y̯ > œ̯**.

2.3. Y-a-t-il des tendances universelles à la nasalisation ?

A partir de ces deux cas qui, rappelons-le, sont à la base des généralisations émises à propos du développement de la nasalisation, trois remarques s'imposent :

a. Tenter de dégager des généralisations à partir d'un nombre d'exemples aussi réduit est pour le moins prématuré.

b. Certains facteurs jouant un rôle dans l'un des exemples (rôle du lieu d'articulation de la consonne nasale en chinois et de la position antérieure/postérieure en français) semblent inopérant dans l'autre.

c. Les analyses que nous avons résumées ci-dessus ne font pas l'unanimité chez les spécialistes de la phonétique historique. Ainsi, pour le français, Haudricourt(1947) attribue à une opposition de quantité vocalique ce qui a été interprété par d'autres comme une opposition de nasalité. Rochet(1974), quant à lui, insiste sur le rôle moteur de la morphologie (par opposition à un conditionnement phonétique) dans le développement des voyelles nasales du français.

Face à de telles réserves, il nous semble évident que la tâche la plus urgente est d'élargir notre base de données, c'est-à-dire d'examiner d'autres cas de développement de la nasalisation, afin de pouvoir établir s'il y a véritablement un conditionnement phonétique de ce processus.

3. Nasalisation en teke

Une très large majorité des langues du monde n'a pas de système d'écriture. On ne peut donc, dans ces langues, étudier l'évolution des changements phonétiques en comparant des textes de différentes époques. Il est toutefois possible d'obtenir des informations précieuses sur les différentes étapes de l'évolution diachronique des langues dites "à tradition orale". Pour cela, il suffit d'examiner l'état synchronique d'un grand nombre de parlers linguistiquement aussi proches que possible. Chaque parler a eu son évolution propre, c'est-à-dire que, par rapport à certains changements phonétiques attestés dans d'autres parlers proches, il a été conservateur (i.e. a préservé la forme archaïque) alors que pour d'autres changements phonétiques il a pu être novateur et évoluer rapidement. Pour une évolution phonétique donnée, l'ensemble des états synchroniques des parlers actuels peut être considéré comme autant d'étapes qui ont conduit de la forme originelle (présente dans la langue mère) à la forme attestée dans le parler le plus novateur. C'est dans cette optique que nous présentons maintenant l'évolution de la nasalisation dans trois parlers Teke du Congo et du Gabon.

Le tableau 1 récapitule les correspondances entre les structures $C_1 V_1 mV_2$ (cinq premières lignes) et $C_1 V_1 mb V_2$ (cinq dernières lignes) de la langue mère (proto-teke), et les reflexes attestés dans trois parlers actuels : atege, ibali et ndzindziu[8]. La colonne correspondant à l'atege ne comporte pas de formes nasalisées mais elle permet de constater que le passage de la mi-nasale **[mb]** à la nasale **[m]** a provoqué un allongement compensatoire et parfois une diphtongaison de V_1. En ibali la nasalisation n'a eu lieu que lorsque la voyelle qui précédait la consonne nasale était longue et non fermée. En ndzindziu par contre, elle s'est généralisée aux voyelles fermées et aux brèves. Nous n'avons considéré ici que la nasale labiale ; la nasale dentale n'a pas encore nasalisé dans ces trois parlers[9] (ex. *commencer* **báánà** en atege, ibali et ndzindziu - et non pas **baa̯** ou **ba ɔ̯**). Quant à la nasale vélaire, elle s'est amuïe sans laisser de traces de nasalisation (ex. *genou* **bóŋgó** en proto-teke mais **bʊ́ʊ́** en atege, ibali et ndzindziu).

En résumé, les paramètres qui conditionnent le développement de la nasalisation dans la zone teke sont :

- le lieu d'articulation de la consonne nasale : la labiale vocalise avant la dentale (mais **ŋ** est tombé sans nasaliser) ;
- la quantité vocalique : les voyelles longues sont nasalisées avant les voyelles brèves ;
- l'aperture de la voyelle : les non-fermées sont touchées plus tôt que les fermées.

Nous constatons, en outre, que le timbre de la voyelle nasalisée est fortement influencé par l'articulation labiale de la consonne. Remarquons enfin que ces données ne nous permettent pas de préciser si la position (antérieure ou postérieure) de la voyelle favorise ou non sa nasalisation.

Tableau 1. Correspondances entre les formes C1 V1 m(b) V2 du proto-teke et leurs reflexes en atege, ibali et ndzindziu

PROTO-TEKE	ATEGE	IBALI	NDZINDZIU
C_1 i m V_2	C_1 i m V_2	C_1 i m V_2	C_1 i̯ọ
C_1 e m V_2	C_1 e m V_2	C_1 e m V_2	
C_1 a m V_2	C_1 a m V_2	C_1 a m V_2	C_1 ọ
C_1 o m V_2	C_1 o m V_2	C_1 o m V_2	
C_1 u m V_2	C_1 u m V_2	C_1 u m V_2	C_1 oọ , C_1 ọ
C_1 i mb V_2	C_1 ii m V_2	C_1 ii m V_2	C_1 i ọ
C_1 e mb V_2	C_1 ie m V_2	C_1 iọ	
C_1 a mb V_2	C_1 aa m V_2	C_1 a̯a̯	C_1 aọ
C_1 o mb V_2	C_1 uo m V_2	C_1 uọ	C_1 uọ
C_1 u mb V_2	C_1 uu m V_2	C_1 uu m V_2	C_1 uọ , C_1 uọ

Si nous comparons ces résultats avec ceux obtenus pour le chinois et le français, nous voyons que le rôle de certains facteurs est confirmé (e.g.le rôle de l'aperture de la voyelle) alors que d'autres sont remis en cause (e.g.le rôle du lieu d'articulation de la consonne). L'exemple teke nous permet d'introduire un facteur qui jusqu'alors n'avait pas, à notre connaissance, été mentionné comme jouant un rôle dans la chronologie du développement des voyelles nasales : la quantité vocalique. Notons enfin un fait probablement lié au rôle de la quantité vocalique dans l'évolution de la nasalisation : les voyelles ouvertes qui, comme nous l'avons vu, sont les premières à se nasaliser sont aussi phonétiquement plus longues que les voyelles de plus faible aperture.

4. Explications phonétiques

L'exemple du teke montre bien comment une étude détaillée du développement de la nasalisation peut remettre en cause les "généralisations" existantes. Il est donc important de collecter des données provenant d'autres zones linguistiques afin de faire apparaître clairement les facteurs phonétiques qui conditionnent le développement de la nasalisation. Pour l'instant, seule l'aperture de la voyelle a joué un rôle dans les trois cas présentés ici, or il a été montré à plusieurs reprises[10] qu'il y avait une corrélation entre la position du voile du palais et l'aperture de la voyelle : plus la voyelle est ouverte, moins le voile du palais est relevé. Ceci implique qu'une fuite nasale, et par conséquent une nasalisation involontaire de la voyelle, pourra se produire plus facilement lors de la réalisation d'une voyelle de grande aperture.

Terminons en examinant si l'évolution du timbre des voyelles lors de leur nasalisation peut-être expliqué par des considérations phonétiques. Beddor(1983) résume les positions de nombreux auteurs quant aux effets de la nasalisation sur le timbre vocalique ; certains indiquent que la nasalisation ferment la voyelle alors que d'autres prétendent le contraire. En réalité, ces deux positions ne sont pas aussi contradictoires qu'elles peuvent le paraître à priori. Comme le montre Beddor, la nasalisation n'affecte pas de la même manière les voyelles fermées et les voyelles ouvertes ; on constate en fait une tendance à la centralisation : les voyelles fermées s'ouvrent alors que les voyelles de grande aperture se ferment. En outre, il semble qu'il y ait une influence différente avant et après la perte de la consonne nasale qui conditionne le développement de la nasalisation : la voyelle a tendance à se fermer lorsque la nasale postvocalique est encore présente, et s'ouvre au contraire lorsque la consonne s'amuït. Kawasaki (1986) présente les résultats d'une expérience perceptuelle qui peut permettre de comprendre comment on peut aboutir à ces évolutions apparemment divergentes. Elle montre qu'un abaissement progressif de l'amplitude des consonnes nasales dans les séquences mVm - sans modification de la voyelle - est interprété par certains sujets comme une augmentation de la nasalisation de la voyelle interconsonantique alors que, pour d'autres sujets, cet abaissement de l'amplitude des nasales est interprété comme une diminution de la nasalisation de la voyelle.

Ceci illustre bien que des locuteurs d'une même communauté linguistique peuvent avoir des stratégies perceptuelles très différentes[11] qui peuvent les amener à opérer des décodages divergents, en particulier lorsqu'il s'agit de séquences fortement coarticulées et que l'un des segments est en train de s'amuïr.

5. Conclusion

Nous espérons avoir montré que les données sur lesquelles reposent les "universaux" diachroniques de la nasalisation étaient trop restreintes, et qu'il était urgent de collecter des données supplémentaires afin de voir émerger de véritables tendances. Quant aux explications phonétiques associées à ces tendances, nous pensons qu'elles passeront par un modèle de codage/décodage de la parole prenant en compte les spécificités articulatoires et conceptuelles des locuteurs d'une même communauté linguistique.

NOTES

1 Pour le traitement d'un changement phonétique de ce type, voir Hombert, Ohala et Ewan(1978).

2 Pour une présentation plus complète de l'origine des changements phonétiques, voir Hombert(1984).

3 Pour une synthèse de ces travaux, voir Ferguson et al.(1975) et Ruhlen(1978).

4 Le développement de voyelles nasales par assimilation progressive n'est attesté que dans quelques rares exceptions (parlers Min du Sud).

5 Voir par exemple Straka(1955), Fouché(1969).

6 Les dates de ces étapes de la nasalisation pendant la période de l'ancien français peuvent évidemment varier suivant les régions.

7 Par souci de clarté, nous n'avons pas présenté ici l'évolution des diphtongues.

8 Pour une présentation plus complète de la nasalisation dans les parlers teke, voir Hombert(1986) et Hombert et Paulian(1987).

9 Mais elle a nasalisé dans d'autres parlers voisins tel que le ngungwel (voir références de la note précédente).

10 Voir par exemple Bell-Berti et al.(1979).

11 Voir Hombert(1984).

REFERENCES

P.S. BEDDOR, "Phonological and phonetic effects of nasalization on vowel height", Reproduced by the Indiana University Linguistics Club, Bloomington, Indiana, 1983.

F. BELL-BERTI, T. BAER, K.S. HARRIS et S. NIIMI, "Coarticulatory effects of vowel quality on velar function", Phonetica 36, pp. 187-193, 1979.

M. CHEN, "Nasals and nasalization in Chinese : explorations in phonological universals", Ph.D., University of California, Berkeley, 1972.

M. CHEN, "An areal study of nasalization in Chinese", in: C.A. FERGUSON et al., "Nasálfest", pp. 81-124, 1975.

C.A. FERGUSON, L.M. HYMAN and J.J. OHALA (eds.), "Nasálfest: papers from a symposium on nasals and nasalization", Stanford University: Language Universals Project, 1975.

P. FOUCHE, "Phonétique historique du français", Klincksieck, 1969.

A.G. HAUDRICOURT, "EN/AN en français", Word 3, pp.39-47, 1947.

J.M. HOMBERT, "Reflexion sur le mécanisme des changements phonétiques", Pholia 1, LAPHOLIA-CRLS-Univ. Lyon2, pp. 87-112, 1984.

J.M. HOMBERT, "The development of nasalized vowels in the Teke language group (Bantu)", in: K. BOGERS, H. van der HULST and M. MOUS (eds.) "The Phonological representation of suprasegmentals", Foris Publications, pp. 359-373, 1986.

J.M. HOMBERT, J.J. OHALA and W.G. EWAN, "Phonetic explanations for the development of tones", Language 55,1, pp. 37-58, 1979.

J.M. HOMBERT, M. MAZAUDON, B. MICHAILOVSKY, F. OZANNE-RIVIERRE, A. RIALLAND, J.C. RIVIERRE et L. SAGGART, "Universals of nasalization revisited", Communication présentée à la Conférence Internationale de Linguistique Historique, Lille, Septembre 1987.

J.M. HOMBERT et C. PAULIAN, "Spreading of nasalization in the Teke area", Communication présentée au Colloque International de Linguistique Africaine, Leiden, Pays-Bas, Septembre 1987.

H. KAWASAKI, "Phonetic explanation for phonological universals : The case of distinctive vowel nasalization", in J.J. OHALA and J.J. JAEGER (eds), "Experimental Phonology", Academic Press, pp. 81-103, 1986.

J.J. OHALA, "Phonetic explanations for nasal sound patterns", in FERGUSON et al.(eds) "Nasálfest", pp. 289-316, 1975.

B. ROCHET "About a pseudo-linguistic universal : that nasal vowel have a tendency to lower", in L. HEILMANN (ed), "Proceedings of the 11th International Congress of Linguistics, Bologna, Societa Editrice Il Mulino, pp. 727-730, 1974.

M. RUHLEN, "Nasal vowels", in: J.H. GREENBERG (ed.), "Universals of Human Language" vol. 2, Stanford University Press, pp.203-242, 1978.

G. STRAKA, "Remarques sur les voyelles nasales et leur évolution en français", Revue de Linguistique Romane 19, pp. 245-274, 1955

Extraction automatique de caracteristiques dynamiques du signal de parole. Application à l'analyse des voyelles nasales.

P. F. Marteau, J. Caelen & M.T. Janot-Giorgetti

Laboratoire de la Communication Parlée - ICP, unité associée au CNRS
INPG/ENSERG 46, Av. F. Viallet 38031 Grenoble Cedex

ABSTRACT

We propose a model based on a factorial analysis to describe phonetic transitions. This model is used to segment the acoustic signal and to characterize the transitions as the motion of some masses of energy in both time and frequency domains. We have tested such a model within the framework of a study of the French nasal vowels. We show that this model allow us to relocate roughly the results of articulatory and psychoacoustic analysis. We propose and discuss then the existence of dynamic acoustical cues for the phonetic feature of nasality.

1-INTRODUCTION

Le signal de parole est un signal temporel non stationnaire qui posséde une structure fréquentielle riche. L'observation de ce signal à travers des outils tels que la représentation spectrographique (basée sur la transformée de Fourier à court terme) a permis d'analyser finement sa texture dans les deux dimensions et par suite d'établir sous forme d'indices une sorte de connaissance qui permet à certains experts de décoder visuellement l'information phonétique transportée par le signal.

Sur de telles représentations, l'oeil de l'expert agit, pour une part tout au moins comme un filtre, ne s'intéressant qu'aux zones les plus significatives, telles que les transitions de formants, l'élargissement de leurs bandes passantes, ou plus généralement, l'apparition d'énergie dans certaines régions du spectre.

Cette faculté de focaliser sur ce qui se déforme et varie effectivement est difficile à modéliser et à transcrire directement en langage informatique. Cette transcription nécessite une étape de simplification afin de réduire le nombre des variables significatives et une étape d'interprétation des nouvelles variables en termes de mouvement de "masses" d'énergie le long des axes temporel et fréquentiel.

Par des techniques d'analyse factorielle, nous proposons une segmentation automatique du signal et sur chaque segment, nous retrouvons les principales masses d'énergie qui ont transité.

L'application de ce modèle de déformation spectrale à l'analyse des voyelles nasales du Français nous permet de mettre en évidence certains paramètres caractéristiques de la nasalité.

2-LE MODELE

Le signal acoustique est transposé dans l'espace des fréquences, à chaque instant k tel que $t_k = k.dt$ (où dt=5ms), par le calcul d'un spectre LPC (14 coefficients), redéfini sur (m=22) bandes fréquentielles de largeur 1 Bark. Considérons une fenêtre temporelle W englobant les instants (1...q).

Notons J={1...q} l'ensemble qui indexe les spectres --considérés comme les individus-- associés à la fenêtre w.

Notons I={1...m} l'ensemble qui indexe les bandes fréquentielles --considérées comme les variables-- qui décrivent les individus. Sur W, le signal est alors représenté par le tableau individus/variables à valeurs réelles :

$\{ S_i^j \mid i \in I; j \in J \}$ où S_i^j est l'énergie de la $i^{ème}$ bande pour le $j^{ème}$ spectre.

Soient $E=R^I$ l'espace vectoriel des fonctions discrètes à valeurs réelles (fonctions fréquentielles) :

$E=\{ f \mid f^i \in R; i \in I \}$

et $G=R^J$ l'espace vectoriel des fonctions discrètes à valeurs réelles (fonctions temporelles) :

$G=\{ g \mid g^j \in R; j \in J \}$

Il y a à ce niveau deux approches duales pour traiter le tableau (S_i^j) selon que l'on considère le nuage C(I) des fonctions S^i dans l'espace G ou le nuage C(J) des fonctions S^j dans l'espace E : C(J) peut être en effet interprété comme une trajectoire temporelle dans l'espace E, tandis que C(I) peut être interprété comme une trajectoire fréquentielle dans l'espace G.

Décrire ces nuages revient à déterminer leurs principales caractéristiques d'inertie, i.e. leurs axes principaux d'inertie. C'est le rôle de l'analyse factorielle (BENZECRI 1973) ou des techniques de décomposition de matrice en valeurs singulières.

Ce type d'analyse appliquée au signal de parole est classique tant pour le traitement de l'information statique (CARTIER & GRAILLOT 1974; CAELEN & VIGOUROUX 1983) que pour le traitement de l'information dynamique (T. MOUSSA 1982; H.KOBATAKE, S. OHTANI 1987, ATAL 1983, MARCUS S. M. 1984, AHLBOM & al. 1987...)

Notre but est de décrire localement la distribution d'énergie dans les deux dimensions temps-fréquence; Cela impose de traiter des spectres issus d'un même segment temporel. La théorie de l'analyse factorielle nous indique que les caractéristiques d'inertie des deux nuages sont duales et ainsi, par la mise en correspondance des deux espaces E et G, nous pouvons interpréter simultanément les transitions du signal comme le déplacement temporel et fréquentiel de certaines masses d'énergie.

Considérons le nuage C(J) dans l'espace E, i.e. la trajectoire temporelle spectrale dans l'espace des bandes fréquentielles.

Soit $\{ \mu^j \mid j \in J \}$ un ensemble de nombres réels positif : μ^j est le poids affecté au $j^{ème}$spectre. En pratique, nous n'utilisons que les pondérations uniformes (fenêtre rectangulaire) ou de hamming (pour réduire les effets de discontinuité aux bornes de la fenêtre).

Notons $\bar{\mu}$ la somme totale des poids.

Si l'on munit l'espace E de la distance euclidienne (d), (E,d) est un espace métrique. -Le choix de la distance euclidienne est justifiée par le caractère homogène des variables (les bandes) qui décrivent les individus (les spectres)- La distance entre deux spectres est alors donnée par :

$$d(j,j')^2 = \sum_{i \in I} (S_i^j - S_i^{j'})^2$$

Notons G le centre de gravité des fonctions S^j de composantes $\{ G_i : i \in I \}$.

$$G_i = \sum_{j \in J} \left(\frac{\mu^j}{\bar{\mu}} S_i^j \right)$$

Soit σ la matrice carrée symétrique représentant la forme quadratique d'inertie associée au nuage C(J) :

$\{ \sigma_{ii'} \mid i \in I; i' \in I \}$

$$\sigma_{ii'} = \sum_{j \in J} \left(\frac{\mu^j}{\bar{\mu}} S_i^j . S_{i'}^j \right) - G_i G_{i'} \mu$$

Soit U l'ensemble des vecteurs propres orthonormés associé à l'application σ : $U=\{ u_i \mid i \in I ; \sigma(u_i)=u_i \}$

Enfin, soit V l'ensemble des composantes principales associées à ces vecteurs : $V=\{v^j \mid j \in J \}$. V^j est la projection de l'individu S^j sur le vecteur u :

$$v^j = \sum_{i \in I} u_i.(S_i^j - G_i)$$

Le tableau S se décompose sous la forme :

$$(S_i^j) \text{ avec : } S_i^j = G_i + \sum_{u \in U;\, v \in V} v^j.u_i$$

Alors, si nous considérons le sous espace U(p) engendré par les p<m premiers vecteurs propres (associés aux p premières valeurs propres ordonnées), nous obtenons une estimation à l'ordre p du tableau S donnée par $\hat{S}=(\hat{S}_i)$ avec :

$$\hat{S}_i^j = G_i + \sum_{h=1}^{p} v^j.u_i \quad (1)$$

où $v^j(h)$ désigne la h^{eme} composante principale associée au h^{eme} vecteur propre. En effet, l'espace {G,U(p)} est le sous espace de dimension p qui minimise le critère.

$$\delta(p) = \sum_{i \in I;\, j \in J} (S_i^j - \hat{S}_i^j)^2$$

où $\delta(p)$ n'est autre que la somme des distances entre les spectres et leur projection dans le sous espace {G,U(p)}.

Les formes u(h) peuvent être interprétée comme des fonctions de masquage fréquentiel qui ont la dimension des indices acoustiques spectraux définis par ROSSI et CAELEN. (ROSSI & al. 1983; CAELEN & CAELEN-HAUMONT 1981). Les fonction v(h) caractérisent l'évolution temporelle .

Soit l'exemple suivant :

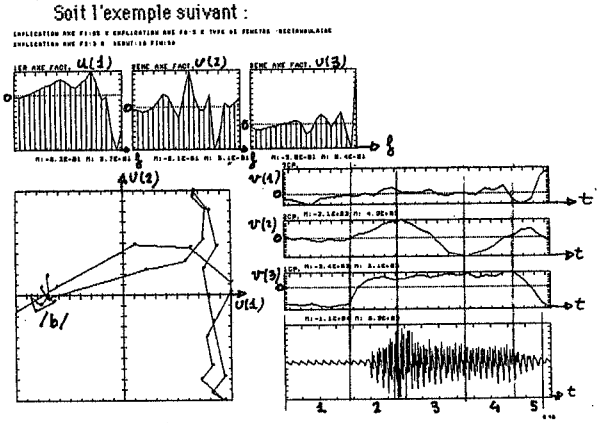


FIG 1 : Représentation graphique des composantes principales et des axes principaux sur la transition /b/-/ã/ et trajectoire des spectres dans le plan des deux axes principaux. On peut observer la présence d'une cible atteinte qui caractérise la présence du /b/, et deux points de rebroussement qui caractérisent deux cibles non atteintes, l'une correspondant à une voyelle quasi orale, l'autre à la cible naso-pharyngale. (CF § 4)

D'après l'équation (1), le tableau S, qui est une fonction des deux variables temps et fréquence, peut être approximé par le tableau $\hat{S}$, où les deux variables sont séparées. L'interprétation du modèle peut se faire par observation simultanée des couples (u,v).L'analyse du diphone /b/-/ã/ dans le plan des deux premiers axes factoriels est reportée sur la Fig. 1. La segmentation à partir des composantes temporelles est ici manuelle. Nous retrouvons les masses d'énergie qui se déplacent sur chaque segment temporel en observant les axes principaux (les composantes fréquentielles). Ainsi, sur le segment n° 3 qui correspond à la nasalisation de la voyelle, l'énergie passe des maxima de u(2) aux minima de u(2) i.e. des canaux --[540-650Hz], [1130-1340Hz], [2060-2350Hz].... -- aux canaux --[260-320Hz], [880-1000Hz] ...-- (u(1) n'intervenant pas dans l'explication de cette transition).

C'est ce type d'observation que nous voulons réaliser automatiquement. Le module d'interprétation que nous décrivons au paragraphe suivant est écrit en Prolog II.

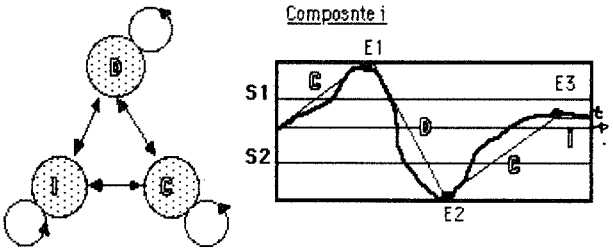
3-LE MODULE D'INTERPRETATION AUTOMATIQUE

Pour interpréter commodément les résultats d'une telle analyse, il faut que le nombre de facteurs ne soit pas trop important . Ce nombre dépend en fait du seuil d'explication choisi.

Ici, on évalue la plus petite dimension (p) du sous espace qui "explique" au moins 80% de l'inertie du "nuage" constitué par la trajectoire. Pour une fenêtre d'analyse de 200ms, le modèle est satisfaisant, (au moins 80% de l'inertie du nuage est "expliquée"), pour un ordre inférieur à 3. Cela signifie que sur cette fenêtre d'analyse la trajectoire est quasiment incluse dans un sous espace de dimension p inférieure à 3.

Les p premières composantes principales sont alors normalisées puis squelettisées par un automate à 3 états, basé sur les seuils S1 et S2: La courbe à squelettiser est représentée par les segments reliant ses extrema successifs, puis, sur chaque segment, l'automate évalue la tendance (croissant, décroissant ou constant) Si, sur deux segments consécutifs, l'automate ne change pas d'état, les segments sont concaténer; Sinon, ils sont conservés.

D : Décroissance
C : Croissance
I : constant



A partir des n squelettes précédents étiquetés, le système découpe la trajectoire en sous-segments homogènes pour lesquels, l'automate ne change pas d'état quelle que soit la composante considérée.

La trajectoire est alors modélisée par une suite de segments :

- rectilignes dans le cas où l'une au moins des composantes varie. Ces segments sont caractérisés par un vecteur directeur que le système peut calculer. Les extréma de ce vecteur donnent une image des masses d'énergie qui transitent le long du segment : les maxima sont les masses qui apparaissent, les minima sont les masses qui disparaissent dans le temps. Le cumul des variations en valeurs absolues des p composantes est une mesure de l'instabilité spectrale locale.
- ponctuels dans le cas ou aucune des composantes ne varie. On considère que de tels segments traduisent la présence d'une cible acoustique atteinte.

```
fichier : banssi.rd1                    debut:12  fin:14 coeff-var:11  deformation
==================

debut_analyse:10 fin_analyse:40          masses
ordre du modele:2   % explication:90     bd-freq:Hz:180.215 croissant enes:11
                                         bd-freq:Hz:540.650 croissant enes:12
debut:1  fin:6 cible-atteinte            bd-freq:Hz:1130.1340 croissant enes:32
                                         bd-freq:Hz:4000.4500 croissant enes:100
debut:6  fin:7 coeff-var:9  deformation  bd-freq:Hz:4500.5000 croissant enes:69
                                          anti masses
          - - -                          bd-freq:Hz:320.380 decroissant enes:8
                                         bd-freq:Hz:880.1000 decroissant enes:43
                                         bd-freq:Hz:1790.2060 decroissant enes:32
                                         bd-freq:Hz:2350.2700 decroissant enes:78

                                         debut:14  fin:21 coeff-var:70  deformation

detection cible naso-pharyngale > deb:14 fin:21
detection d'energie BF > deb:14 fin:21 deb:26 fin:30
detection d'energie autour de 1000Hz > deb:6 fin:7 deb:7 fin:12 deb:12 fin:14
    deb:14 fin:21 deb:21 fin:26

1=<1,6,<cpp1,constant>,<cpp2,constant>>.<6,7,<cpp1,croissant>,<cpp2,constant>>.
  <7,12,<cpp1,croissant>,<cpp2,croissant>>.<12,14,<cpp1,constant>,<cpp2,
  croissant>>.<14,21,<cpp1,constant>,<cpp2,decroissant>>.<21,26,<cpp1,constant>
  ,<cpp2,croissant>>.<26,30,<cpp1,croissant>,<cpp2,croissant>>.nil
```

Fig 2 : Exemple de la transition /b/-/ã/; Sortie du module d'interprétation. Nous retrouvons les principaux évènements de la figure 1.

4-CARACTERISATION DES VOYELLES NASALES DU FRANCAIS

Au niveau acoustique, les voyelles nasales sont le résultat d'un couplage acoustique entre les cavités nasale et orale. Le couplage intervient approximativement au milieu de la cavité orale, entre les lèvres et la glotte. Les principales conséquences de ce couplage sont le déplacement du premier formant oral naturel, l'élargissement de sa bande passante, l'apparition d'un formant basse-fréquence au voisinage de 300Hz et enfin l'apparition d'une paire pôle/zéro dans la fonction de transfert. (FUJIMURA & LINDQUIS 1971, MRYATI 1976, MAEDA 1982, HAWKINS & STEVENS 1985...).

Certaines études ont tenté de mettre en évidence des corrélats dynamiques plus précis : MERMELSTEIN (1977) utilise quatre paramètres acoustiques définis comme étant la variation d'énergie relative des bandes spectrales [0-1KHz], [1-2KHz], [2-5KHz] et [0-500Hz] pour décrire les transitions spectrales. Pour MERMELSTEIN, L'évolution du centroïde de la bande [0-500Hz] est l'indice le plus pertinent pour séparer la transition ANASAL-NON-ANASAL (le modèle utilisé est un modèle statistique.)

FENG & AL. (1985) proposent le concept de cible nasopharyngale

qui est caractérisé par l'apparition de deux formants, l'un au voisinage de 300Hz, l'autre au voisinage de 1000Hz. Ce concept est validé par simulation à partir d'un modèle articulatoire simple.

CHENG & GUERIN (1987) montrent à partir de données psychoacoustiques que la perception de la nasalité peut être représentée par un modèle prenant en compte l'équilibre de deux masses, l'une centrée autour de 300Hz, l'autre autour de 1000Hz.

Ainsi, la nasalité pourait être traduite par la déformation d'un spectre origine correspondant à une voyelle orale, ou quasi-orale, vers un spectre cible caractérisé par une émergence d'énergie au voisinage des fréquence 300 et 1000Hz.

Puisque l'on peut observer une certaine concordance entre les niveaux articulatoires et perceptifs, peut on espérer retrouver au niveau acoustique des propriétés semblables? Nous allons montrer que le modèle proposé apporte quelques éléments de réponse significatifs.

5-L'ETUDE DES VOYELLES NASALES

5.1-LE CORPUS

Le corpus est constitué d'un texte de 45 mots extraits de la revue *Science et vie* , lu par dix locuteurs (5 hommes et 5 femmes). les locuteurs avaient lors de l'enregistrement la consigne de lire naturellement et de maniere compréhensible. Un appareil Radiola de type N 4420 a été utilisé pour l'enregistrement magnétique.

" 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 long vers prospèrent sur le plancher marin des zones sous-marines profondes. Des sources thermales chaudes y maintiennent une température moyenne élevée. "

Le corpus a été étiqueté manuellement par un expert phonéticien, selon les principes d'étiquetage basés sur une analyse spectrale (VIGOUROUX & CAELEN 1985). Le corpus comprend 160 voyelles nasales ({7 /ɑ̃/, 5 /ɔ̃/, 4 /ɛ̃/} x 10 locuteurs). Dans cette étude, aucune distinction n'est faite entre /ɛ̃/ et /œ̃/. La sélection des segments correspondants aux voyelles nasales se fait de manière automatique à partir des fichiers d'étiquettes.

5.2- RESULTATS

Les resultats sont obtenus de manière automatique par filtrage des masses d'énergie qui se déplacent sur chaque segment vocalique. Nous détectons la cible nasopharyngale si des masses d'énergie apparaissent simultanément au voisinage des fréquences 300Hz et 1000Hz. Les résultats présentés sont les résultats de l'analyse en composante principale sur les segments étiquetés par l'expert et correspondants à une voyelle nasale. Ce ne sont pas des pourcentage de reconnaissance. Par exemple, la 1ère ligne du tableau de la figure 2 indique que pour 75% des segments /ɑ̃/ du corpus le système à détecté la présence d'une cible nasopharyngale. Cela ne signifie pas forcément une erreur de la part du système, mais plus vraissemblablement une absence réelle de la cible. Par cette étude, nous essayons de dégager certains indices dynamiques décrivants la nasalité.

Nous notons :
- BF la présence d'une masse d'énergie basse-fréquence, bande fréquentielle [200-400Hz].
- E1000 la présence d'une masse d'énergie autour de 1000Hz, bande fréquentielle [800-1200Hz].
- NF la présence des 2 masses précédentes simultanément. (cible naso pharyngale)

masse-type / phon.	BF	E1000	NF
ɑ̃	94%	91%	75%
ɔ̃	100%	88%	76%
œ̃,ɛ̃	100%	82,5%	72,5%

FIG 3 : Détection des évènements acoustiques pour les 4 voyelles nasales tous contextes confondus.

La figure 3 montre que les résutats ne dépendent pas de la voyelle considérée. La présence simultanée des masses BF et E1000 n'est pas un critère absolu. Dans 15% des cas, on observe un équilibrage séquentiel de ces deux masses. Dans 15% des cas, La presence des deux masses n'est pas détectée.

Dans le corpus utilisé, nous répertorions 8 contextes différents i.e. 8 macro-classes phonétiques:

Fricative sourde : FNV Fricative sonore : FV
Occlusive sourde : ONV Occlusive sonore : OV
Voyelle fermée et aigue : YFA Silence : SI
Nasale : CN Liquide : CL

masse-type / contexte	BF	E1000	NF
CN/OV	100%	85%	74%
CN/FV	94%	100%	88%
YFA/SI	100%	94%	88%
OV/CL	100%	82%	64%
FV/ONV	100%	88%	64%
CL/FNV	94%	82%	58%
ONV/ONV	94%	88%	76%
SI/ONV	100%	80%	60%
SV/OV	100%	90%	70%
FNV/OV	100%	100%	90%
ONV/SI	100%	100%	90%
SI/CN	100%	40%	1%
LI/SI	100%	90%	80%
CN/ONV	100%	80%	70%

FIG 4 : Détection des évènements acoustiques pour les contextes rencontrés dans le corpus, toutes voyelles nasales confondues.

Le contexte ne modifie pas de manière évidente les résultat précédents. Le faible taux pour le contexte silence/consonne nasale, (SI/CN), demande une analyse plus systématique . (Il pourait provenir d'une segmentation ne prenant en compte que la partie orale de la voyelle nasale, compte tenu du contexte nasal qui suit.)

CONCLUSION

Nous avons proposé un modèle qui rend compte des déformations spectrales au cours du temps. Ce modèle squelettise la trajectoire et détecte les cibles et/ou les évènements acoustiques. Les portions de trajectoire entre ces évènements constituent des segments sur lesquels il est possible de déterminer les masses d'énergie qui se sont déplacées. Nous avons montré, dans le cadre d'une étude sur les voyelles nasales du Français, que l'utilisation d'un tel modèle permet de retrouver et d'utiliser des résultats provenant d'études articulatoires et psychoacoustiques. Toutefois, Les résultats présentés ne tiennent pas compte des fausses détections possibles de la cible nasopharyngale sur des phones non nasalisés. (L'étude n'est pas encore terminée; On estime à 30% ce type d'erreur). Le module d'interprétation automatique donne la possibilité de systématiser ce type d'analyse à tout type de transition phonétique, et nous pensons pouvoir utiliser ce genre d'approche non seulement en analyse de la parole (extraction d'indices spectraux dynamiques) mais aussi en segmentation et étiquetage automatique pour le décodage acoustico-phonétique.

BIBLIOGRAPHIE

AHLBOM G., BIMBOT F. & CHOLLET G. (1987),
Modeling spectral speech transitions using temporal decompositions techniques.
ICASSP DALLAS p13-16

ATAL B.S. (1983),
Efficient coding of LPC parameters by temporal decomposition.
ICASSP 2.6 p81-84.

BENZECRI J.P. (1973),
L'analyse des données.
Dunod.

BROAD D.J. & CLERMONT F. (1987),
A methodology for modeling vowel formant contour in CVC context.
J. Acoust. soc. am. ,81 (1), p655-665.

CAELEN J. & CAELEN-HAUMONT G. (1981),
Indices et propriétés dans le projet ARIAL II.
Proceedings GALF-CNRS "Processus d'encodage et de décodage phonétique ".

CAELEN G. & VIGOUROUX N. (1985),
Une base acoustique et phonétique hiérarchisée : des faits aux connaissances.
Actes du symposium Franco-Suédois Grenoble.

CHENG Y.M. & GUERIN B. (1987),
Nasal vowel study : formant structure, perceptual evaluation and neural representation in a model of the peripheral auditory system.
Institut de la Communication Parlée, Bulletin n°0.

DELATTRE P. (1969),
The General Phonetic characteristics : Final report
US Department of Health, Education & welfare.
Office of Education Institute of International Study.

FENG G., ABRY C. & GUERIN B. (1985),
How to cope with nasal vowels? Some acoustic boundary poles.
Actes du symposium Franco/Suedois sur la parole.
Grenoble.

FUJIMURA O. & LINQUIST J. (1971),
sweep-tone measurment of vocal-tract characteristics.
J. acoust. Soc. Am., 49 (2), p541-558.

GAY (1968),
effects of speaking rate on diphtong formant movements.
J. acoust. Soc. Am., 44, p1570-1573.

HAWKINS S. & STEVENS K.N. (1985),
Acoustic and perceptual correlate of non-nasal/nasal distinction for vowels.
J. acoust. Soc. Am., 77, p1560-1575.

KOBATAKE H. & OHTANI (1987),
Spectral transition dynamics of voiceless stop consonnants.
J. acoust. Soc. Am., 81(4), p1146-1151.

LAFFERIERE F. & O'SHAUGHNESSY D. (1986),
Analyse-synthèse et étude de de règles acoustiques de production avec un synthétiseur à formants.
15eme J.E.P. AIX EN PROVENCE P11-14.

LINDBLOM B. & STUDDERT-KENNEDY M. (1967),
On the role of formant transitions in vowel recognition.
J. acoust. Soc. Am., 42, p830-843.

MAEDA S. (1984),
Une paire de pics spectraux comme corrélat acoustique de la nasalisation des voyelles.
13eme J.E.P. BRUXELLE P223-224.

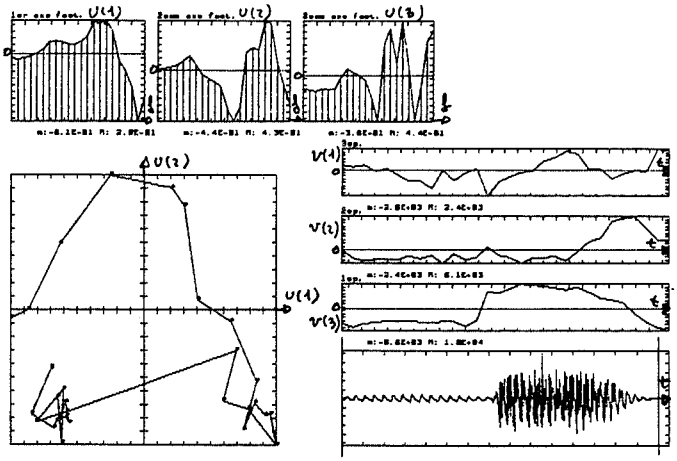
MARCUS S.M. (1984),
Temporal decomposition of speech.
IPO Annual progress report p25-31.

MERMELSTEIN P. (1977),
On detecting nasals in continuous speech.
J. acoust. Soc. Am., 16 (2), p581-587.

MOUSSA T. (1982),
L'analyse factorielle des correspondances et la reconnaissances des formes.
Thèse de Docteur d'Etat ès Science. univ. PARIS 6.

MRYATI M. (1976),
Contribution aux études sur la production de la parole
Thèse de docteur d'état, I.N.P. GRENOBLE

ROSSI M. (1983),
Indices acoustiques multilocuteurs et indépendant du contexte pour la reconnaissance automatique de la parole
Speech comm. 2, p215-217.




```
Fichier : baki_rd1
==================
ordre du modele:12   % explication:90

debut:1  fin:3 coeff-var:19  deformation

 aasses
bd-freq:Hz1340.1550 croissant enes:100
 anti aasses
bd-freq:Hz180.215 decroissant enes:3
bd-freq:Hz450.540 decroissant enes:28
bd-freq:Hz2060.2350 decroissant enes:39
bd-freq:Hz2700.3100 decroissant enes:97

debut:3  fin:12 cible-atteinte

debut:12  fin:14 coeff-var:142  deformation

 aasses
bd-freq:Hz650.760 croissant enes:17
bd-freq:Hz2060.2350 croissant enes:52
 anti aasses
bd-freq:Hz180.215 decroissant enes:7
bd-freq:Hz215.260 decroissant enes:8
bd-freq:Hz4000.4500 decroissant enes:100
bd-freq:Hz4500.5000 decroissant enes:69

debut:14  fin:17 coeff-var:41  deformation

 aasses
bd-freq:Hz1340.1550 croissant enes:100
 anti aasses
bd-freq:Hz180.215 decroissant enes:9
bd-freq:Hz260.320 decroissant enes:13
bd-freq:Hz450.540 decroissant enes:16
bd-freq:Hz3100.3550 decroissant enes:96
bd-freq:Hz4000.4500 decroissant enes:80
bd-freq:Hz4500.5000 decroissant enes:13

debut:17  fin:21 cible-atteinte

debut:21  fin:22 coeff-var:18  deformation

 aasses
bd-freq:Hz450.540 croissant enes:40
bd-freq:Hz2060.2350 croissant enes:97
bd-freq:Hz2700.3100 croissant enes:100
 anti aasses
bd-freq:Hz1340.1550 decroissant enes:88

debut:22  fin:30 coeff-var:173  deformation

 aasses
bd-freq:Hz180.215 croissant enes:8
bd-freq:Hz215.260 croissant enes:9
bd-freq:Hz4000.4500 croissant enes:100
bd-freq:Hz4500.5000 croissant enes:59
 anti aasses
bd-freq:Hz650.760 decroissant enes:20
bd-freq:Hz1550.1790 decroissant enes:53

bd-freq:Hz2060.2350 decroissant enes:45

detection cible naso-pharyngale :
detection d'energie BF : deb:22 fin:30
detection d'energie autour de 1000Hz :

l=<1,3,<cpp1,constant>,<cpp2,decroissant>>,<3,12,<cpp1,constant>,<cpp2,constant
  >>,<12,14,<cpp1,croissant>,<cpp2,croissant>>,<14,17,<cpp1,constant>,<cpp2,
  decroissant>>,<17,21,<cpp1,constant>,<cpp2,constant>>,<21,22,<cpp1,constant>,
  <cpp2,croissant>>,<22,30,<cpp1,croissant>,<cpp2,croissant>>,nil
```

ANNEXE :

Résultats de l'analyse de la transition /b/-/a/.
Représentation graphique et sorties du module d'interprétation.

ETUDE DU COMPORTEMENT PHONETIQUE DES DISSIMILARITES

Haiyan YE*, Denis TUFFELLI*, Louis-Jean BOE**

Institut de la Communication Parlée, UA CNRS no. 360

*Laboratoire de la Communication Parlée
INPG-ENSERG, 46 avenue Félix Viallet, 38031 GRENOBLE CEDEX

**Institut de Phonétique de Grenoble
Université Grenoble III, 38400 St. Martin d'Hères

ABSTRACT

The aim of this work is to evaluate, in the light of an a priori phonetic knowledge, the behaviour of dissimilarities used in case of speech recognition. To do this, we used as "references" synthetic vocalic stimuli for which we have a structural phonetic representation. The distance matrix calculated using different methods permitted us, with the help of multidimensional scaling specially adapted to perceptual data, to reconstruct the output spaces. The behaviour of the preprocessing and the comparison stages of the recognition system, is in this case sketchily assimilated with an auditor using a perceptual judgment. The appraisal of two spaces at a time - that of reference and that of output - calculated in this way allows us to compare these preprocessing and to make an extrinsic (phonetic) judgment on their behaviour. This evaluation could permit and outlet on new strategies for processing, pattern matching and decision making in speech recognition.

I. INTRODUCTION

Le but de ce travail est d'évaluer, à la lumière de connaissances phonétiques a priori, le comportement de distances (ou dissimilarités) utilisées dans le cadre de la reconnaissance automatique de la parole : distance d'ITAKURA, de PLOMP, de KLATT (WSM : Weighted Slope Metric), APS, MFCC et cepstrale.

Pour ce faire, nous avons utilisé des stimuli vocaliques synthétiques "de référence" - il s'agit de voyelles du français dont la fonction de transfert est calculée à partir de fonctions d'aires proposées par BOE (1973) adaptées par MRAYATI (1976) et améliorées par FENG (1986) - pour lesquels on dispose d'une représentation phonétique structurelle, (articulatoire, acoustique et perceptive). Nous appelerons ces données : l'espace de référence.

Les matrices de distances calculées avec différentes méthodes nous ont permis, grâce à une analyse multidimensionnelle (KRUSKAL, 1979) spécialement adaptée aux données perceptives, de reconstruire des espaces de sortie de deux ou trois dimensions. Le comportement des étages de pré-traitement et de comparaison du système de reconnaissance étant, dans ce cas, très schématiquement assimilé à un auditeur opérant un jugement perceptif.

Les appréciations deux à deux des espaces - de référence et de sortie - ainsi calculés permettent de comparer ces traitements et de porter sur leurs comportement un jugement extrinsèque (phonétique).

Cette évaluation pourrait permettre de déboucher sur de nouvelles stratégies pour les traitements, la mise en correspondance (Pattern Matching) et la prise de décision dans les systèmes de reconnaissance automatique de la parole.

II. LA COMPARAISON DE STRUCTURES

Pour pouvoir comparer deux systèmes (le système auditif et un opérateur machine) opérant sur les mêmes objets, mais de façons très différentes, il faut disposer - d'une manière ou d'une autre - de résultats de même nature ou tout au moins susceptibles d'être évalués avec les mêmes critères.

C'est tout l'objet des techniques de "scaling" permettant de reconstruire un espace mathématique à partir de matrices de distances fournies par le sujet (résultats de tests perceptifs) ou induites des dissimilarités et/ou similarités. Il est très important de pouvoir analyser la structure interne de cet espace (SINGH & WOODS, 1970).

Afin de pouvoir porter un jugement sur un opérateur machine fournissant des distances (qui définissent une relation entre des objets dans un espace) nous allons reconstruire un autre espace mathématique et le comparer à un ou des espaces que nous appelerons "de référence" (opératoires et classificatoires) que sont l'espace articulatoire (défini en termes de lieu et d'aperture) acoustique (le bien connu plan F1/F2) et perceptif (issu d'analyse multi-dimensionnelle de données perceptives), voir par exemple LONCHAMP (1978). Pour construire cet espace mathématique à comparer, nous avons utilisé l'analyse multidimensionnelle qui est rarement appliquée à un opérateur machine.

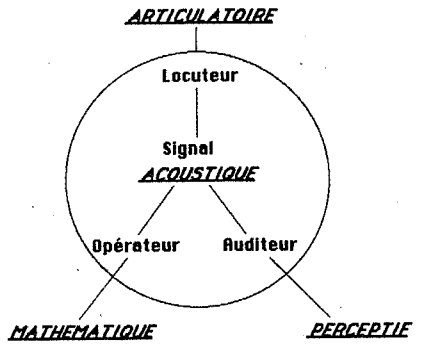


Fig.1 L'emetteur/recepteur, le signal un opérateur et les espaces associés.

Fig.1

Cette technique a été récemment utilisée pour analyser un système de reconnaissance (YUCHTMAN et al., 1986).

III. ANALYSE MULTI-DIMENSIONNELLE (MDS)

L'analyse multidimensionnelle est en fait constituée par un ensemble de méthodes qui permettent de construire la configuration d'objets dans un espace à faible dimension en utilisant les distances inter-objets ou des informations sur les ordres de grandeur de distances bruitées.

L'entrée typique de cette analyse consiste en une ou partie d'une matrice, qui est interprétée comme l'ensemble de distances interpoints bruitées et ayant peut-être subies une transformation inconnue (mais dont on suppose qu'elle conserve l'ordre de l'espace).

Une sortie typique est constituée par les coordonnées de la projection des points dans un espace à deux ou trois dimensions.

Une analyse de type INDSCAL "Individual Difference Scaling", permet d'utiliser plusieurs matrices de distances interpoints pondérées. TORGERSON est l'un des premiers chercheurs à avoir travaillé sur cet ensemble de méthodes dès 1927. C'est lui qui a proposé le terme de "Multidimensional Scaling" et qui a largement contribué à sa diffusion. Plus tard, SHEPARD (1962), CARROL et CHANG (1970), KRUSKAL (1977) ont proposé des conceptions nouvelles et différentes; ils ont réalisé de nombreux programmes. Cette analyse a trouvé nombre d'applications dans les analyses de données psychophysiques.

De nombreuses études sur la perception de la parole ont utilisé l'analyse multidimensionnelle (MDS) avec succès, pour la mise en évidence de structures intrinsèques.

Nous allons donner un bref résumé de cette méthode.

Soit une matrice de dissimilarité (tij) à dimension I*I : le modèle le plus simple consiste à interpréter ces dissimilarités comme des distances dij entre I points dans $\mathbb{R}^R$. On suppose qu'il y a I points Xi=(Xi1,Xi2,...,XiR) satisfaisant à :

$$t_{ij} \approx d_{ij} = \sum_{r=1}^{R} (X_{ir} - X_{jr})^2 \qquad \text{pour tous les (i, j) disponibles}$$

L'égalité approximative indique bien que dij n'est qu'une estimation de tij qui peut être entâchée d'erreurs aléatoires (bruit).

En fait, il n'est pas forcément nécessaire de disposer d'une matrice de dissimilarité; une matrice de similarité convient parfaitement. Pour cette analyse il suffira d'opérer quelques transformations simples.

Une première approche de modèle a consisté à transformer linéairement les données d'entrée.

$$a + b\, t_{ij} \approx d_{ij} \qquad \text{pour tous les (i, j) disponibles avec a, b} > 0$$

D'une façon générale, on peut avoir avec ce genre de modèle :

$$f(t_{ij}) \approx d_{ij}$$

où f est une fonction polynomiale monotone croissante dans la zone comprenant tij.

Une mesure d'erreur (stress) est définie pour évaluer l'aproximation des données par le modèle. Plus le "stress" est petit, plus le modèle est proche des données. Une procédure récursive est utilisée afin de trouver la solution qui minimise le "stress".

IV. Quelques résultats de projections

Nous présentons ici quelques résultats de projection, après l'analyse MDS, dans un espace à deux dimensions. Nous avons utilisé un logiciel expérimenté par F. LONCHAMP, implanté par J. GENIN & D. PASCAL et adapté au microVAX par YE.

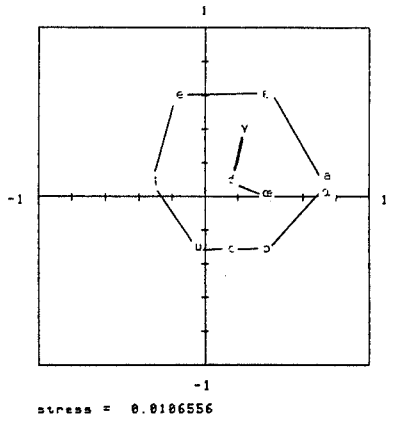


Fig.2a
Projection de la distance d'ITAKURA sur 2 dimensions après MDS

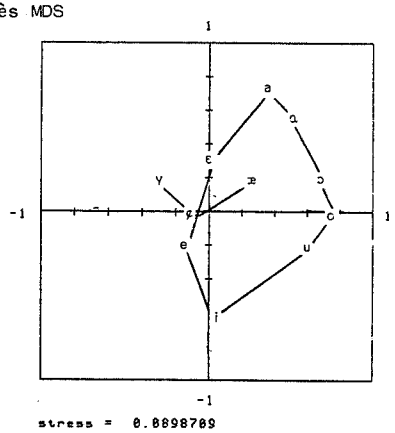


Fig.2b
Projection de la distance cepstrale sur 2 dimensions après MDS

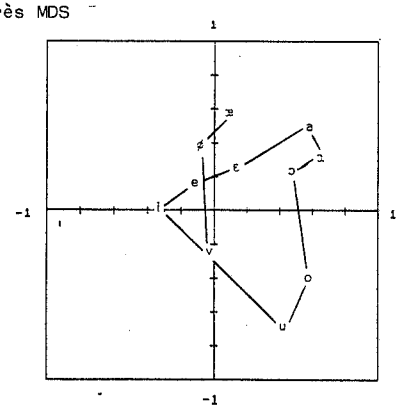


Fig.2c
Projection de la distance de PLOMP.

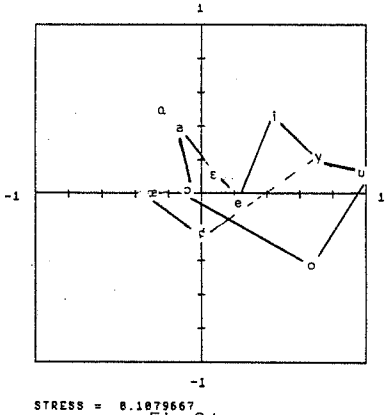


Fig.2d
Projection de la distance de WSM.

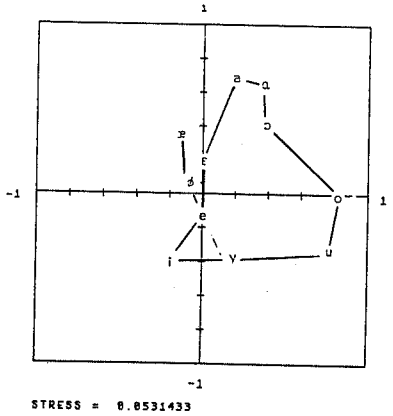


Fig.2. e
Projection de la distance de MFCC.

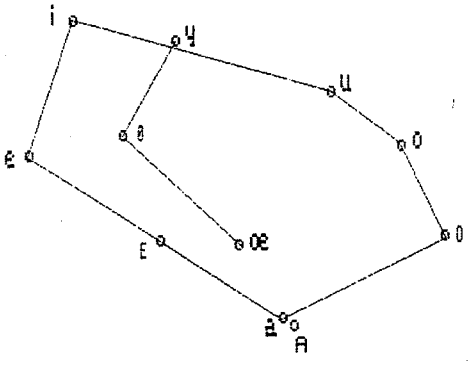


Fig.2. f
Projection de la distance APS.

V. DISCUSSION ET CONCLUSION

Pour évaluer les performances des dissimilarités et pour pouvoir apprécier leur discriminabilité phonétique, il faut définir des critères de comparaison. KANE-ESRING et al. (1987) ont proposé deux critères assez globaux :

1) la variance de distances entre plusieurs répétitions de mêmes mots doit être petite et celle entre des mots différents doit être grande,

2) la "forme" de distances doit être bien corrélée avec celle de la similarité phonétique.

Pour notre part nous allons définir quatre critères structurels plus phonétiques. Il va s'agir "d'apprécier" :

1) La bonne description de l'espace maximum décrit par les trois voyelles cardinales extrêmes /i, a, u/.

2) L'opposition entre les voyelles arrières et les voyelles d'avant, associée à F2, ou à F'2-F1 (le "spread" proposé par FANT (1983)).

3) La disposition des voyelles arrondies /y, EU, OE/ par rapport à la série antérieure correspondante. Ces deux classes des voyelles sont souvent difficiles à discriminer. Le critère F'2+F1 appelé "flatness" permet de séparer ces deux catégories (FANT, 1983).

4) La position des voyelles de timbre intermédiaire par rapport aux extrêmes (fermées/ouvertes): elle est associée au paramètre F1.

Avec ces critères, nous pouvons évaluer qualitativement les performances des dissimilarités: étant bien entendu que nous privilégierons la bonne tenue structurelle à la seule efficacité discriminatoire. Trois valeurs +1 0 -1 vont nous permettre de juger les espaces de sorties avec ce critère : +1 correspond à une bonne discrimination, 0 à une discrimination médiocre et -1 à une mauvaise discrimination.

On a constitué le tableau I avec ces valeurs ternaires en fonction des dissimilarités et des projections issues de l'analyse multidimensionnelle. Pour la distance cepstrale on a utilisé les trois possibilités suivantes de "liftering" : abscence de pondération (sans), pondération linéaire (linéaire), pondération quadratique (quadra.). LPCZwi représente une intégration spectrale sur des bandes 1 Bark de ZWICKER (ZWICKER, 1981) obtenues par une analyse LPC du signal. LPCKla correspond à une intégration spectrale sur des bandes de KLATT (KLATT, 1979) précédée d'une analyse LPC du signal. Avec FFTZwi est opéré une intégration spectrale sur des bandes de ZWICKER avec avant une analyse FFT du signal. FFTKla est associé à une intégration spectrale sur des bandes de KLATT obtenues par une analyse FFT du signal (voir Tableau 1).

La figure 2a présente une projection de la distance d'ITAKURA sur un espace à deux dimensions. On voit bien que l'espace maximal est assez bien représenté. Les voyelles arrières se distinguent des voyelles d'avant. Par contre les voyelles arrondies et non-arrondies sont un peu confondues. L'ordre des voyelles dans la dimension fermée/ouverte est très mal respecté, ce qui n'est pas une surprise car la distance d'ITAKURA n'a pas un bon comportement pour les formants bas (YE & TUFFELLI, 1987).

Pour l'ensemble des dissimilarités, les voyelles avant/arrière sont les plus faciles à séparer et les ordres des voyelles fermés/ouvertes sont les moins bien respectés.

Si l'on examine de près ces projections, on peut noter d'abord que les voyelles arrondies et non-arrondies sont souvent confondues, la relation structurelle entre ces deux séries est mal respectée. Le fait que ces deux séries de voyelles soient assez proches acoustiquement, articulatoirement et algorithmiquement mais bien distinguées perceptivement donne à penser que le système auditif utilise une pondération spéciale des paramètres pour renforcer certains contrastes privilégiés.

Parmi toutes les dissimilarités, et sur l'ensemble des critères, la distance de APS est très prometteuse, cela n'est pas très étonnant car c'est la seule distance utilisant des paramètres proches des corrélats acoustico-phonétiques. On trouve également que la distance de PLOMP avec le traitement LPCKla et FFTZwi donne une meilleure représentation structurelle. La distance WSM avec le traitement LPCZwi respecte très mal cette structure. L'ordre de respect structurel est le suivant :

APS		+4
PLOMP	LPCKla	+3
PLOMP	FFTZwi	+3
CEP	linéaire	+3
MFCC	LPCZwi	+2
MFCC	FFTZwi	+2
ITAKURA		+1
MFCC	LPCKla	+1
CEP	sans	+1
WSM	FFTZwi	0
CEP	quadra.	-1
PLOMP	LPCZwi	-1
MFCC	FFTKla	-1
CDT		-1
WSM	FFTKla	-2
PLOMP	FFTKla	-3
WSM	LPCKla	-3
WSM	LPCZwi	-4

Ce classement donne une information plus détaillée et plus riche qu'un simple taux de reconnaissance à travers un système de reconnaissance. On voit bien ici qu'un "liftering" linéaire des coefficients ceptraux améliore les résultats. Comparativement le traitement MFCC permet d'obtenir de bons résultats.

Tableau 2.4.5. L'évaluation des performances

Dissimilarités		cardinales extrêmes	voyelles d'avant	voyelles arrondies	voyelles fermées	total
ITAKURA		+1	+1	0	-1	+1
APS		+1	+1	+1	+1	+4
CEP	sans	+1	+1	0	-1	+1
	linéaire	+1	+1	0	+1	+3
	quadra.	0	+1	-1	-1	-1
PLOMP	LPCZwi	0	+1	-1	-1	-1
	LPCKla	+1	+1	+1	0	+3
	FFTZwi	+1	+1	0	+1	+3
	FFTKla	0	-1	-1	-1	-3
WSM	LPCZwi	-1	-1	-1	-1	-4
	LPCKla	-1	0	-1	-1	-3
	FFTZwi	+1	0	-1	0	0
	FFTKla	+1	-1	-1	-1	-2
MFCC	LPCZwi	0	+1	+1	0	+2
	LPCKla	0	+1	0	0	+1
	FFTZwi	+1	+1	-1	+1	+2
	FFTKla	-1	+1	-1	0	-1
CDT		0	+1	+1	-1	+1
somme		+6	+10	-5	-5	

Références :

/1/. BOE LJ. (1973) "Etude Acoustique du Couplage Larynx-Conduit Vocal" Revue d'Acoustique 27, 235-244

/2/. CARATY M.J. & RODET X. (1985) "Distance Interspectrale à Critères Perceptifs" 14èmes JEP, Paris, 87-90

/3/. CARROLL D. & CHANG J.J. (1970) "Analysis of Individual Differences in Multidimensional Scaling Via an N-Way Generation of Eckart-Young Decomposition" Psychometrika 35(3), 283-319

/4/. FANT G. (1983) "Feature analysis of swedish vowels - a revisit" STL-QPSR 2-3/1983, 1-19.

/5/. FENG G. (1986) "Modélisation Acoustique et Traitement du Signal de Parole" Thèse nouveau régime INPG.

/6/. LONCHAMP F. (1978) "Recherche sur les Indices Perceptifs des Voyelles orales et Nasales" Thèse 3ème Cycle, Laboratoire de Phonétique, Université de Nancy II.

/7/. KANE-ESRING Y., STREETER L.A., KAMM C., DEVLIN S. & MACCHI M. (1987) "Evaluating Spectral Distance Measures with Reference to Human Perception" J. Acoust. Soc. Am. Suppl.1, 81, S95.

/8/. KRUSKAL J.B. (1977) "Multidimensional Scaling and other Methods for Discovering Structure" In Mathematical Methods for Digital Computers. Vol. III Statistical Methods for Digital Computers Vol.III, Ed. by ENSLEIN K., RALSTON A., WILF H.S. John Wiley, New-York, 296-339.

/9/. MRAYATI M. (1976) "Contribution aux Etudes sur la Production de la Parole" Thése Docteur d'Etat I.N.P. Grenoble

/10/. PASCAL D. (1984) "Analyse de Données Perceptives par les Méthodes d'Echelle Multidimensionnelle INDSCAL et MDSCAL ou Analyse des Proximités" Séminaire Traitement de Données Phonétique, 17-21 Sept. 1984, Grenoble.

/11/. POLS L.C.W., Van Der KAMP L.J.Th. & PLOMP R. (1969) "Perceptual and Physical Space of Vowel Sounds" J. Acoust. Soc. Am. 49(2), 458-467

/12/. SHEPARD R.N. (1962) "Analysis of Proximities : Multidimensional Scaling with Unknown Distance Function" Psychometrika 27(2), 125-140; 27(3), 219-246.

/13/. SINGH S. & WOODS D. (1970) "Perceptual Structure of 12 American English Vowels" J. Acoust. Soc. Am. 49(6), 1861-1866

/14/. TORGERSON W.S. (1958) **THEORY AND METHODS OF SCALING** John Wiley, New York

/15/. YE H. & TUFFELLI D. (1987) "Deterministic Characteristics of the LPC Distances: an Inconsistency with Perceptual Evidence" IEEE, Int. Conf. ASSP 1987, 1.1, Dallas

/17/. YUCHTMAN M., NUSBAUM H.C. & DAVIS C.D. (1986) "Multidimensional Scaling of Confusion Produced by Speech Recognition System" J. Acoust. Soc. Am. Suppl. 1. 79, S95

RECONNAISSANCE DE /VCV/ (CYCLES VOCALIQUES) EN PAROLE CONTINUE PAR COMPARAISON DYNAMIQUE

Marie-José CARATY, Xavier RODET

LAFORIA, Université Pierre et Marie Curie, PARIS 6
4, place Jussieu - 75005 PARIS
(43 36 25 25, poste 47-57)

ABSTRACT

This paper presents a method for acoustico-phonetic decoding of connected speech using anchor points (steady state vowels) and /VCV/ pattern identification around anchor points.
/VCV/ pattern matching uses a dynamic time warping algorithm without explicit segmentation of test patterns.

The method has been used for identification of /a C i/, C ∈ {/p/, /t/, /k/} in connected speech.

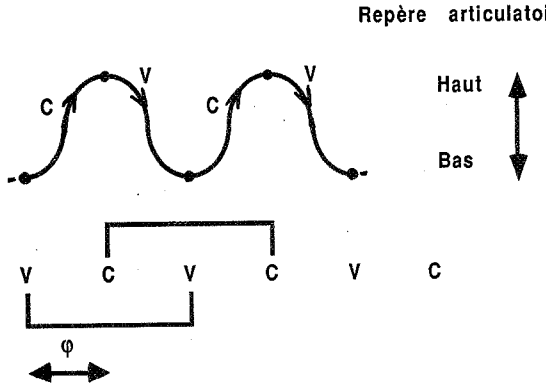
I. INTRODUCTION

Plusieurs raisons motivent le choix du cycle vocalique /VCV/ comme unité de référence pour le décodage acoustico-phonétique.

— Compte tenu des travaux sur la coarticulation [OHM-67], on sait que l'influence d'une consonne porte sur l'entourage vocalique. Par conséquent une séquence /VCV/ comporte le **maximum de cohérence intra segmentale** [OHM-66].
Il faut cependant remarquer la contrepartie du choix d'une telle unité : il existe "potentiellement" environ 15x19x15 combinaisons /VCV/, limitées heureusement par des contraintes phonotactiques. A titre d'exemple, il y a 34^3 triphones possibles mais sur comptage de plus de 300000 sons, 12000 triphones environ, soit 30 % seulement ont été observés [TUB-85].

— Nous privilégions les cycles vocaliques /VCV/ plutôt que les cycles consonantiques /CVC/ parce que, dans l'objectif particulier de reconnaissance d'**îlots de confiance** amorçant la reconnaissance, il est bien moins difficile de segmenter et d'identifier des **noyaux vocaliques** que des consonnes.

— Nous aurions pu appeler nos unités des triphones, par extension du terme diphone introduit pour la synthèse par éléments, le diphone allant de la partie centrale d'un son à un autre son.
Mais cette appellation :
- ne renvoie pas à la nature de nos unités, car il existe, en effet, des triphones /VVV/ et /CCC/,
- ne met pas l'accent sur la séquence privilégiée VC, et nous cantonne à un champ d'application qui n'est pas relié à tous les travaux sur l'organisation temporelle de la parole.



Les unités de référence sont des segments qui, dans une séquence ...VCVC..., vont de la partie centrale d'une voyelle à la suivante. En se référant aux travaux sur la parole, en relation avec la psycho-motricité ([TUL-84], [MUN-85], [BEN-86]), il s'agit là d'un cycle vocalique (le terme renvoie à une séquence indéfinie VCVCV...), le cycle consonantique allant d'une consonne à la suivante, et la phase (φ) caractérisant le décalage entre les deux cycles. Sans entrer ici dans les détails, cette approche de la parole se révèle depuis plusieurs années excessivement productive pour l'étude de l'organisation temporelle de la parole, en relation avec les mécanismes de production et leurs effets sur le signal acoustique.
Dans la séquence ...VCVC... c'est d'ailleurs le champ VC qui se révèle le **lieu** par excellence des **contrôles temporels**, mis en évidence dans des variations de débit et pour des conditions différentes, d'allongements vocaliques, d'accent et d'intonation.

II. PRESENTATION DE L'ETUDE

Notre but fondamental est de reconnaître, à partir d'îlots de confiance (les noyaux vocaliques) [CAR-87], des formes acoustiques relativement invariantes, les cycles vocaliques, par comparaison avec des formes de référence.
Nous avons développé et implanté une procédure de comparaison des formes-test et des formes de référence, avec **déformation non linéaire**, et **sans connaissance précise** des positions temporelles des **extrémités**.

III. PRESENTATION DU CORPUS

Compte tenu de la complexité de la tâche et du très grand nombre des cycles vocaliques de référence, nous nous limitons dans un premier temps à l'idendification des cycles $/V_1CV_2/$ dont la consonne intervocalique est une occlusive sourde (C ∈ {p, t, k}) dans le contexte vocalique de deux voyelles cardinales extrêmes (V_1 = /a/ et V_2 = /i/).

Le **corpus d'apprentissage** est constitué d'une série de trois phrases porteuses contenant les unités **/a C i/** en position début, milieu et fin de phrase. Les phrases sont donc de la forme :

"**/a C i/** final".
"C'est l'**/a C i/** final".
"C'est l'**/a C i/**".

Exemple : "C'est l'aki final", etc...

Les consonnes intervocaliques considérées à l'apprentissage sont les suivantes :

/p/, /t/, /k/, /b/, /d/, /g/, /m/, /n/, /l/, /v/, /z/, /s/, /r/, /f/.

Le **corpus-test** est un ensemble de 36 phrases :
- **3** cycles vocaliques (/api/, /ati/, /aki/),
- **3** positions dans la phrase (début, milieu, fin),
- **4** phrases différentes de chaque type.

Les deux corpus sont enregistrés dans un environnement peu bruité, la fréquence d'échantillonnage est de 10 kHz, la digitalisation sur 12 bits. Le locuteur (masculin) prononce chaque phrase de mémoire, après une lecture préalable ; un temps de pause de 10 secondes est observé entre chaque enregistrement.

IV. ANALYSE

Pour une meilleure définition des caractéristiques spectrales, nous avons préféré une analyse pseudo-synchrone à la période fondamentale, plutôt qu'une analyse centiseconde asynchrone [RAB-77].

IV.1. Analyse pseudo-synchrone

Les fenêtres d'analyse de durée fixe (22ms) sont centrées sur un échantillon présentant une amplitude maximum. Ainsi, dans les segments de parole voisés, les fenêtres sont positionnées de façon sensiblement constante par rapport à la période fondamentale et donc par rapport aux portions "glotte ouverte".

Le **schéma 1** représente le signal de parole du cycle vocalique de référence /api/ (en début de phrase). Les positions des échantillons-centre des fenêtres d'analyse ainsi déterminées sont visualisées par les lignes pointillées.

IV.2. Conditions et technique de l'analyse

Rappelons brièvement les conditions d'analyse :
- durée des fenêtres de 22ms,
- prétraitement par préemphase et fenêtrage de Hamming,
- analyse LPC d'ordre 16 (méthode d'autocorrélation),
- détection, à partir de l'enveloppe du spectre LPC, des caractéristiques des pics spectraux.

Le **schéma 2** représente les trajets des pics spectraux obtenus par analyse du signal du cycle de référence /api/. La position verticale d'un triangle indique la fréquence centrale du pic, la taille du triangle est représentative de son amplitude.

V. LOCALISATION DES SEGMENTS DE PAROLE

V.1. Fonction d'instabilité

Le rôle de la fonction d'instabilité du signal vocal est de mesurer la **variation** entre deux fenêtres consécutives.

— Une variation fondamentale dans le signal est l'**énergie**, elle est illustrée par l'enveloppe du signal temporel. La courbe de variation d'énergie $\{e_i\}$ considérée est obtenue à partir des énergies moyennes, calculées en dB, sur les fenêtres d'analyse successives, valeurs lissées sur 3 points.

— La seconde variation considérée est la variation **spectrale**. La courbe $\{d_{i,i+1}\}$ correspondante est obtenue à partir des valeurs calculées par les distances inter-spectres **APS** (par **A**justement de **P**ics **S**pectraux) [CAR-85,86,87] entre deux fenêtres consécutives, valeurs lissées sur 3 points.

Remarque : nous appliquerons la mesure APS sous sa forme non discriminante en choisissant, pour forme invariante, la forme minimale : n=0.

La **fonction d'instabilité** est alors définie à partir des valeurs des deux courbes précédentes par :

$$f_{ins}(i) = emoy_{i,i+1} * d_{i,i+1}$$

où, $emoy_{i,i+1}$ est l'énergie moyenne en dB des deux fenêtres temporelles consécutives (i et i+1), d'énergie e_i et e_{i+1}.

La **courbe d'instabilité** est obtenue à partir des valeurs, lissées sur 3 points, de la fonction d'instabilité.
Les minima relatifs de cette fonction coïncident avec des **parties stables**, les maxima relatifs correspondent à des **instants de transition**.

Le **schéma 3** représente les courbes d'énergie et d'instabilité du cycle de référence /api/.

V.2. Segmentation des références et des îlots de confiance

— Le principe de segmentation d'un cycle de référence est de définir, pour les deux voyelles le délimitant, les zones, considérées comme parties stables, des noyaux vocaliques à l'intérieur desquelles doivent être ajustées les frontières d'un cycle-test.

L'analyse des trajectoires de pics et des courbes d'instabilité des cycles vocaliques permet rapidement de constater que la partie stable, généralement assimilée à un noyau vocalique, est en réalité une "zone d'évolution spectrale" plus ou moins lente et importante selon le rythme d'élocution et la coarticulation. En s'y rapportant, compte tenu du **rôle important** de la **segmentation**, on ne peut qu'estimer la difficulté d'une identification des noyaux vocaliques sur une seule fenêtre. **Une prise en compte**, de l'**évolution spectrale** ou de la **modélisation des cibles vocaliques**, semble impérative pour **une reconnaissance optimale** des **voyelles** dans la **parole continue**.

Le **principe de segmentation intra-segmentale** choisi est le suivant :
- une chute de -2.5dB (seuil expérimental) au voisinage d'un noyau vocalique, relative au maximum de la courbe d'énergie, définit une zone de scrutation à partir de laquelle on repère les minima locaux de la fonction d'instabilité. Nous choisissons alors de délimiter la "zone de stabilité" (ou "zone d'ajustement") par les deux minima qui définissent la plus grande plage à l'intérieur de la zone de scrutation.

Ce principe ne permet pas de segmenter tous les segments de référence :
- certains (très rares) ne présentent aucun minimum local, nous avons alors choisi de délimiter la zone d'ajustement des frontières par les deux fenêtres présentant maximum d'instabilité et qui correspondent généralement, pour la fonction d'instabilité considérée, à des énergies maximales,
- d'autres, principalement les cycles admettant pour consonnes intervocaliques les liquides ou les nasales, ne permettent pas de délimiter le (ou les) noyau(x) vocalique(s) par une chute d'énergie. A défaut d'une segmentation adaptée, nous avons alors effectué une segmentation manuelle.

— Si le principe de reconnaissance d'un signal-test, ne nécessite pas une segmentation explicite des cycles, il nécessite la **segmentation intra-segmentale** des **îlots de confiance** qui amorcent la reconnaissance. Ces îlots sont des noyaux vocaliques, qui seront segmentés de la même façon que les références.
On remarquera que le rythme d'élocution des phrases-test est plus rapide et que les noyaux présentent, plus souvent que dans le cas des références, un seul minimum, nous choisissons alors de définir la zone d'ajustement par les deux fenêtres adjacentes à ce minimum.

La segmentation du cycle de référence /api/ est illustrée sur les **schémas 1, 2 et 3**.

VI. IDENDIFICATION DES SEGMENTS DE PAROLE

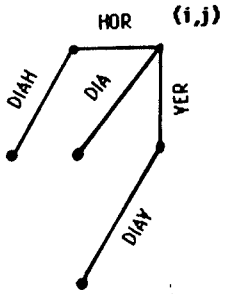
Nous nous inspirerons de **techniques de reconnaissance globale** dont les **performances** sont **unanimement reconnues**.
En effet, une variante de l'**algorithme de comparaison dynamique** est particulièrement intéressante pour notre application puisqu'elle permet une reconnaissance à partir d'un **îlot de confiance** (le noyau vocalique) :
- **sans segmentation explicite** du segment /VCV/ dans le signal-test,
- avec **prise en compte**, dans la représentation des unités de référence, de l'**incertitude** concernant la **position des frontières**.

VI.1. Choix de l'algorithme de comparaison dynamique

— Différentes publications semblent confirmer que des contraintes locales symétriques présentent un avantage en reconnaissance, sur leurs homologues asymétriques ([SAK-79], [MYE-80]), et nous ont conduits à retenir un **principe de contraintes locales symétriques** avec une **pente de 2** [SAK-78].

— La technique d'analyse pseudo-synchrone (intervalles d'analyse irréguliers) a conduit à redéfinir la **fonction de pondération** donnée par Sakoe et Chiba qui, rappelons-le, est à **valeurs positives** et doit permettre la **normalisation** de la distance d'un test à une référence.

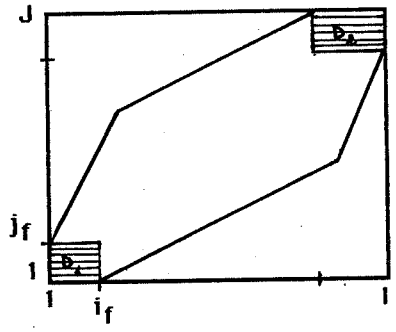
Nous illustrerons ici les pondérations choisies pour un principe de contraintes locales symétriques de pente 2.



HOR = $x_i - x_{i-1}$
VER = $y_j - y_{j-1}$
DIA = $x_i + y_j - (x_{i-1} + y_{j-1})$
DIAH = $x_{i-1} + y_j - (x_{i-2} + y_{j-1})$
DIAV = $x_i + y_{j-1} - (x_{i-1} + y_{j-2})$

où x_i (resp. y_j) représente l'échantillon-centre de la fenêtre i (resp. j) du cycle de référence (resp. de test).

— Nous adopterons le principe de **relâchement aux frontières sur les 2 axes** ([MAR-83], [BOY-87], [GAU-82]) qui donne une plus grande liberté pour la positions des points C(1) et C(K) que le relâchement sur un seul axe.



Les contraintes aux frontières deviennent alors :

$C(1) \in D_1$ et $C(K) \in D_2$ que nous définirons ici pour C(1) :

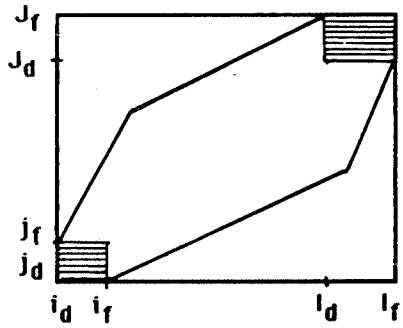
$C(1) \in D_1 : 1 \leq i(1) \leq i_f$ et $1 \leq j(1) \leq j_f$

L'**équation locale** est **redéfinie** puisqu'elle est fondée sur deux hypothèses qui ne sont plus vérifiées :
- contraintes aux frontières (1,1) et (I,J),
- longueur constante des sous-chemins autorisés, dont l'extrémité est un point de la matrice de comparaison, par le choix même des valeurs de pondération.

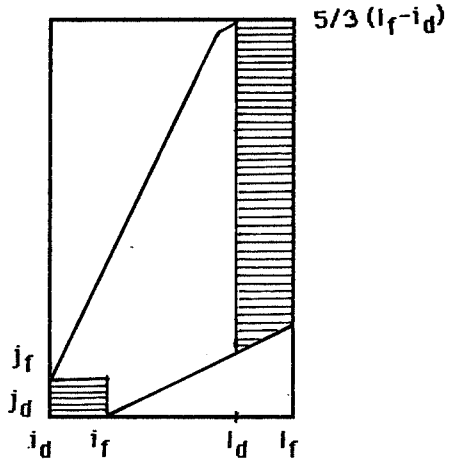
Le principe d'optimalité locale est alors modifié en tenant compte de la longueur variable des chemins de recalage, c'est-à-dire en normalisant, dans l'équation locale, g(i',j') par rapport à la longueur du chemin optimal d'extrémité (i',j').

Le principe de **comparaison** dynamique avec **relâchement** aux frontières sur les **deux axes** est **très adapté** à la reconnaissance de **cycles vocaliques**. En effet, si la segmentation des mots enchaînés peut être difficile, cela l'est d'autant plus pour des segments de taille inférieure dans un flot de parole continue, où l'on observe généralement un rythme d'élocution rapide, une moins bonne "tenue" des voyelles et une cible vocalique qui n'est pas toujours atteinte.
L'identification d'un cycle vocalique $/V_1CV_2/$ est amorcée après la **reconnaissance** d'une au moins des deux voyelles V_1 et V_2 ; ces voyelles sont des îlots de confiance, elles fixent une partie au moins du contexte vocalique des références à comparer et limitent ainsi la combinatoire des comparaisons.
Deux cas peuvent se présenter que nous illustrerons par les domaines de recherche respectifs du chemin optimal et des relâchements aux frontières sur les deux axes :

— V_1 et V_2 sont identifiés



— V_1 seul est identifié



VII. RESULTATS EXPERIMENTAUX

Dans l'expérience de reconnaissance, nous considérons identifiée sur le signal-test, comme îlot de confiance, la voyelle du cycle vocalique $/V_1CV_2/$ présentant la plus forte énergie et la meilleure tenue.
Le domaine de recherche du chemin optimal dans l'algorithme de programmation dynamique est par conséquent celui illustré par le schéma correspondant à une seule voyelle identifiée.
Dans la phase d'identification, le cycle vocalique est comparé aux 14 cycles de référence suivant leurs 3 contextes d'occurrence, soit 42 comparaisons, et identifié à la référence pour laquelle la distance calculée par l'algorithme de comparaison dynamique est minimum.

Les résultats sont les suivants :

	Cycle vocalique	Nombre de cycles identifiés (sur 4 occurrences)
DEBUT	/api/	3
	/ati/	3
	/aki/	4
MILIEU	/api/	2
	/ati/	2
	/aki/	1
FIN	/api/	4
	/ati/	3
	/aki/	1
		= 23/36

Les résultats de cette expérience ne prétendent pas valider la méthode d'un point de vue qualitatif ; il est clair que cette méthode ayant été implantée nécessite désormais une étude approfondie de son comportement.

— En effet, certaines confusions des cycles vocaliques (en milieu de phrases) avec des références du cycle /avi/ sont tout-à-fait surprenantes et demandent une **analyse** systématique des **chemins optimaux** déterminés par l'algorithme de comparaison dynamique, qui n'a pas encore été faite.

— Si le principe de segmentation tient compte, ou plus exactement -comme nous le pensons -doit tenir compte, de la variation d'énergie et d'une fonction d'instabilité, la comparaison dynamique, elle, n'intégre que la variation spectrale, une meilleure identification pourrait être obtenue par **ajout** de l'**information d'énergie**, qu'elle intervienne, soit en tant que sur-opérateur de la mesure APS dans l'algorithme, soit dans un principe de comparaisons dynamiques parallèles (à deux niveaux).

— Une fois cette mise au point "terminée", le **comportement** de la méthode devra être **étudié** sur un plus **large corpus** et une plus grande **variété des cycles vocaliques**, tant au niveau de la consonne intervocalique que du contexte vocalique.
Alors seulement pourra-t-on prendre en considération les scores de reconnaissance obtenus.

VIII. CONCLUSION

— **La reconnaissance des cycles vocaliques**, ouvre la voie à des études plus spécifiques, sur les diverses méthodes composantes, d'une procédure de reconnaissance fondée sur une comparaison dynamique tenant compte d'un ajustement aux frontières.

— La méthode de comparaison globale des formes spectro-temporelles utilisée pour la reconnaissance des /VCV/ suppose, en quelque sorte, que l'information utile à la reconnaissance est uniformément répartie sur l'axe temporel. Remarquons que si nous avons critiqué, quant à l'axe fréquentiel, les distances qui fondaient cette hypothèse, nous estimons, également, qu'en utilisant une **distribution de "poids" sur l'axe temporel des références** (et donc, par ajustement temporel sur celui du test), on pourrait, en quelque sorte, **favoriser les instants** (nous pensons à l'explosion de /d/ dans la transition /di/) qui, **du point de vue perceptif**, semblent **plus pertinents** que d'autres.

Enfin, il ne fait pas de doute que ces méthodes **devront être testées** sur des **corpus** beaucoup **plus larges, laissant varier** et le **locuteur**, et les **contextes**, et les conditions d'**élocution** et de **bruit**. C'est alors seulement, que l'on pourra mesurer, avec beaucoup plus de précision, les gains apportés par ces méthodes au décodage acoustico-phonétique de la parole continue.

BIBLIOGRAPHIE

[BEN-86] C. BENOIT, C. ABRY
"Vowel-consonant timing across speakers"
12th ICA, paper A6-1, Montreal, 1986.

[BOY-87] A. BOYER
"Application des techniques de programmation dynamique et de quantification vectorielle à la reconnaissance des mots isolés et des mots enchaînés"
Doctorat de l'Université de NANCY I en Informatique, 1987.

[CAR-85] M.J. CARATY, X. RODET
"Distance interspectrale à critères perceptifs",
14èmes JEP, pp. 87-90, Paris,1985.

[CAR-86] M.J. CARATY, X. RODET
"Etude comparative de mesures de distorsion spectrale"
15èmes JEP, pp. 297-300, Aix-en-Provence, 1986.

[CAR-87] M.J. CARATY
"Contribution au décodage acoustico-phonétique : études de distances interspectrales et reconnaissance de cycles vocaliques"
Doctorat de l'Université de PARIS VI en Informatique, 1987.

[GAU-82] J.L. GAUVAIN
"Reconnaissance de mots enchaînés et détection de mots dans la parole continue"
Thèse de 3ème cycle en Electronique, PARIS-SUD, 1982.

[MAR-84] J. di MARTINO
"Contribution à la reconnaissance globale : mots isolés et mots enchaînés"
Thèse de Docteur-Ingénieur en Informatique, NANCY I, 1984.

[MYE-80] C.S. MYERS, L.R. RABINER, A.E. ROSENBERG.
"Performance tradeoffs in Dynamic Time Warping algorithms for Isolated Word Recognition"
IEEE TASSP, vol. AS-28, pp. 623-635, 1980.

[MUN-85] K.G. MUNHALL
"An examination of intra-articulation relative timing"
JASA, no. 78, pp. 1548-1553,1985.

[OHM-66] S.E.G. OHMAN
"Coarticulation in CVC utterances : spectrographie measurments"
JASA, vol. 39, no 1, p. 151, 1966.

[OHM-67] S.E.G. OHMAN
"Numerical Model of Coarticulation"
JASA, vol. 41, pp. 310-320, 1967.

[RAB-77] L.R. RABINER, B.S. ATAL, M.R. SAMBUR
"LPC prediction error-analysis of its variation with the position of the analysis frame"
IEEE TASSP, vol. ASSP-25, pp. 434-442, 1977.

[SAK-78] H. SAKOE, S. CHIBA
" Dynamic programming algorithm optimization for spoken word recognition"
IEEE TASSP, vol. ASSP-26, pp. 43-49, 1978.

[SAK-79] H. SAKOE
"Two-level DP-Matching- A dynamic programming-based pattern matching algorithm for connected word recognition"
IEEE TASSP, vol. ASSP-27, pp. 588-594, 1979.

[TUB-85] J.P. TUBACH, L.J. BOE
"Un corpus de transcriptions phonétiques : constitution et exploitation statistique"
Rapport ENST, avril 1985.

[TUL-84] B. TULLER, J.A.S. KELSO
"The timing of articulatory gestures : evidence for relational invariants"
JASA, no.76, pp.1534-1543,1984.

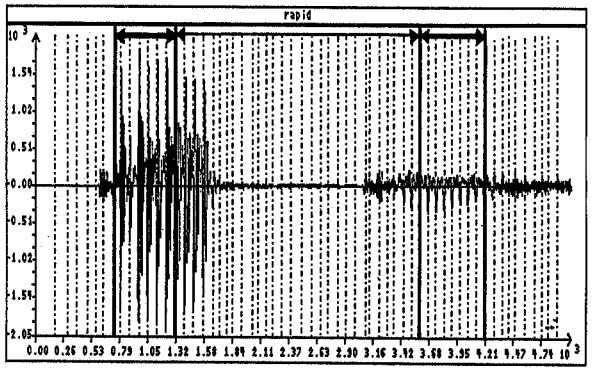


Schéma 1

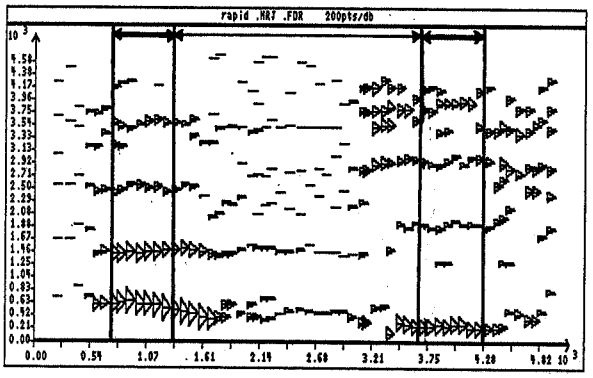


Schéma 2

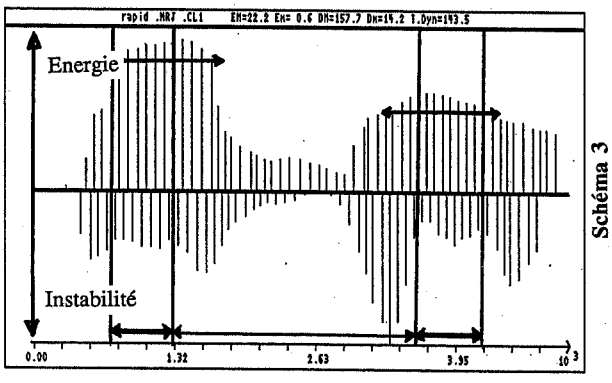


Schéma 3

APPLICATION DU "FORMANT EFFECTIF" F'2 A LA CLASSIFICATION DES VOYELLES ANTERIEURES DU FRANCAIS

Marios MANTAKAS, Jean-Luc SCHWARTZ, Pierre ESCUDIER

Laboratoire de la Communication Parlée
I.C.P. - Unité Associée au C.N.R.S., I.N.P.G. - E.N.S.E.R.G.
46 avenue Félix Viallet, 38031 GRENOBLE CEDEX

ABSTRACT

We test the "effective formant" F'2 as a classification parameter for the rounding opposition in French vowels (natural speech). For this purpose, we measure the first five formant frequencies and levels of a corpus of vowels [i, y, e, ø] spoken by 3 men and 2 women. We next estimate F'2 from these acoustic data. We report several acoustic aspects of the opposition [i] vs. [y] and [e] vs. [ø]. We discuss some shortcommings of our F'2 model [1], comparing it breefly with a spectral integration process [6]. We finally evaluate the classifying power of F'2.

1. INTRODUCTION

Dans cette étude, continuation de [1], nous testons le pouvoir classificateur du "formant effectif" F'2 dans la parole naturelle. Nous nous intéressons aux voyelles antérieures, pour lequelles F'2 est significativement supérieur à F_2 et nous étudions l'apport de ce paramètre à la séparation des deux classes de voyelles antérieures du Français, non arrondies et arrondies, à savoir la classification [i]-[y], [e]-[ø], [ɛ]-[œ].

A cette fin, nous analysons acoustiquement un corpus des voyelles [i,y,e,ø] (mesure des formants). A partir de ces données acoustiques, nous estimons F'2 en utilisant le modèle proposé dans [1].

2. REMARQUES SUR LE MODELE DE F'2

Le modèle de F'2 (voir [1]) a été fondé sur les données des expériences d'ajustement (matching) [2] et [3].

Il repose sur des données formantiques: fréquences et niveaux de F2, F3 et F4. F'2 est calculé comme un "centre de gravité" des fréquences F_2, F_3, F_4 pondérées par les niveaux respectives L_2, L_3, L_4 (idée principale du modèle de Bladon et Fant [3]), si ces formants sont regroupables au sens de la "distance critique" d'environ 3 Bark [4], selon la suggestion de Bladon [5].

Bien que le modèle utilise les formants comme paramètres d'entrée, leur extraction par le système auditif ne fait pas nécessairement partie de nos hypothèses sur F'2 [1]. Une étude préliminaire [6] a montré qu'un traitement global du spectre par intégration large-bande (fenêtre de largeur 3-3.5 Bark glissée sur l'axe des fréquences) suivie d'une extraction de maximum peut rendre compte des résultats d'expériences d'ajustement de F'2 pour les voyelles antérieures. La mise au point d'un tel traitement et l'estimation automatique de F'2 ainsi envisagée fera l'objet de la prochaine étape de nos travaux.

Notre formule actuelle [1] pose a priori trois problèmes principaux:

a/ Le second formant y joue un rôle central. F'2 est d'habitude considéré comme F2 plus une "correction" positive rendant compte des formants supérieurs. Est-ce correct pour les voyelles aiguës et ouvertes ? ([i] en particulier). Nous donnons quelques éléments de reponse dans les § 3 et 4.

b/ En ce qui concerne les niveaux des formants, les expériences de type "centre de gravité" mettent en évidence le rôle des nivaux relatifs. Des expériences d'identification des voyelles donnent de résultats contradictoires quant au rôle des niveaux (voir [7]). Mais le rôle des niveaux pour le calcul de F'2 n'a jamais été directement étudié. Dans notre modèle nous adoptons l'idée de Bladon et Fant [3] (voir ci-dessus) sans vérification expérimentale.

c/ Les prédictions de F'2 présentées dans les tables 1 et 2 de [1] montrent une surestimation pour [ø] dans les expériences de Carlson et al. (1970) et [ʉ, ɨ, ø] dans les expériences de Bladon et Fant (1978), malgré la diminution du poids de F3 dans le modèle ("correction" empirique de niveau par le coefficient a3). Cette situation pose un problème de fond actuellement non resolu: Pourquoi dans cette zône du triangle vocalique la proximité de F3 à F2 (F3-F2 inférieur à 3 Bark) n'élève-t-elle pas la valeur de F'2 ?

3. ANALYSE ACOUSTIQUE D'UN CORPUS

Le corpus que nous avons analysé acoustiquement (mesures des formants) a été élaboré et utilisé par Abry et al. [8] dans une étude articulatoire de la labialité vocalique et consonantique. Il comprend les voyelles [i,y,e,ø] précédées par les consonnes [s,z,ʃ,ʒ] en position finale (donc accentuée) et ouverte. Les 48 phrases du corpus ont été prononcées une fois par 5 locuteurs, 3 hommes (2: J-L.C., 3: D.A., 5: P.C.) et 2 femmes (1: D.L., 4: F.E.).

Nous avons mesuré "manuellement" la fréquence et le niveau des 5 premiers formants sur le spectre lissé par cepstre. L'étiquetage des formants a été effectué selon les propositions de C.Abry [9].

Dans la suite, nous discutons le rôle des paramètres formantiques dans la réalisation des oppositions [i]-[y] et [e]-[ø].

Opposition [i]-[y]

Le principal corrélat acoustique pour le passage de [y] à [i] pour les locuteurs 2, 3 et 5 est la forte hausse de F_3, illustrée par la figure 1. Pour ces locuteurs la tendance consisterait donc à passer de [y] à [i] par une bascule de la "masse" F2-F3 à la "masse" F3-F4. Pour la locutrice 4, pour laquelle la hausse de F_2 est cette fois importante, la tendance est la même. Par contre, dans le cas de la locutrice 1, l'opposition serait réalisée par un transfert en bloc de la "masse" F2-F3 vers les hautes fréquences (figure 2). On remarque sur cette figure, dans le cas de [i], le manque de F4 et la valeur très élévée de F_2.

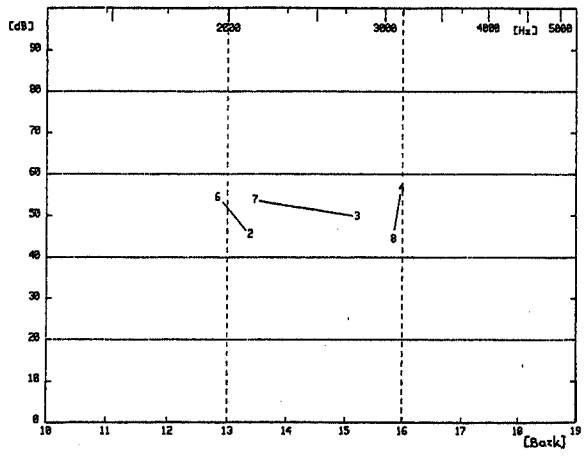


figure 1
Locuteur 5 (P.C.)
Stratégie acoustique pour l'opposition [i]-[y]
Niveaux (dB) vs. fréquences (en Bark) de F2, F3, F4
Symboles: 2,3,4 moyennes (en niveau et en fréquence) de F2, F3 et F4 pour [i]. 6,7,8, idem pour [y]

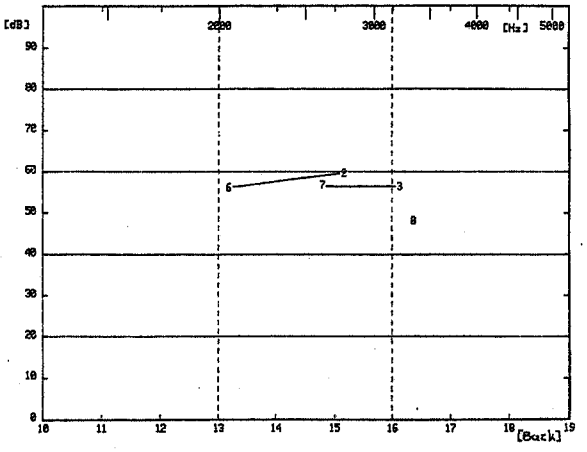


figure 2
Locuteur 1 (D.L.)
Stratégie acoustique pour l'opposition [i]-[y]
(Symboles: voir figure 1)

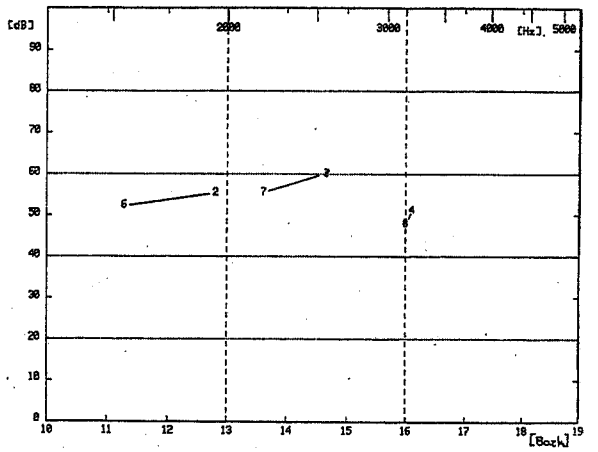


figure 3
Locuteur 5 (P.C.)
Stratégie acoustique pour l'opposition [e]-[ø]
Niveaux (dB) vs. fréquences (en Bark) de F2, F3, F4
Symboles: 2,3,4 moyennes (en niveau et en fréquence) de F2, F3 et F4 pour [e]. 6,7,8, idem pour [ø]

Pour les locuteurs 3 et 5, L_2 est très bas par rapport à L_3 et L_4. Dans ces cas, F2 ne semble pas être un paramètre pertinent pour la description phonétique et la perception.

En conséquence, les voyelles [y] et [i] de ce corpus présentent une proéminence spectrale respectivement vers 2000 Hz et 3000 Hz.

Opposition [e]-[ø]

Le passage de [e] à [ø] est caractérisé par une baisse surtout de F_2, mais aussi de F_3 et, en général, une baisse de L_2 et de L_3 (figure 3). Les seuls cas de proximité de formants sont ceux de F2/F3 pour les locutrices 4 et surtout 1. Il nous paraît, ainsi, difficile de décrire l'opposition [e]-[ø] en termes d'une (et une seule) forte proéminence spectrale dans la plage de 10 à 18 Bark, comme c'était le cas pour [i] et [y].

4. ESTIMATION DE F'2

Le problème majeur pour l'adaptation de nos données formantiques au modèle concerne les niveaux. Le modèle demande des niveaux absolus, mais nous ne disposons que de mesures de niveau relatif. Nous avons constaté que a) les mesures des niveaux avaient des valeurs réalistes pour la parole naturelle et b) un large décalage de ces valeurs sur l'échelle des niveaux n'affectait pas essentiellement les résultats. Nous avons, donc, directement utilisé les valeurs mesurées.

Les paramètres du modèle ont les valeurs proposées dans [1]: distance critique de 3-3.5 Bark et coefficients de correction empirique des niveaux a3= 0.7 et a4= 0.6.

Nous présentons les résultats de l'estimation de F'2 sur les figures 4 (pour les 5 locuteurs, classes [i], [y]) et 5 (classes [e], [ø])

Trois problèmes principaux des estimations apparaissent sur ces figures:

1/ Les valeurs de F'2 pour [y] pour la locutrice 1 sont dispersées et certaines sont très élévées (figure 4). La cause de ce phénomène est illustrée par la figure 6. F4 est "intégrable" avec F2 et F3 au sens de la distance critique à une exception près. F3 est tantôt plus proche de F2 et tantôt de F4, ce qui provoque un fonctionnement "bimodal" du modèle (voir [1]). Ainsi, notre idée de départ, à savoir décrire [y] (par opposition à [i]) par le fait que F_3 est plus proche de F_2 que de F_4, s'avère quantitativement insatisfaisante.

C'est le **rôle des niveaux** qui apparait ici également important. En effet, L_4 est dans ce cas nettement inférieur à L_2 et L_3, et le modèle ne tient pas compte de ce fait. Toutefois, un modèle à intégration spectrale large-bande [6] ne présenterait pas cette faiblesse et ne donnerait pas des valeurs de F'2 supérieures à 14.5 Bark.

2/ Les valeurs de F'2 pour [i] pour le locuteur 3 sont divisées en deux groupes autour de 14 et 15.5 Bark (figure 4). Cela provient de la dispersion de la distance F_4-F_2 aux alentours de 3.5 Bark (situation illustrée par la figure 7). Dans le cas où F_4-F_2 dépasse les 3.5 Bark, le quatrième formant ne contribue pas au calcul de F'2, ce qui donne le groupe des valeurs basses.

Malgré le niveau très bas de F2 par rapport à L_3 et L_4 (§3), notre modèle, s'ancre sur le deuxième formant et ne peut pas rendre compte de la proéminence spectrale vers 15-16 Bark créée par F3 et F4. Une fois encore, notre modèle sous-estime le rôle des niveaux et un modèle à intégration large-bande fonctionnerait sans doute beaucoup mieux.

Pour ce locuteur, comme pour le locuteur 5, F'2 ne peut pas être imaginé comme F2 + contribution des formants supérieurs.

Il est ici intéressant de critiquer la figure 2 de [1]. Pour éviter les deux problèmes précédents, nous avions diminué dans cette simulation la valeur de la distance critique : 3-3.3 au lieu de 3-3.5 Bark. F4 n'intervenait alors pas dans le calcul, ce qui donnait des valeurs de F'2 trop basses pour [i].

3/ Considérons maintenant les estimations de F'2 pour [ø] pour les 5 locuteurs (figure 5). Nous visualisons sur la figure 8 les valeurs en fréquence du deuxième formant pour les voyelles [e,ø] pour tous les locuteurs. Pour les trois hommes (locuteurs 2, 3 et 5), ces valeurs tombent dans la zône fréquentielle de F2, pour laquelle F'2 est à peine supérieur à F_2, selon les expériences d'ajustement des études [2] et [3], alors que notre modèle effectue des surestimations (§2).

Les deux locutrices (1 et 4) ont des valeurs de F2 supérieures à celles que l'on obtient pour les locuteurs masculins de 0.8 Bark environ, se recouvrant avec les valeurs de F2 pour [e] pour les hommes. Le fondamental peut-il normaliser (d'une part la zône fréquentielle de surestimation et d'autre part la classification [e]-[ø], voir suite) ? Est-ce que l'augmentation de fondamental peut engendrer un passage de la perception de [ø] à celle de [e] ? (cf [10]).

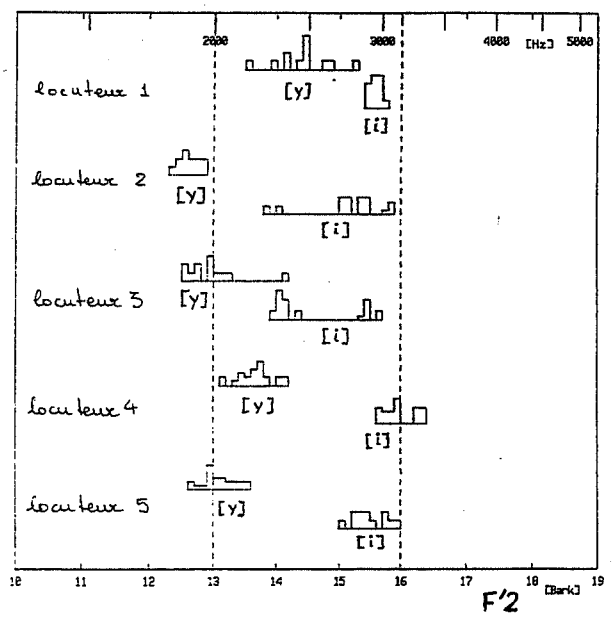


figure 4
Résultats d'estimation de F'2 pour [i] et [y] pour les cinq locuteurs (abscisse linéaire en Bark).

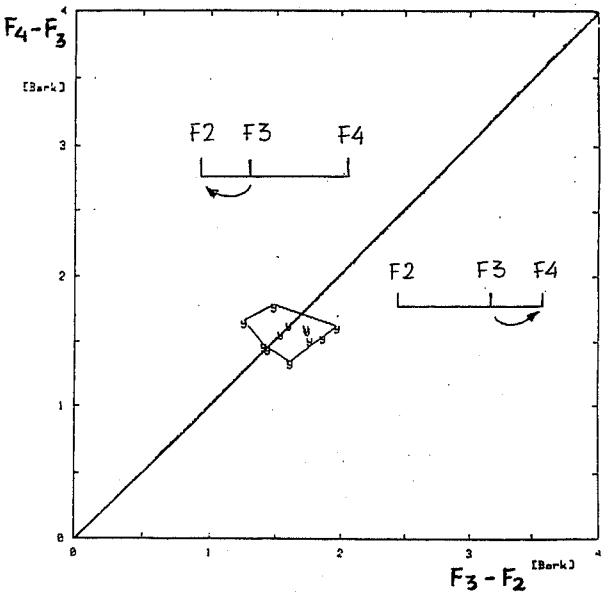


figure 6
Locuteur 1 (D.L.)
Représentation de la voyelle [y] dans l'espace F_4-F_3 vs. F_3-F_2 (axes en Bark)

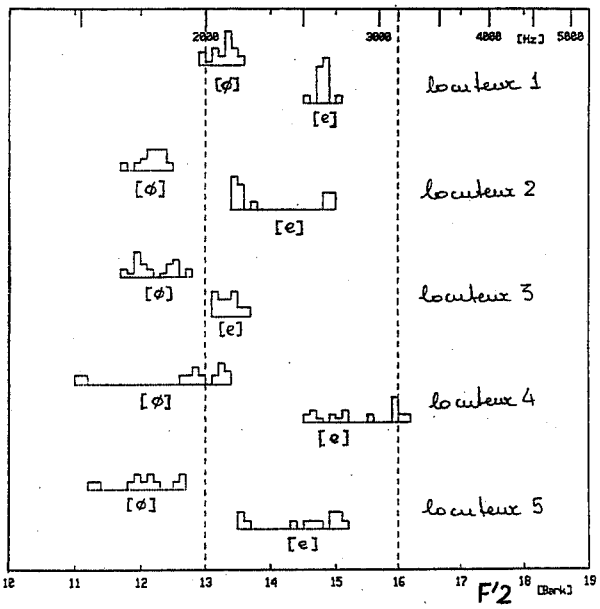


figure 5
Résultats d'estimation de F'2 pour [e] et [ø] pour les cinq locuteurs (abscisse linéaire en Bark).

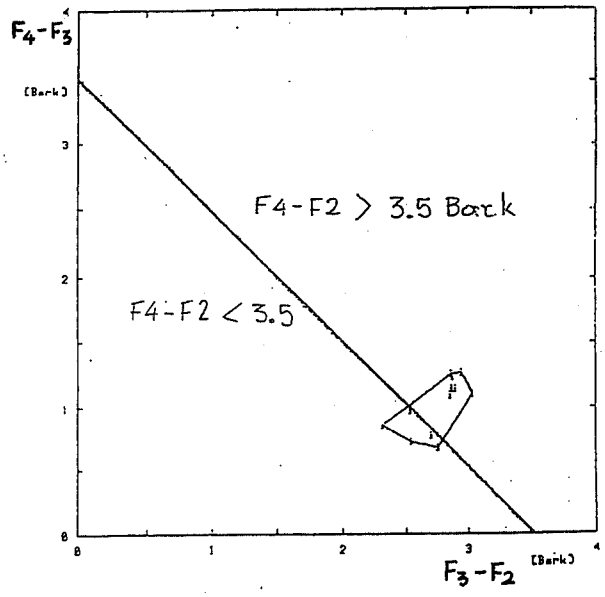


figure 7
Locuteur 3 (D.A.)
Représentation de la voyelle [i] dans l'espace F_4-F_3 vs. F_3-F_2 (axes en Bark)

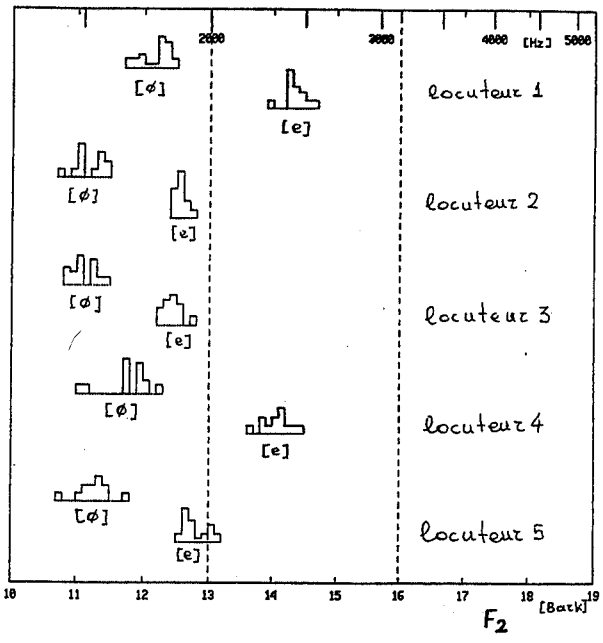


figure 8
Fréquences de F2 pour [e] et [ø] pour les cinq locuteurs (abscisse linéaire en Bark).

5. POUVOIR CLASSIFICATEUR DE F'2

Opposition [i]-[y]

En se rappelant que la valeur de F'2 calculée par un modèle à intégration large bande ne dépasserait pas 14.5 Bark pour [y] pour la locutrice 1 et serait aux alentours de 15.5 Bark pour [i] pour le locuteur 3, nous pouvons dire que la classification donne des résultats intéressants. De plus, la séparation est normalisatrice, la frontière commune pour les 5 locuteurs étant à 14.5-15 Bark environ.

Il est intéressant de remarquer que le paramètre F'2-F0, F0 étant le fondamental, pourrait être aussi normalisateur (cf. [11])

Il faut souligner que, par contraste, les classes [i] et [y] se recouvrent légèrement selon la dimension F2 pour les locuteurs 3, 4 et 5.

Opposition [e]-[ø]

F'2 sépare les voyelles [e], [ø] à 100%. Il est normalisateur séparement pour les hommes (frontière 13 Bark) et pour les femmes (frontière aux environs de 14 Bark). Ainsi, le paramètre F'2-F0 est très probablement normalisateur pour tous les 5 locuteurs.

REMERCIEMENTS

Ont participé / aidé à cette étude: MM. C.Abry (discussion sur l'étiquetage des formants et la production de [i]), L-J.Boë (fourniture du corpus et de divers logiciels, numérisation du signal, discussion sur la production de [i]), P.Badin (aide en informatique, discussions sur la méthode d'analyse acoustique et les modèles articulatoires), G.Murillo (aide en informatique, discussions sur la méthode d'analyse), P.Perrier (explications sur le logiciel pour les mesures), J.Caelen (discussion sur les méthodes d'analyse acoustique). Cette étude a été soutenue par le C.N.E.T. Lannion (convention GR 705159).

REFERENCES

[1] **MANTAKAS M., SCHWARTZ J.L., ESCUDIER P. (1986)** "Modèle de prédiction du deuxième formant effectif F'2 et application à l'étude de la labialité des voyelles avant du Français" Actes des 15e JEP du GALF, pp. 157-161, AIX EN PROVENCE

[2] **CARLSON R., GRANSTRÖM B., FANT G. (1970)** "Some studies concerning perception of isolated vowels" STL-QPSR 2-3/1970, pp.19-33

[3] **BLADON R.A.W., FANT G. (1978)** "A two-formant model and the cardinal vowels" STL-QPSR 1/1978,pp.1-8

[4] **CHISTOVICH L.A., SHEIKIN R.L., LUBLINSKAYA V.V. (1979)** "'Centers of Gravity' and Spectral Peaks as the Determinants of Vowel Quality" in Frontiers of Speech Communication Research, B.Lindblom and S.Ohman eds, Academic Press, London, pp. 143-158

[5] **BLADON A. (1983)** "Two-formant models of vowel perception : Shortcommings and enhancements" Speech Comm. 2, pp. 305-313

[6] **ESCUDIER P., SCHWARTZ J.L., BOULOGNE M. (1985)** "Perception de voyelles stationnaires: représentation interne des formants dans le système auditif et modèles à deux formants" Actes du Symposium Franco-Suédois sur la Parole, pp. 143-174, Guérin et Carré eds, GRENOBLE.

[7] **SCHWARTZ J-L., ESCUDIER P. (1986)** "Does human auditory system include large-scale integration?" présenté à "The Psychophysics of Speech Perception", NATO Advanced Research Workshop, UTRECHT, juillet 1986

[8] **ABRY C., BOË L.J., CORSI P., DESCOUT R., GENTIL M., GRAILLOT P. (1980a)** "Labialité et phonétique, données fondamentales et études expérimentales sur la géométrie et la motricité labiale" Publication de l'Université des Langues et Lettres de Grenoble.

[9] **ABRY C. (1987)** Communication personnelle sur l'étiquetage des formants, Institut de Phonétique de Grenoble

[10] **FANT G., CARLSON R., GRANSTRÖM B. (1974)** "The [e]-[ø] ambiguity" Speech Communication Seminar, Stockholm, pp.117-118

[11] **TRAUNMÜLLER H. (1981)** "Perceptual dimension of openness in vowels", J.Acoust.Soc.Am.,Vol.69, pp.1465-1475

PHYSIOLOGIE

ETUDE AERODYNAMIQUE DES CONSONNES FRANCAISES
VALEURS DE REFERENCE ET PROFILS CARACTERISTIQUES

Denis AUTESSERRE et Bernard TESTON

INSTITUT DE PHONETIQUE UA 261 CNRS AIX-EN-PROVENCE

ABSTRACT:
We attempt to estimate the mean values for French consonnants of the following aérodynamic parameters:Oral and nasal airflow,the corresponding air volume,and the intra oral air pressure.This research is a step towards the development of an integrated aerodynamic theory of speech production.We also hope to define reference values for diagnostic aids and for functional rehabilitation.

INTRODUCTION:

Les variations du débit d'air à l'interieur du conduit vocal,lors de la production de la parole,réflètent les changements de mode d'articulation des consonnes et des voyelles émises.On comprend donc tout l'intérêt que suscite leur étude,depuis bientot un siècle,d'abord chez les spécialistes de Phonétique physiologique,puis parmis les cliniciens.L'achèvement récent,grace à un contrat de l'INSERM (génie biologique et médical),de deux prototypes du Poliphonomètre III,destinés à la mesure des paramètres aérodynamiques,nous a conduit à entreprendre une étude systèmatique des tracés de débits d'air caractéristiques des consonnes et des voyelles du Français,préalable à l'interprétation des données cliniques.En effet le second prototype se trouve actuellement au service d'exploration neurologique fonctionnelle de l'hopital de la Salpètrière à Paris (Dr Claude CHEVRIE MULLER).Il convient de fournir aux cliniciens disposant du même type d'instrument que notre laboratoire,un ensemble de valeurs de débits d'air buccal de référence par rapport auquel pourront être appréciées les déviations pathologiques.Nous présentons les premiers résultats de ces investigations menées sur un échantillon de population normale.

PROCEDURE EXPERIMENTALE:

1-Le corpus:

Un soin tout particulier a été apporté à la constitution du corpus destiné à mettre en évidence les différents profils caractéristiques des consonnes et des voyelles du Français.Il comprend 40 phrases comprenant chacune deux à trois groupes accentuels.Les consonnes de même mode articulatoire sont toujours placées en position intervocalique (contextes symétriques /a/,/i/,/u/,ou symétriques /a-i/,/a-u/),et regroupées à l'interieur d'un même énoncé.Ainsi les oclusives non voisées, /k/,/t/,/p/,,apparaissent dans la séquence "Léa l'attaqua,le rata puis l'attrapa".Le corpus est prononcé dix fois par un même locuteur au cours de trois ou quatre séances situées à plusieurs jours d'intervalle.Il porte sur dix locuteurs (cinq femmes et cinq hommes),ce qui représente quatre milles réalisations.

2-Appareillage et paramètres enregistés:

Deux enregistrements sont conservés simultanément sur les deux pistes d'un magnétophone REVOX:Le phonogramme buccal,et les vibrations au larynx,captées au moyen d'un laryngophone.Le Vu-Mètre du phonogramme buccal,est utilisé comme controle du niveau d'émission acoustique par les sujets.Une transcription phonétique,effectuée à partir de ces enregistrements rend possible la délimitation temporelle des unités phoniques successives.Les signaux aérodynamiques sont recueillis séparément à la sortie des orifices de la bouche et du nez.Ainsi,le débit d'air buccal (DAB) et le volume d'air buccal (VAB)sont captés à l'aide d'une embouchure buccale spécialement adaptée la morphologie faciale du sujet,ceci pour maitriser les fuites d'air tout en ne contraignant pas le locuteur,pour ne pas gener les mouvements des lèvres et l'ouverture inter maxillaires.Deux embouts introduits dans chaque narine,permettent l'enregistrement du débit d'air nasal (DAN),et du volume d'air nasal (VAN).On note la sensibilité particulière du capteur nasal:le tracé de DAN met en évidence des micro-phénomènes affectant l'écoulement de l'air nasal au moment des phases d'implosion et d'explosion des consonnes oclusives.

Ces différents tracés sont enregistés sur un oscillographe à jet d'encre SIEMENS Oscillomink à dix canaux,à la vitesse de défilement de 250 mms par seconde.Sur cinq réalisations du corpus,les paramètres suivants sont enregistés:Vibrations au larynx,Phonogramme buccal,Débit d'air buccal,Volume d'air buccal expiré,Phonogramme nasal,Débit d'air nasal, Volume d'air nasal expiré.Sur les cinq réalisations restantes,les paramètres suivants sont enregistrés:Vibrations au larynx,Phonogramme buccal,Intensité acoustique du phonogramme buccal,Débit d'air buccal,Volume d'air buccal, expiré,Pession intra orale (PIO).La dynamique d'enregistrement des divers paramètres aérodynamiques est maintenue constante lors des dix enregistrements.

3-Dépouillement des données et mesures:

Les mesures sont effectuées manuellement à partir des tracés oscillographiques.Nous limitons notre étude à la mesure du débit d'air buccal en millilitres par seconde.La segmentation du tracé du débit d'air buccal est faite,non pas en projetant sur lui une segmentation à partir des indices acoustiques des différents phonogrammes,mais directement sur le tracé lui même,en tenant compte de ses variations morphologiques (les phonogrammes permettent cependant un repèrage des macroclasses vocaliques et consonantiques).On peut ainsi localiser six points remarquables:

-1:Correspond à la valeur la plus basse du.DAB situé sur la partie vocalique qui précède la consonne.

-2:Le sommet du DAB à la partie finale de la voyelle.

-3:Le point le plus bas du tracé du DAB,dans la première partie de la consonne;point de rencontre de la chute du DAB et de la tangente à l'occlusion.

-3':Le deuxième point le plus bas avant la remonté du DAB ;intersection avec la tangente à la pente de l'occlusion.

-4:Le sommet du DAB avant le début de la voyelle subséquente.

-5:La valeur la plus basse atteinte sur la voyelle.

L'attribution d'un numéro 3' à la fin de l'occlusion facilite la comparaison avec les consonnes constrictives.Elle permet en outre, de rendre compte des différents degrès de modification contextuelle des consonnes occlusives (lénition),et de rechercher des seuils ou l'on passe de la perception d'une obstruante occlusive à une obstruante constrictive. De même,il est attribué des numéros indicés, type 4' et 4",à des modifications morphologiques de l'explosion de la cosonne:montée suivie d'une dépression avant une nouvelle montée,puis une chute sur la voyelle (ce phénomène se produit préférentiellement lors de certaines réalisations de la consonne /p/.A partir des points ainsi repérés on délimite des segments de 1 à 4,dont on calcule la durée (figure 1).Les voyelles ont une morphologie plus hétérogène:elles combinent des segments influencés par les consonnes adjacentes.Elles peuvent être difficilement délimitables,au seul niveau aérodynamique,dans certains contextes (consonnes vocaliques,par exemple).Cette segmentation du débit d'air buccal,est validée par comparaison avec le tracé de la PIO:la correspondance peut être établie avec les points d'inflexion des tracés de PIO (figure 2).

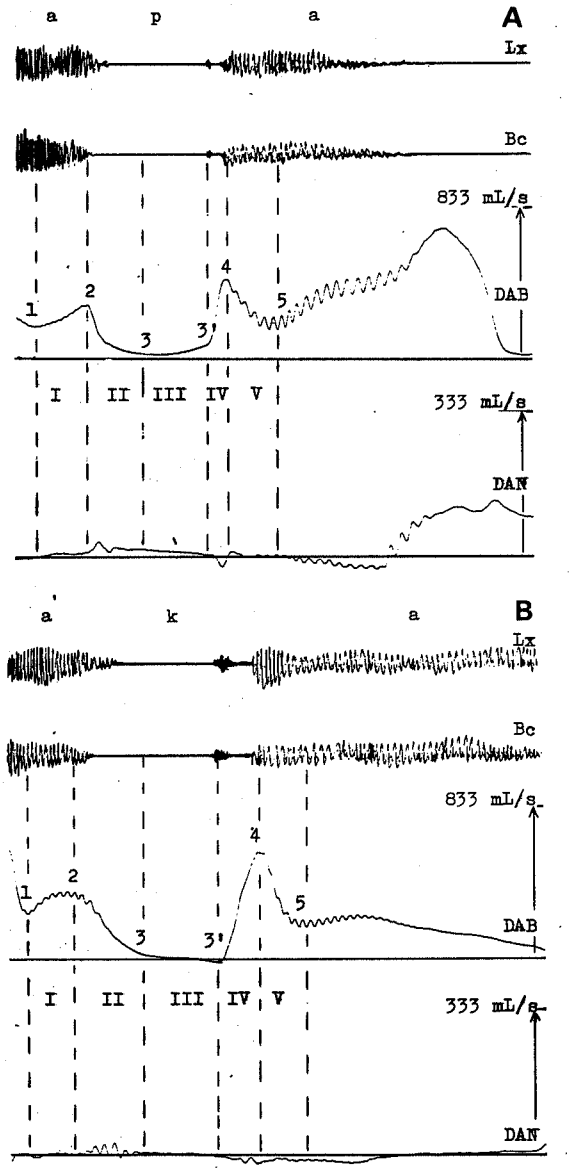


Figure 1 A et B:Consonnes occlusives /p/ et /t/ en contexte vocalique symétrique /a/. Points remarquables pour le calcul des amplitudes de DAB et segments correspondants pour les mesures de durée.

RESULTATS:

1-Malgré les multiples précautions prises, les amplitudes des différents points des tracés du débit d'air buccal pour les dix réalisations d'un même locuteur apparaissent assez dispersées. Les valeurs des écarts-types sont toujours importantes.

2-Les valeurs plus homogènes obtenues pour la durée des segments, attestent que les fortes variabilités de débit ne sont pas en relation avec de fortes variations de la vitesse d'énonciation.

3-Des amplitudes différentes sont associées à des courbes d'intensité parfaitement comparables: on ne peut minimiser des variations d'intensité de la voix.

4-Si l'on compare les dix réalisations de chacune des trois consonnes occlusives /p/, /t/,

NUMERO DE SERIE DES SEQUENCES [a p a] MESURES	POINTS DE REFERENCE POUR LES MESURES DE D. A. B. (en mL/s)							
	1	2	3	3'	4	4'	4''	5
1	168	288	24	72	432			192
2	334	363	0	24	216	12	360	240
3	145	203	-29	-29	305	189	247	160
4	145	218	0	0	174	160	203	145
5	242	299	-12	12	552			288
6	58	102	0	15	421			189
7	131	189	0	15	522			218
8	116	218	-15	0	493			261
9	107	178	0	0	195			-89
10	178	178	-18	0	195			89
MOYENNE	162	223	-9,8	16,7	367			169
ECART-TYPE	73	70	10,7	20,9	130			94

NUMERO DE SERIE DES SEQUENCES [a k a] MESURES	POINTS DE REFERENCE POUR LES MESURES DE D. A. B. (en mL/s)					
	1	2	3	3'	4	5
1	252	348	24	-12	576	216
2	360	372	48	-36	528	168
3	131	290	-15	-58	392	102
4	261	377	15	-15	493	160
5	322	357	12	0	518	173
6	145	305	15	0	377	131
7	232	319	29	0	406	174
8	174	348	0	-15	435	160
9	266	355	0	-36	604	124
10	337	337	0	-18	515	142
MOYENNE	248	340,8	12,8	-19	484,4	155
ECART-TYPE	75	26,8	15,6	18	74,7	30

Valeurs des amplitudes de DAB pour dix réalisations des consonnes occlusives /p/ et /k/ en contexte vocalique symétrique /a/, chez un même locuteur (D.A).

NUMERO DE SERIE DES SEQUENCES [a p a] MESURES	SEGMENTS DE REFERENCE POUR LES MESURES DE DUREES (en ms)				
	I	II	III	IV	V
1	46	50	60 / 128	18	46
2	46	54	64 / 156	38	26
3	32	64	58 / 148	26	26
4	34	72	50 / 146	24	24
5	42	78	50 / 148	20	36
6	38	60	54 / 132	18	30
7	28	52	68 / 140	20	30
8	38	42	74 / 140	24	32
9	46	58	72 / 140	10	14
10	28	50	80 / 146	16	14
MOYENNE	37,8	58	63 / 142	21,4	27,8
ECART-TYPE	6,7	10,3	9,8	7	9

NUMERO DE SERIE DES SEQUENCES [a k a] MESURES	SEGMENTS DE REFERENCE POUR LES MESURES DE DUREES (en ms)				
	I	II	III	IV	V
1	42	62	68 / 168	38	42
2	54	56	74 / 168	38	34
3	46	70	54 / 174	46	36
4	28	84	52 / 174	38	38
5	46	62	60 / 162	40	32
6	48	58	52 / 142	32	40
7	56	52	52 / 140	36	30
8	46	54	50 / 142	38	28
9	54	48	76 / 160	36	34
10	46	50	66 / 150	34	36
MOYENNE	46,6	59,6	60,4 / 158	37,6	35
ECART-TYPE	7,5	10,2	9,3	3,5	4,1

Durées des différentes portions (1 à 5) du tracé de DAB délimitées par les points de référence (1 à 5), pour les mêmes consonnes occlusives /p/ et /k/, chez un même locuteur .

/k/,on remarque que les modifications d'amplitude des tracés de DAB ne remettent pas en cause la morphologie particulière du tracé propre à chaque consonne.Plutot que de se limiter à une étude de l'amplitude des pics de DAB,il convient peut être

DAB,il convient,peut-être,de s'interesser à la permanence de certaines formes caractéristiques dont on pourrait modéliser la variabilité.

DISCUSSION:

1-La dispersion des valeurs de DAB engage à une certaine prudence dans l'interprétation de ces données.Elle conduit aussi à prendre quelques précautions au moment de l'examen clinique:il convient de prévoir un certain nombre de répétitions,pour savoir si la dispersion observée est de même type que celle que l'on a dégagée dans l'échantillon de population normale.

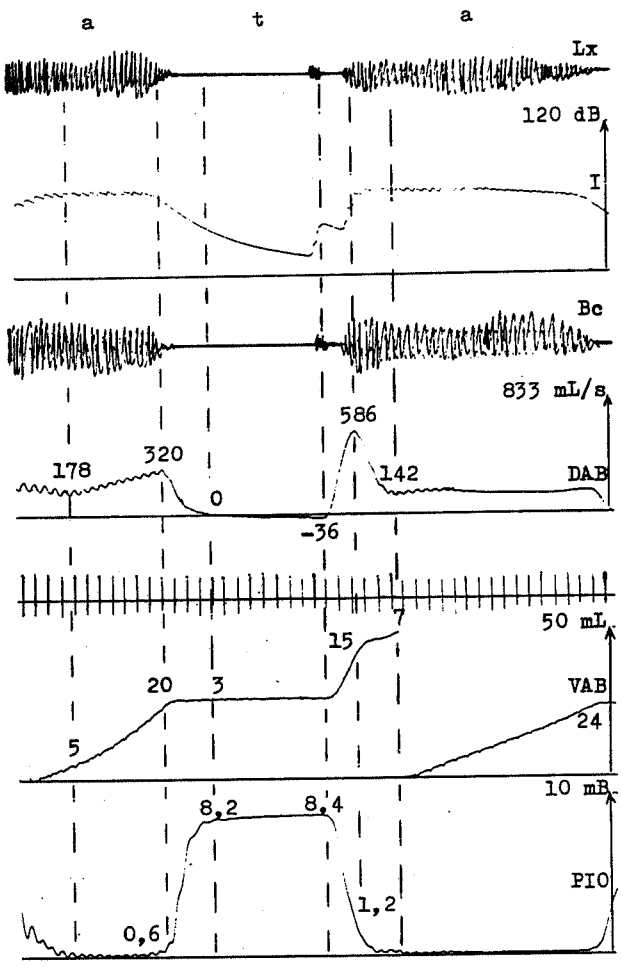
2-Les résultats démontrent la nécessité de standardiser ce genre d'investigation (comme le souligne à juste raison DEJONCKERE et Al).

3-Dans les cas douteux,il convient d'enregistrer simultanément d'autre paramètres dont les conditions de variations particulières, donnent une signification à des faits qui ont pu passer inaperçus:confrontation des tracés de DAB et de DAN,ou DAB et PIO(figure 2).

CONCLUSION:

La dispersion observée des paramètres aérodynamiques de débit d'air buccal,dans des conditions satisfaisantes de controle des variables,semble être en relation avec des variations de degrés d'attention au message difficilement maitrisables.Cependant ces faits ne doivent pas être perçus négativement:ils méritent même d'être exploités positivement dans des approches thérapeutiques fondées sur le controle rétroactif biologique.Dans ce cas la reproduction de formes modélisées à partir des profils types,dégagés pour chaque consonne,apparait plus satisfaisant qu'un simple controle artificiel des pics d'amplitude qui trés vite aboutir à la réalisation de shémas stéréotypés..

Nous n'avons dépouillé pour l'instant que les réalisations individuelles.Nous allons terminer dans un avenir proche,l'acquisition du corpus dans sa totalité (qui représente environ cent milles données).Nous allons pouvoir ainsi traiter au moyen de toutes les méthodes de traitement statistiques que nous avons à notre disposition pour aller "à la pèche" aux informations contenues dans ces données.



Valeurs de trois paramètres aérodynamiques; débit d'air buccal (DAB),volume d'air buccal (VAB) et pression intra-orale (PIO),pour la réalisation de la consonne occlusive /t/ dans la séquence /ata/(niveau d'intensité constant I).

Figure 2

BIBLIOGRAPHIE:

BAKEN,R.J.Cinical measurement of speech and voice,Taylor and Francis,London,518 p,1987.

-BAKEN,R.J.,Clinical measurement of speech and voice,Tailor and Francis,London,518p,1987.

-DEJONCKERE,P.H.,GREINDL,M. et SNEPPE,R.,"Débimètrie aérienne à paramètres phonatoires standardisés",Folia Phoniatrica,37,58-65,1985.

-TESTON,B. et AUTESSERRE,D.,"Réalisation d'une unité d'analyse polyphonomètrique"C.L.O.S. 5-6,Hommage à Georges MOUNIN,415-437,1975.

-TESTON,B."A system for the analysis of the aérodynamic parameters of speech:The Polyphonometer model III",10° Int. Cong. Phon. sc. Utrecht,Sec.5,457,1983.

SYNTHESE ARTICULATOIRE DYNAMIQUE DES TRANSITIONS MAXIMALES VOCALIQUES

MAHMOUD HELAL & LOUIS-JEAN BOE

Institut de la Communication Parlée
Institut de Phonétique de Grenoble
Université III
38400 St Martin d'Hères.

ABSTRACT

This paper proposes a method in synthesis of vocalic maximal transitions with the help of a temporal simulation model of the vocal tract . The model adopted , (MAEDA, 1982) , uses an area function in the input , with the possibility of varying at every moment the length and dimension of each section . To generate vocalic transitions between the three extreme cardinal vowels [i, a, u] , we used, within these targets , two types of interpolations . Results are given acoustically and controlled perceptually.

INTRODUCTION

Ce travail se situe dans le cadre d'une étude à long terme visant à commander un modèle de simulation utilisant des paramètres de type articulateur (mâchoire, langue, lèvres) . Nous avons d'ailleurs choisi le modèle articulatoire de MAEDA [2] dont la validité a déjà été testée [3] . En fait dans l'état actuel , nous ne disposons pas de données temporelles permettant de générer , ne serait-ce que des transitions vocaliques : les paramètres de commande de la langue étant pour le moins difficiles à évaluer (on s'oriente à la fois vers des mesures ultra - sons KELLER [4] et des procèdures d'inversion du modèle CHARPENTIER [5]) . Cette étude préliminaire est donc menée sur un ensemble de paramètres plus faciles à induire , les fonctions d'aires , qui ne nécessitent pas de modèle sous - jacent . Nous avons choisi de générer des transitions vocaliques ; ici seront présentés des résultats concernant les voyelles cardinales extrêmes [i, a, u] . Notre approche va consister à interpoler entre des " cibles fonctions d'aires " , linéairement ou en utilisant une évolution sinusoïdale (la réponse d'un système du second ordre non amorti à un échelon de commande) . A partir d'une excitation glottique, dont on peut fixer la pression de l'onde de débit et l'évolution de Fo le modèle dynamique proposé par MAEDA [1] permet de générer un signal acoustique pour une fonction d'aire décrite au cours du temps par une succession de tubes élémentaires de longueur et de section variables. Le modèle de simulation (fig 1) inclut les pertes et permet de brancher les cavités nasales si besoin . La figure 2 montre comment on approxime le conduit vocal (avec ou non les cavités nasales) par une concaténation d'un nombre fini de q quadripôles type T : chacun d'entre eux représente un tube élémentaire cylindrique de longueur li et de section Ai (avec i= 1, 2, ..., q) [1, 6] . Les quantités li et Ai sont considérées comme les paramètres qui décrivent la configuration géométrique du conduit vocal au cours du temps .

SIMULATION TEMPORELLE

Pour notre simulation nous avons adopté deux méthodes d'interpolation (linéaire et non linéaire) entre les aires pour générer des transitions maximales voyelle-voyelle . Nous avons choisi des fonctions d'aires des voyelle du français proposées par BOE (1973) et affinées par FENG (1986) . Pour chaque son , la durée totale a été fixée à environ 275 ms (60 ms pour chaque cible vocalique et environ 155 ms pour la transition) . La fréquence fondamentale Fo présente une évolution (fig 3) montante pendant 50 ms et descendante jusqu'à la fin du stimulus . Nous avons choisi une fréquence de calcul permettant de générer un signal échantillonné à 20 kHz . Dans ce cas la distorsion spectrale devient pratiquement négligeable en dessous de 4 kHz [1] .

1-INTERPOLATION LINEAIRE

Dans un premier temps et pour tester le programme , nous avons généré les transitions en interpolant chaque section (aire et longueur) linéairement entre les 2 cibles vocaliques . Une surprise , bien que purement artificielle , cette approche présente malgré tout, des résultats acceptables aussi bien dans les trajets formantiques que lors du contrôle auditif . (fig 4) . Les figures 5 , 7 donnent la détection automatique des formants pour les transitions suivantes i--->a, A--->u et le signal synthétique . Les figures 6 et 8 présentent les transitions dans le plan acoustique F1/F2 dont l'allure est conforme à ce que l'on peut observer pour la parole naturelle .

2-INTERPOLATION NON LINEAIRE

On utilise comme évolution temporelle de fonctions d'aires et de longueurs , une variation sinusoïdale de type : (fig 9)

$$A(t) = (Af-Ad)/2 * (1+\sin\Pi(t-td)/(tf-td))$$

Avec Ad = aire de départ de la cible
td = instant de départ de la cible
Af = aire finale de la cible
tf = instant final de la cible
tp = temps de passage d'une cible à l'autre
tt1 = temps de tenue de la 1er cible

tt2 = temps de tenue de la 2ème cible

Les figures 10 à 13 montrent le suivi automatique des formants et les transitions dans le plan acoustique F1/F2 . Des mesures sur les mouvements des membres , des articulateurs (mâchoire) permettent en effet de considérer comme satisfaisante , en première approximation , des évolutions sinusoïdales pour des transitions rapides . C'est d'ailleurs ce que nous avons vérifié pour les mouvements de la main (grâce à une tablette d'acquisition graphique) et pour le déplacement de la mâchoire (à l'aide d'un kinésiographe , cf. LALLOUACHE & WORLEY dans ces mêmes JEP) . Au niveau du suivi des formants et des trajectoires formantiques les différences entre les 2 types d'interpolation sont très faibles (d'ailleurs comment les évaluer ?) . A l'écoute les différences sont notables , les stimuli paraissent dans le deuxième cas plus naturels .

CONCLUSIONS

Ce premier travail nous a permis de mettre en oeuvre un modèle de synthèse dynamique à partir de fonctions d'aires . Nous avons testé les transitions voyelle-voyelle , les résultats présentés ici ont été limités aux transitions maximâles entre [i, a , u]. Deux types d'interpolation ont été testés : l'interpolation non linéaire (sinusoïdale) permettant de générer des stimuli plus naturels qu'une simple interpolation linéaire qui malgré tout permet d'obtenir une qualité très acceptable . La suite de cette étude va consister à piloter temporellement le modèle articulatoire de MAEDA [2] pour obtenir des fonctions d'aires mais cette fois-ci comme simples données intermédiaires .

REMERCIEMENTS

Nous tenons à remercier vivement C. ABRY pour ses précieuses suggestions dans le domaine de l'articulatoire.

REFERENCES

[1] MAEDA S. (1982) . A Digital Simulation Method of the Vocal Tract. Speech Communication, vol. 1, N°. 3, 4, 199-229.

[2] MAEDA S. (1979) . Un modèle articulatoire de la langue avec des composantes linéaires 10èmes JEP, GALF-GCP, 27-28 .

[3] PERRIER P. BOE L.J. MAJID R. & GUERIN . (1985) . Modèlisation articulatoire du conduit vocal exploration et exploitation . 14èmes JEP, GALF-GCP , 55-58

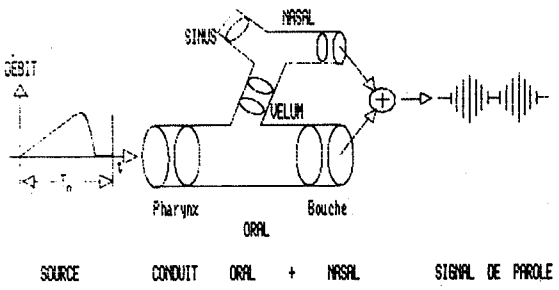
[4] KELLER E. & OSTRY D.J. (1983) . Computerized Pulsed Echo Ultra Sound Measurements of Tongue Dorsum Movements . J. Acoust. Soc. Am.

[5] CHARPENTIER FR. (1984) . Determination of the Vocal Tract Shape from the Formants by Analysis of the Articulatory to Acoustic Nonlinearities . Speech Communication , Vol. 3, N°. 4, 291-308 .

[6] MRAYATI M. (1976) . Contribution aux études sur la production de la parole . Doctorat ès- sciences Physique , USM Grenoble.

[7] BOE L.J. (1973) Etude acoustique du larynx conduit vocal (fréquence laryngienne des productions vocaliques). Revue d'Acoustique 27, 235-244.

[8] FENG G. (1986) Modélisation acoustique et traitement du signal, le cas des voyelles nasales. Doctorat ès-sciences , INP Grenoble



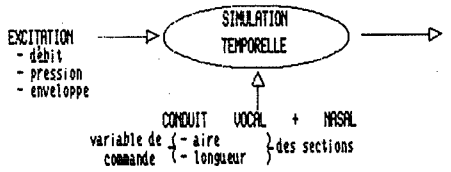


fig 1 : modèle de simulation dynamique du conduit vocal

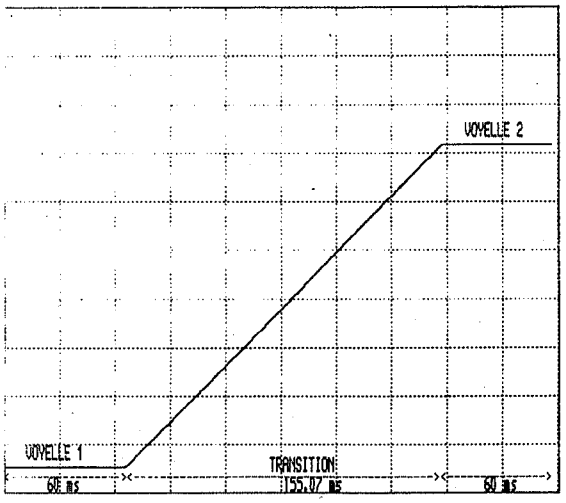


fig 4 : transition linéaire entre les aires

Tube élémentaire de section A_j et de longueur l_j et son équivalent électrique

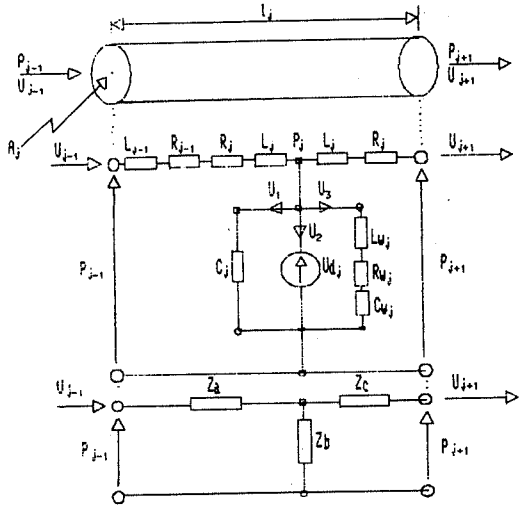


fig 2 : modélisation avec pertes

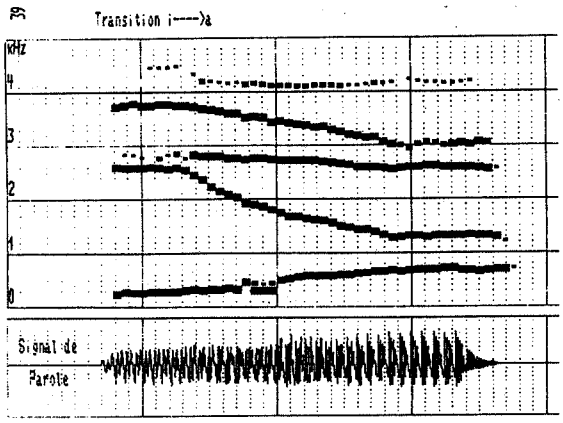


fig 5 : détection automatique des formants

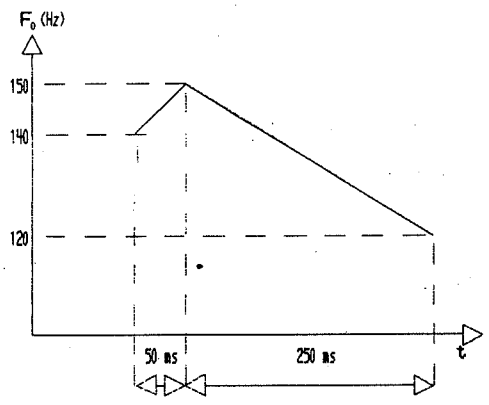


fig 3 : variation de Fo

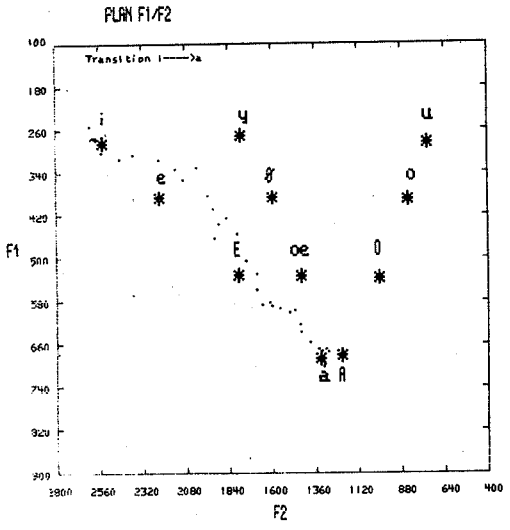


fig 6 : transition i---->a dans le plan F1/F2

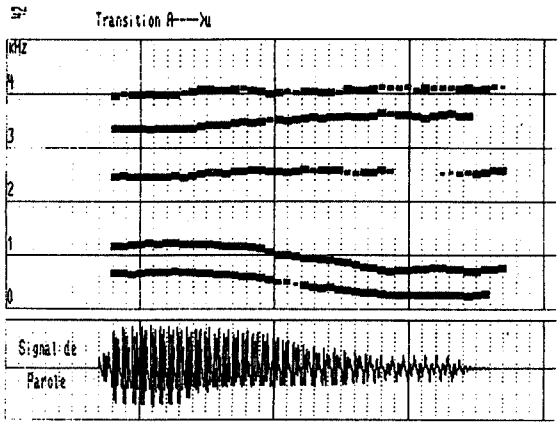


fig 7 : détection automatique des formants

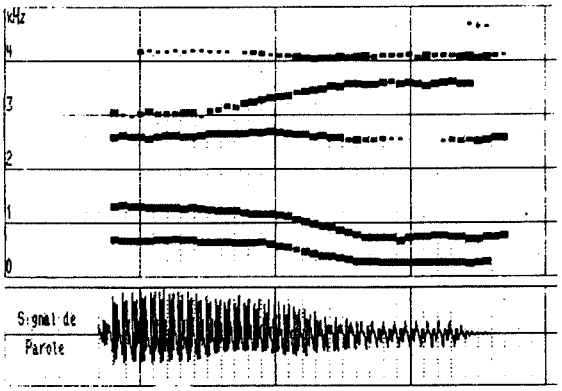


fig 10 : détection automatique des formants

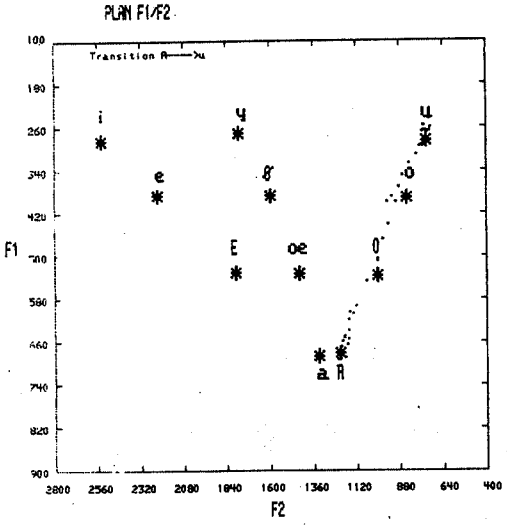


fig 8 : transition A---->u dans le plan F1/F2

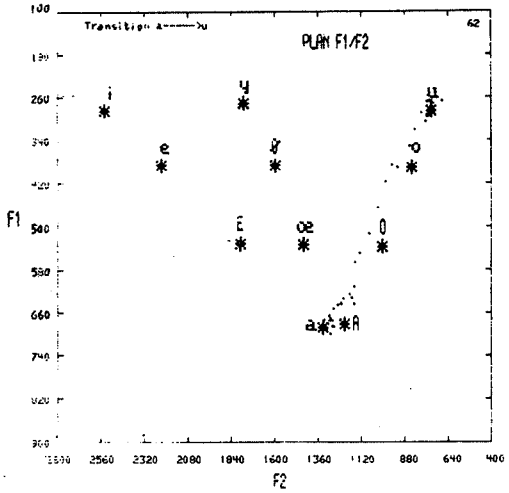


fig 11 : transition a---->u dans le plan F1/F2

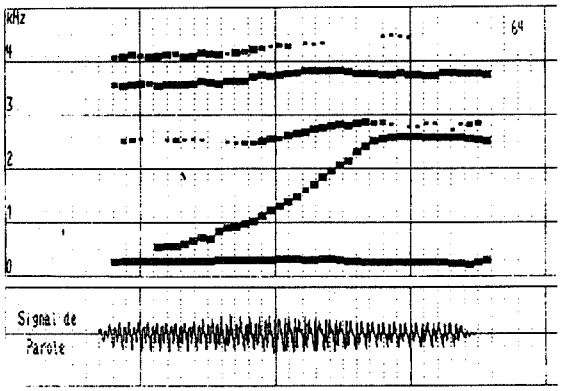


fig 12 : détection automatique des formants

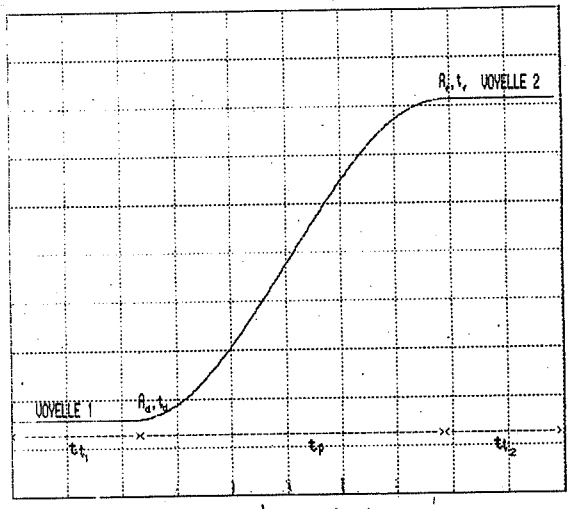


fig 9 : transition non linéaire entre les aires

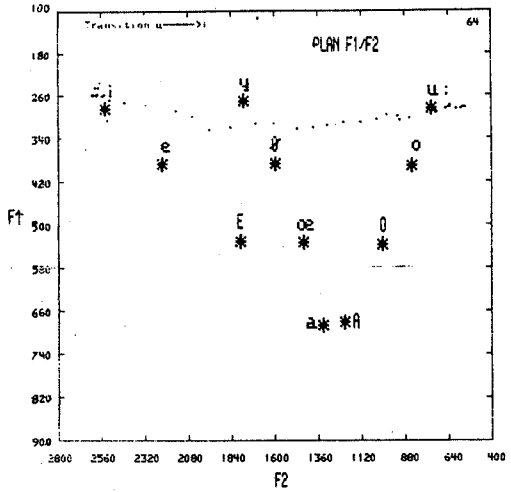


fig 13 : transition u---->i dans le plan F1/F2

SAISIE, EDITION ET TRAITEMENT D'IMAGES ET DE SIGNAUX ARTICULATOIRES LEVRES ET MACHOIRE

Tahar-Med LALLOUACHE & Cecil WORLEY

Institut de la Communication Parlée
Institut de Phonétique de Grenoble
Université de Grenoble III
38040 GRENOBLE CEDEX

ABSTRACT

In the study of speech production, it is crucial to have the speech signal together with some physiological data. In this paper, we describe a system that is used to collect data on lip shapes and jaw movements.

INTRODUCTION

Pour l'étude de la production de la parole, on connait toute l'importance de la saisie simultanée des paramètres physiologiques et du signal acoustique correspondant.

Notre travail a consisté à mettre en place une chaîne d'acquisition et un ensemble d'édition et de traitement de signaux physiologiques [1]. Les paramètres physiologiques que nous avons traités dans un premier temps sont d'origine supraglottique : fournis par un kinésiographe (MYOTRONICS, K5AR) pour le mouvement de la mâchoire et par un ensemble "Vidéo-Parole" [2] pour l'étude des lèvres.

L'ensemble des fonctions temporelles issues du kinésiographe comme du traitement d'images est pris en charge par un logiciel specialisé qui a été adapté pour éditer simultanément ces signaux (8 voies) et opérer d'autre part les traitements classiques (intensité, FFT, Fo, cepstre, filtrage etc.)

Le KINESIOGRAPHE MANDIBULAIRE (MODELE K5AR)

Cet appareil a été conçu pour détecter et suivre en trois dimensions la position de la mâchoire par rapport au crâne. Il se compose de six capteurs, montés sur un cadre très léger, qui sont sensibles aux changements d'un champ magnétique créé par un aimant fixé sur la mandibule, entre la lèvre inférieure et les dents.

Il permet de disposer au cours du temps du déplacement de la mâchoire dans les trois plans sagittal, frontal et horizontal (ou deux plans plus une vitesse de déplacement).

D'autres systèmes sont actuellement plus ou moins disponibles sur le marché notamment Movetrack [3] qui permet d'obtenir le suivi de plusieurs points sur la mâchoire, les lèvres ou la langue. Notre option étant d'obtenir pour les lèvres non pas le suivi d'une ou de deux fonctions temporelles mais de toutes celles que nous permet l'analyse des images. Nous avons confié au kinésiographe le suivi de l'organe articulateur dont la déformation est la moins complexe, gardant pour la vidéo la saisie complète des lèvres.

Le stockage FM

Avant numérisation, il est possible de stocker sur enregistreur FM le signal de parole, les signaux délivrés par le kinésiographe et tout autre signal physiologique.

La numérisation

Tous ces signaux peuvent être directement numérisés grâce à un système à multiplexage (8 signaux plus le signal de parole) [4]. Les huit voies ont été conçues pour les signaux de basse fréquence; l'échantillonnage peut se faire à 100 Hz ou à 200 Hz, le signal de parole est échantillonné à 10 kHz. Le signal multiplexé est acquis par le système informatique grâce à un calculateur spécialisé et stocké dans un fichier qui peut être traité par logiciel.

Les premiers résultats obtenus concernent la mise en correspondance des événements acoustiques [5] et des événements articulatoires; la figure 1. nous montre : en haut, un des signaux du MKG, le déplacement vertical; en bas, le signal de parole (sur lequel nous avons superposé le signal de mâchoire), le tout pour l'énoncé " C'est ça? " répété. Nous pouvons en extraire le passage entre cibles vocaliques de, /ɛ/ à /a/, (cycle vocalique), auquel se surimpose le cycle consonnantique de /s/, consonne pour laquelle le couplage langue-mâchoire est maximal [6].

LE POSTE VIDEO-PAROLE

But

Ce poste de travail est destiné à l'étude des images-parole. Son but étant en premier lieu l'analyse automatique des images vidéo (pour l'instant du visage) afin d'éffectuer un suivi de points de référence. La sélection des images étant opérée à partir du signal synchrone de parole. Cette méthode pourra être étendue à des images endoscopiques ou radiographiques [7].

Nous avons tenu compte dans notre travail des précédents systèmes d'exploitation automatique de films pour les données labiales [8],[9].

La prise de vues

La prise de vues est normalement faite dans des conditions optimales d'éclairage et de maquillage (lèvres en blanc, dents en noir). Les sujets sont filmés de face, il est également possible d'avoir l'image de profil avec un miroir placé à 45°, d'un côté du locuteur. Notons que quand les sujets portent le cadre du MKG il est difficile de les filmer de profil même en utilisant deux caméras.

Nous avons utilisé une caméra CCD, noir et blanc, haute résolution, munie d'un téléobjectif, à la cadence de 25 images/seconde. (Nous prévoyons la possilité d'utiliser une caméra 200 images/seconde.) La prise de vues est complètement séparée de l'exploitation. Les signaux (image + parole) sont systématiquement stockés sur un magnétoscope. Mais ils pourront aussi provenir d'un disque vidéo ou d'une numérisation opérée à partir d'un film classique 16 ou 35 mm.

Unité de traitement

Il s'agit d'un PC AT doté pour le moment d'un disque de 50 Mo (30 seulement utilisables sous MS.DOS version 3.1)

L'interface signal vidéo-PC est constituée par une carte Matrox entièrement programmable : 4 plans mémoire 512 X 512 X 8 bits (256 niveaux de gris), numérisation en temps réel, gain et offset programmables, accès direct mémoire, visualisation sur moniteur RVB.

Un seuilleur réglable manuellement nous permet de faire l'acquisiton d'images préalablement binarisées.

Résultats

La comparaison avec les résultats manuels [2] a montré que la procédure de détection automatique est opératoire et précise.

Nous donnons pour deux types de traitement les exemples de problèmes rencontrés.

CAS 1 (fig. 2)

Le seuil de binarisation est déterminé à partir de l'histogramme de la fenêtre d'analyse, celle-ci entourant largement l'ouverture labiale. Le résultat du seuillage est placé aux coordonnées (0,0) de l'écran vidéo. Deux projections (suivant OX et OY) permettent d'isoler la forme labiale des ombres et des points-repères de maquillage déterminant ainsi un nouveau cadre de travail.

Une première analyse en inertie nous donne les axes principaux de la forme puis une optimisation polynomiale de degré quatre [10] permet de tracer l'arc supérieur et d'éliminer ainsi les ombres portées sur la lèvre supérieure.

Une seconde analyse permet de déduire la valeur de l'étirement labial. L'aire intéro-labiale et l'aperture (repérée à partir du barycentre de la forme) sont mesurées directement grâce un comptage de pixels.

Cas 2 (fig. 3)

Pour les petites ouvertures labiales, il n'y a pas d'ombre portée sur les commissures des lèvres, alors la procédure de traitement est différente : après détermination d'un seuil de binarisation et projection de l'image seuillée suivant OX et OY, les trois paramètres sont déterminés par un simple comptage de pixels.

L'automatisation des mesures a montré que les résultats pouvaient être meilleurs que ceux obtenus à partir des tracés manuels, en particulier pour les petites fentes labiales. En effet les deux procédures reposent sur des principes de traitement différents : par exemple le calcul de l'aire à partir d'un pèrimètre polygonal induit quelque soit l'algorithme des erreurs importantes sur les petites aires. Le comptage de pixels ne pose évidemment pas ce genre de problèmes.

Il nous reste sans aucun doute à améliorer le temps de traitement et la précision grâce à une meilleure prise de vues (contraste et élimination des ombres). Nous pourrons ainsi opérer le même traitement quelque soit la dimension de la forme intéro-labiale.

REMERCIEMENTS

Nous remercions tout particulièrement J.P. CHARRAS (Laboratoire de Traitement d'Images et de Reconnaissance des Formes, ENSERG) qui nous a fait bénéficier de son savoir expert en traitement d'images; T. GAY (Department of Oral Biology of Farmington, Connecticut) pour nous avoir transmis son expérience en kinésiographie; enfin L.J. BOE et C. ABRY (Institut de Phonétique de Grenoble) qui nous ont guidé dans ce travail.

BIBLIOGRAPHIE

[1] WORLEY C. (1986), Acquisition, visualisation et traitement de signaux physiologiques et/ou acoustiques de parole. Rapport de D.E.A Systèmes Electroniques, ENSERG

[2] LALLOUACHE M. T. (1987) Détection Automatique du contour de lèvres et extraction des paramètres constitutifs. Rapport DEA Systèmes Electroniques, ENSERG.

[3] BRANDERUD P. (1985) Movetrack, a Movement Tracking System. Symposium franco-suédois sur la parole (B. GUERIN et R. CARRE eds.) Grenoble 22-24 Avril 1985, II 113-122.

[4] AL HAKAWATI AL DAKKAK O. (1985) Acquisition de signaux pour kinésiographe. Rapport de DEA Traitement de l'Information, ENSERG.

[5] ABRY C. BOE L.J. CORSI P. DESCOUT R. GENTIL M. & GRAILLOT P. (1980) Labialité et phonétique : Données fondamentales et études expérimentales sur la géométrie et la motricité labiale. Université des Langues et Lettres de Grenoble.

[6] BOGNAR E. (1982) Espaces et cibles mandibulaires. Etude de l'espace articulatoire et des positions cibles de la mâchoire dans les réalisations de séquences consonne-voyelle chez deux locuteurs français. Thèse de 3ème cycle de phonétique, Université de grenoble III.

[7] TESTON B. & AUTESSERE D., (1986) Description d'un dispositif simultané des mouvements des organes articulateurs. 15ème JEP du GCP du GALF 65-68.

[8] GOURRET J.P PAILLE J. BOE L.J. DESCOUT R. (1985) Système d'exploitation automatique de labiofilms. 14ème JEP du GCP du GALF, 61-64.

[9] LOUVION J.R. Détection automatique du contour de lèvres et extraction de paramètres constitutifs. Application à l'analyse automatique de labiofilms. Thèse de 3ème cycle, Université de Rennes I.

[10] LINDBLOM B. E. F. & SUNDBERG J. E. F. (1971) Acoustical Consequences of Lip, Tongue, Jaw and Larynx Movement. J.A.S.A 50, 1166-1179.

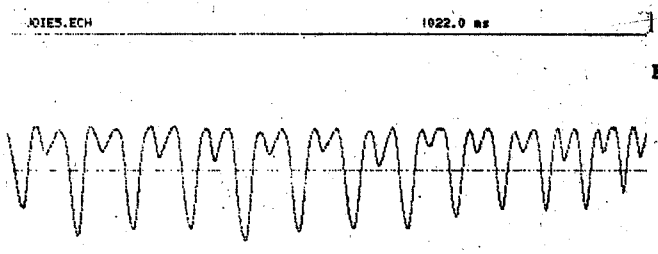


Fig. 1 : Répétitions de " C'est ça ? ".
En haut, le déplacement vertical de la mâchoire issu du kinésiographe (échantillonné à 100 Hz); en bas le signal de parole (à 10 kHz)avec le signal mâchoire surimposé.

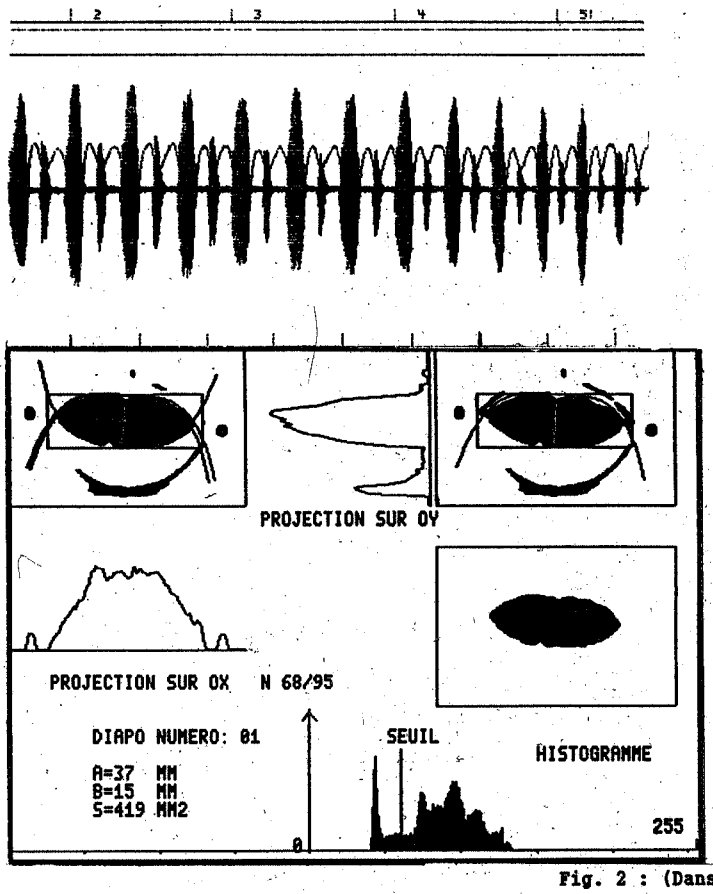


Fig. 2 : (Dans le cadre) les différentes étapes du traitement automatique : en haut, à droite le résultat de la première analyse ainsi que le tracé de l'arc supérieur; a gauche, le résultat de la seconde analyse; en bas, à droite le contour détecté est replacé sur le visage à des fins de contrôle.
Les résultats de la détection sont affichés : A : étirement; B : aperture S : aire intéro-labiale.
A droite, les résultats des tracés manuels

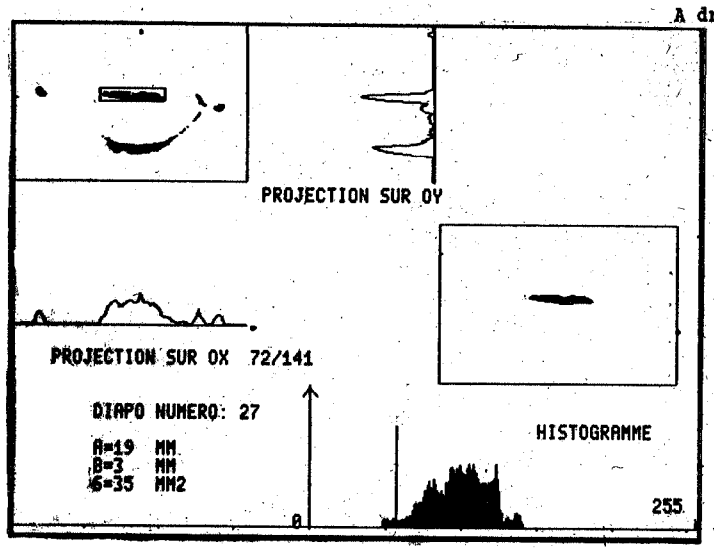


Fig. 3 : Mêmes indications que pour la fig. 2 mais, sans analyse en inertie.

L'asymétrie des appuis linguo-palatins

Alain Marchal & Robert Espesser

Institut de phonétique, UA 261 CNRS
29, avenue R. Schuman, 13621 AIX-EN-PROVENCE

Introduction

La radiocinématographie en levant l'obstacle de l'inaccessibilité de la langue à une observation directe a permis d'obtenir des informations dynamiques capitales sur l'articulation. On doit à cette technique l'essentiel des informations sur les voyelles et sur les consonnes du Français (1, 2, 3, 4)).
Il s'agit de vues latérales de la langue. Par extrapolation, on a utilisé ce type de données articulatoires bidimensionnelles pour alimenter des modèles articulatoires de synthèse de la parole. Les résultats, bien que dignes d'intérêt, laissent toutefois à désirer si on prend le qualificatif "articulatoire" au pied de la lettre. Outre les problèmes de source bien connus, il semble bien aussi que les fonctions d'aires dérivées de films aux rayons x soient loin de prendre en compte les réalités anatomiques et physiologiques. Nous en voulons pour témoin par exemple les abérrations que fournissent les transformées inverses (Fig.1).

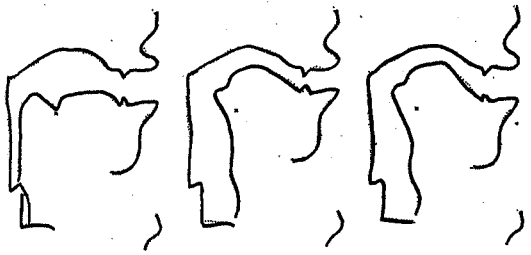


Fig.1 Transformée inverse de la transition de /k/ à /a/ dans /katga/.

Celles-ci étaient stigmatisées dans ces termes par Fant (5) au congrés de phonétique de Copenhague "The so-called inverse transforms generate a "pseudo area-function" that can be translated back to high quality synthetic speech but which remains fictional in the sense that they do not necessarily ressemble natural area function. Their validity is restricted to non-nasal, non constricted articulation and even so, they at best retain some major aspects of the area function shape rather than its exact dimensions".
Il faut reconnaître que les méthodes d'investigation couramment utilisées jusqu'à présent pouvaient facilement par leurs limites propres induire un tel biais. Ainsi la radiographie comme la radiocinématographie et ses techniques dérivées ne fournissent qu'une vue latérale du bord le plus élevé de la langue; Elle renseignent donc sur le degré de constriction
maximale, mais ne permettent pas de préciser si toute la masse linguale ou non est concernée par le mouvement d'élévation de la langue.
La palatographie directe ou indirecte avait contribué à montrer qu'il pouvait exister un patron asymétrique d'appui lingual; mais cette technique n'offrait qu'une représentation composite de toute une série de mouvements depuis le début de l'élévation de la langue jusqu'à la fin de l'articulation. Cette technique, malgré les tentatives intéressantes de Firth (6) avec la "word palatography" ne permettait d'étudier que des articulations plus ou moins isolées, et partant, assez artificielles.
L'électropalatographie (7) a permis de répondre à la demande d'informations dynamiques sur l'évolution des appuis linguo-palatins dans leur largeur. Nous avons utilisé cette technique pour examiner la coarticulation dans les groupes d'occclusives en Français (8). Nous avons pu observer qu'il existait souvent un appui préférentiel de la langue au palais (Fig. 2).

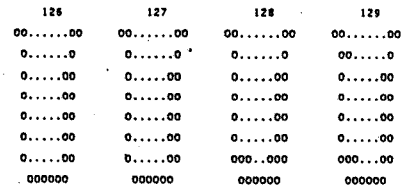


Fig. 2 Les palatogrammes dynamiques montrent une réalisation de /d/ caractérisée par un appui préférentiel à gauche.

Nous proposons pour essayer de décrire et de quantifier ce phénomène d'utiliser un indice d'asymétrie. Nous discuterons également dans cette communication des implications de cette observation.

L'indice d'asymétrie

$$I_a = \frac{Na - Nb}{Na + Nb}$$

où N = Nombre de contacts,
a = côté droit, b = côté gauche

La figure 2 indique la répartition sur 8 rangées (1 à 8) des 64 électrodes sur le palais artificiel. Nous avons divisé le palais en 4 zones : antérieure et postèrieure, droite et gauche. Nous pouvons ainsi calculer un indice d'asymétrie global ou un indice d'asymétrie pour les parties antérieure et postérieure.

Etude

Les questions que nous allons examiner seront les suivantes :
- Peux-t'on définir une asymétrie globale ou doit-on distinguer l'asymétrie dans les parties antérieure et postérieure ?
- l'asymétrie concerne-t'elle tous les phonémes de la même manière ou existe-t'il des phonémes plus "asymétriques" que d'autres ?
- la coarticulation contraint-elle le degré d'asymétrie ?

Les résultats que nous présentons portent sur 3000 palatogrammes de tenues des occlusives /t, d, k, g/ obtenus à l'aide du système d'électropalatographie de Montréal. Ils concernent l'articulation d'un sujet et ont de ce fait un caractère tout à fait préliminaire. Ils invitent à s'interroger sur la validité de l'indice proposé et sur l'intérêt potentiel de l'application de la méthode à une population importante.

Résultats

L'asymétrie globale

Le calcul de l'indice sur l'ensemble du corpus fait apparaître un appui préférentiel significatif ($F(1, 5020) = 7.38$, $\ll .007$) du côté gauche (Fig. 3) .
La population regroupée par classe d'indice exprimée en % montre qu'il existe une légère asymétrie à gauche

Notre corpus est constitué de consonnes occlusives antérieures et postèrieures, il faut donc pour essayer de réduire la dispersion observée affiner l'analyse et examiner l'influence du lieu d'articulation.

L'effet du lieu d'articulation

Pour les alvéodentales, on observe une légère asymétrie (taux moyen = -4.7%, asymétrie significative: = « .01) qui avantage le côté gauche (Fig.4).
Pour /k, g/, on ne peut valablement examiner que les appuis dans la zone postèrieure : en effet, les rares contacts dans la zone antérieure, dûs selon toute vraisemblance à des effets de coarticulation vocalique auraient un poids trop important dans le calcul d'un indice d'asymétrie. Il est raisonnable de comparer les appuis de 15 à 20 contacts se répartissant du côté droit et du côté gauche mais cette analyse n'a plus guère de sens lorsqu'on oppose par exemple 2 contacts à droite à un contact à gauche; à moins d'alourdir la procédure et de pondérer l'indice, ce qui reste possible bien sûr mais ne serait nécessaire seulement que pour quelques cas isolés.
On observe pour /k, g/ un appui préférentiel à gauche dans la zone palatale et postpalatale (Fig. 5, taux moyen = -7.9%, asymétrie significative: «.01).

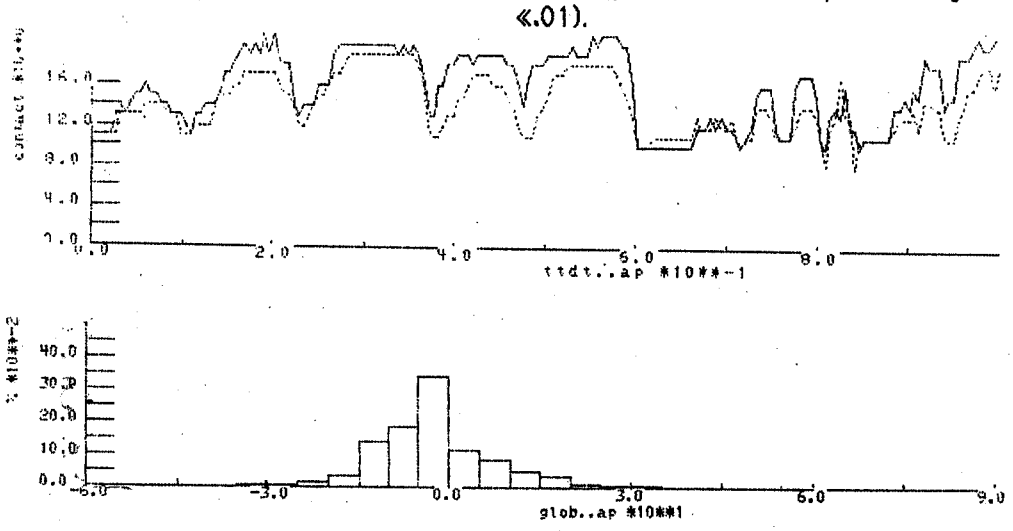


Fig. 3 Distribution des contacts à droite (pointillé) et à gauche (continu) pour toutes les consonnes.

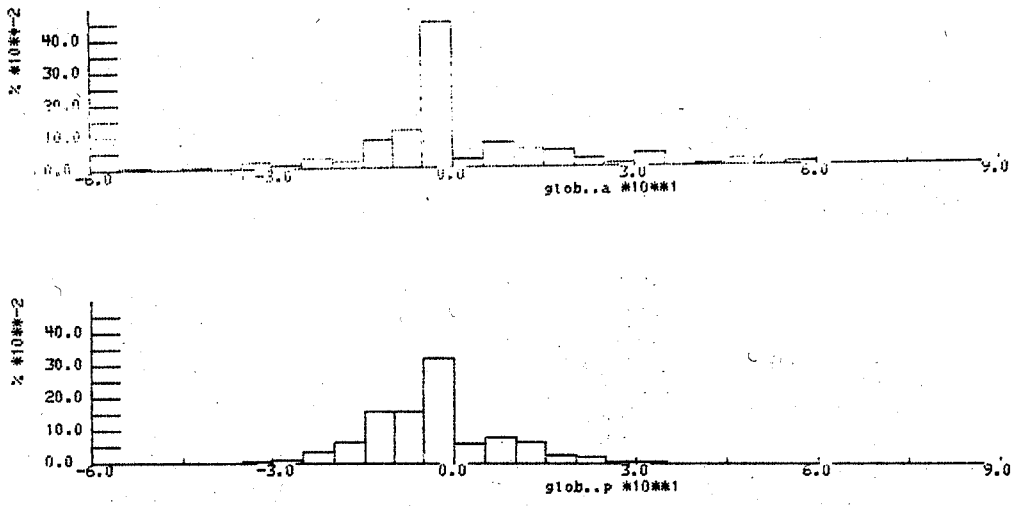


Fig. 4 Effet du lieu d'articulation sur l'asymétrie de l'appui linguo-palatal.
(ap: palais total, a: palais antérieur, p: palais postèrieur)

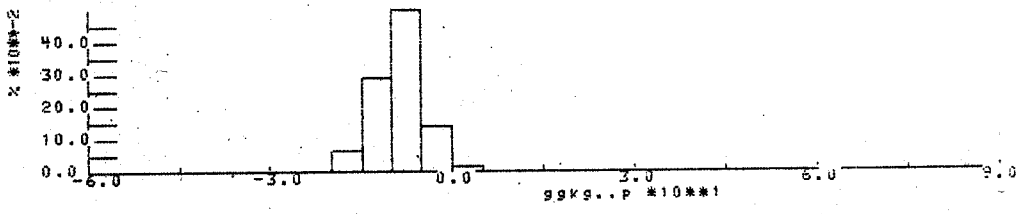


Fig. 5 Asymétrie à gauche pour /k. g/

L'effet de coarticulation

Il nous reste maintenant à examiner le problème de la coarticulation et nous allons l'aborder par l'étude de l'appui lingual dans les groupes d'occlusives. On pourrait croire a priori que les patrons observés pour les consonnes simples se retrouveraient plus ou moins identiques. Il n'en est rien comme le montre graphiquement la figure 6.

On y voit en effet qu'il existe une dispersion trés importante. On peùt se poser la question de savoir si la position de la consonne dans le groupe exerce une influence.
Nos observations sur 800 palais sont résumées dans le tableau suivant.

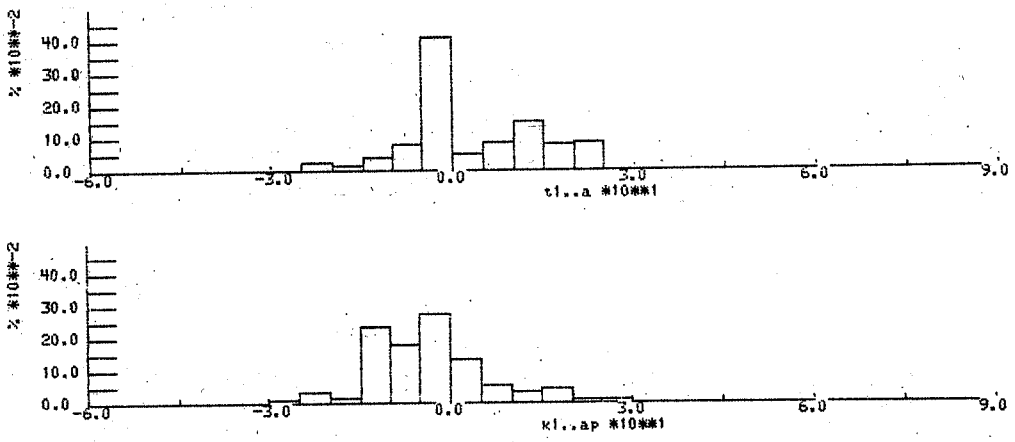


Fig. 6 L'asymétrie de /t: et /k/ en position initiale dans un groupe consonantique apparaît variable.

t1, d1= Asymétriques à droite et à gauche.
t2, d2= peu asymétriques
k1, g1= Asymétriques à gauche.
k2, g2= Asymétriques à droite et à gauche.

Il est manifeste que les mouvements de la langue sont trés différents selon que l'on a affaire à une consonne simple ou à la même consonne dans un groupe consonantique. L'écart entre les taux de ces deux classes est significatif («.01). Cela semble bien indiquer que les mouvements du dos et de l'apex ne sont pas aussi indépendants qu'on a tendance à le croire généralement. Il est par ailleurs intéressant de noter que le mouvement de la langue semble moins précis pour l'articulation d'un groupe consonantique. L'étude des faits de coarticulation linguale nous avait amené à postuler l'existence d'un phénomène de coproduction. Celui-ci se manifestait en particulier par un mouvement d'élèvation de la langue relativement indifférencié par rapport à la cible de C1 ou C2. Il est intéressant de voir qu'une étude apparemment aussi éloignée de la première, comme celle de l'asymétrie des appuis de la langue au palais aboutisse à la confirmation des premières conclusions.

Conclusion

Les observations sur l'asymétrie des appuis de la langue au palais appellent plusieurs commentaires. Elles montrent les limites des méthodes d'investigation de l'activité linguale qui ne permettent d'examiner que deux dimensions à la fois. L'idéal serait de pouvoir obtenir immmédiatement des informations tridimensionnelles comme celles fournies par les images scanner et RMN (9). En attendant de pouvoir disposer de ces nouveaux moyens techniques ou d'autres procédés produisant des données similaires (10) la radiocinématographie et la palatographie dynamique doivent être utilisées de manière complémentaire.
Les patrons de contacts linguaux montrent que le lieu d'articulation exerce une influence sur le degré d'asymétrie. A un appui préférentiel à droite dans la zone antérieure peut correspondre un appui préférentiel à gauche dans la partie postérieure : ceci semblerait indiquer, soit que la langue ne s'élève pas uniformément ou régulièrement dans l'axe alvéo-palatal, soit que la morphologie palatine du sujet est elle-même asymétrique favorisant ainsi pour un même degré d'élèvation de la langue un appui préférentiel d'un côté. Pour examiner ces deux hypothéses, il faudra reprendre cette étude sur un plus grand nombre de locuteurs et en multipliant les mesures anatomiques.
L'influence de la position de la consonne dans un groupe d'occlusives apparaît importante. Elle nous a amené à rappeler notre hypothése sur la coproduction.

Références bibliographiques

(1) Simon, P. (1967).: *Les consonnes françaises, mouvements et positions articulatoires à la lumière de la radiocinématographie.* Klincksieck, Paris.

(2) Brichler, C. (1970).: *les voyelles françaises. Mouvements articulatoires à la lumière de la radiocinématographie.* Klincksieck, Paris.

(3) Santerre, L. (1971).: *Les voyelles orales dans le français parlé à Montréal.* Thése de doctorat d'état, Strasbourg.

(4) Rochette, CL. E (1973).: *Les groupes de consonnes en Français.* Klincksieck, Paris.

(5) Fant, G. (1979).: "The relation between area function and the acoustical signal" *Proc. 9th. Int. Congr. Phon. Sci.*, Vol II, pp. 79-108.

(6) Firth, J.R. (1948).: "Word-palatograms and articulation;" Bull. Sch; of oriental and African studies, *12*: 857-864.

(7) Hardcastle, W.J; (1972).: "The use of electropalatography in phonetic research." Phonetica, *25*:197-215.

(8) Marchal, A. (1985).: *L'électropalatographie, contribution à l'étude de la coarticulation*. Thése de doctorat d'état, Nancy II.

(9) Rossi, M. & Autessere, D. (1981).: "Movements of the hyoid bone and the larynx and the intrinsic frequency of vowels." J. Phon., *9*:233-249.

(10) Morrish, K.A.; Stone, M.; Shawaker, TH.H. & B.C. Sonies. (1985).: "distinguisability of tongue shape during vowel production. J. Phon., *13* : 189-203.

VERS UNE MODELISATION DES MOUVEMENTS DU DOS DE LA LANGUE.

Pascal PERRIER[1] Christian ABRY[2] & Eric KELLER[3]

1 Laboratoire de la Communication Parlée, ENSERG/INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex.
2 Institut de Phonétique de Grenoble, Université III, Domaine Universitaire, 38400 St Martin d'Hères.
3 UQAM & Centre de Recherche du Centre Hospitalier Côtes des Neiges, 4565 Queen Mary, Montreal.

ABSTRACT.

To model tongue highest point trajectories in [ka] syllables (stressed and unstressed, fast and slow), a second order system has been used with the following characteristics :

* stiffness is the control parameter ;
* it is time swept ;
* the model is an agonist-antagonist distributed one ;
* targets are of the equilibrium-point type (specified by the agonist/antagonist stiffness ratio) ;
* they are timed targets.

Merit and pitfalls of this modelling are evaluated.

I-INTRODUCTION.

Dans le cadre d'une réflexion théorique sur les aspects dynamiques de l'articulation , nous proposons une modélisation du mouvement du dos de la langue (du "point" le plus haut), sous la forme d'un système du second ordre, où la raideur serait le paramètre contrôlé au cours du temps. L'introduction des paradigmes venus de la psychomotricité dans les études sur la parole, tout en ravivant parfois de vieux et vifs débats (J. Phonetics 14, 1986), nous amène à situer nos choix parmi les modèles possibles. Ces choix cruciaux portent notamment sur :

* le statut du temps dans les programmes moteurs (Schmidt vs. Fowler & Kelso) ;
* la nature des cibles (Polit & Bizzi vs. MacNeilage) ;
* les paramètres de contrôle (force, longueur ou raideur ?) (Fujimura vs. Abbs vs. Cooke ou Feldman).

Pour nous, le temps sera une composante contrôlée de la parole, non seulement en ce qui concerne les variations "volontaires" de la vitesse d'élocution, mais aussi dans l'orchestration des initiations (onset) du mouvement vers les cibles. En ce domaine, la référence sera donc davantage Schmidt [1] que Kelso & Tuller [2].

La spécification des cibles en termes spatiaux [3] a trouvé une concurrence importante dans la notion de "cible-équilibre" [4] (voir à ce sujet la revue de questions in Berkinblit et al. [5]). La capacité d'atteindre, pour un singe déafférenté, son objectif, quelle que soit la perturbation apportée au cours du mouvement (que ce soit en opposition ou en facilitation), a été, à cet égard, un résultat majeur, du moins pour l'efficacité des cibles atteintes en mono-articulateur. Cela ne signifie pas, bien entendu, que, pour la parole, nous ayons affaire, dans tous les cas, au même type de cibles : pour les plosives par exemple, opposées en cela aux voyelles, on doit évidemment tenir compte de la notion d'impact inter-articulateur.

Enfin, le contrôle de paramètres classiques comme la force, la longueur, la raideur, la viscosité ou la vitesse, fait l'objet d'un débat toujours ouvert. Et même en ce qui concerne la raideur ici choisie - qu'une revue de questions du début des années 80 [6] donnait gagnante (23% des auteurs, contre 3 à 6% pour la force, la longueur et la vitesse, p. 569) - la forme du contrôle oppose deux conceptions : celle de Cooke [7], qui change directement la raideur pour passer d'une cible à l'autre, et celle de Feldman [8] qui peut agir simultanément sur le seuil d'activation (déterminé par le réflexe myotatique) et la raideur.

II-LES MODELISATIONS ANTERIEURES.

Parmi les modélisations déjà proposées, nous ne mentionnerons, pour la parole, que les recherches pilotes qui se sont définies explicitement par rapport à la spécification de la raideur.

Nelson [9] a montré que les mouvements de la mâchoire suivaient un principe de minimisation des coûts en énergie et en "jerk" (variation de l'accélération) ; ce qui est aussi le cas d'un système du second ordre, où, pour un amortissement nul, la vitesse maximale Vmax atteinte en cours de mouvement, et l'amplitude du mouvement, Amp, sont liées par la relation : $Vmax/Amp = c/T$, c étant une constante et T la durée du mouvement.

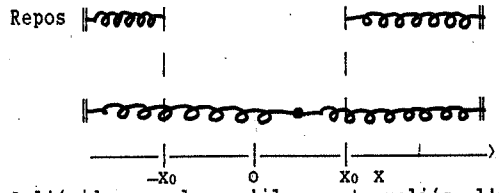
Pour le mouvement de la langue, les études de Ostry, Keller, Parush et Munhall [10] [11] ont mis en évidence la validité de cette relation pour différentes variations d'accent, de débit et pour différentes voyelles. Mais Nelson [9] comme Munhall & Ostry [12] soulignent bien évidemment les limites d'une modélisation par un simple ressort sans amortissement.

C'est à cette même conclusion que sont arrivés Browman & Goldstein [13] qui utilisent, pour les ajuster au mouvement de la lèvre inférieure, des portions de sinusoïdes, dont la fréquence et l'amplitude sont modulées respectivement par la raideur et la position au repos d'un ressort non amorti. Ils constatent effectivement que les plus grandes erreurs se situent là où le mouvement donne un plateau, là où la vitesse s'annule (consonne ou voyelle accentuée).

De notre point de vue, s'il est certain que le système doit être bien plus élaboré, ce n'est pas dans la direction de la prise en compte d'un amortissement ad hoc, mais plutôt dans l'introduction directe de la variable "temps", à la fois pour déterminer le passage d'une cible à l'autre et pour décrire l'évolution temporelle de la raideur.

III-NOTRE MODELISATION.

III.1. Au plan mécanique : Notre point de départ était le modèle masse-ressort amorti de Cooke [7], mais il s'est rapidement avéré que les différences dans les origines mécaniques des forces intervenant dans le mouvement (raideur d'un côté, poids de l'autre) induisaient de regrettables dissymétries dans les profils de déplacement et de vitesse. Les résultats sur le mouvements du bras sont, à cet égard, tout à fait révélateurs [7]. Par conséquent, nous avons opté pour une modélisation à deux ressorts amortis, qui autorise une représentation symétrique des rôles des muscles agonistes et des muscles antagonistes. Pour nous affranchir des problèmes liés à la masse située à la jonction des deux ressorts, nous ne considérons que les mouvements dans la direction horizontale, et la masse est normalisée à 1.



A l'évidence, lorsqu'ils sont reliés l'un à l'autre, les ressorts sont étirés par rapport à leur longueur au repos, et les forces en présence seront exclusivement des forces de rappel. Considérons alors un axe horizontal Ox dont l'origine est située au milieu du segment unissant les extrémités des ressorts, lorsqu'ils sont au repos. L'équation du mouvement sur cet axe est alors très simplement :

$$(1) \qquad \ddot{x} = -f.\dot{x} - (k_1+k_2).x - (k_1+k_2).x_0$$

où f (viscosité), k_1 (raideurs) et x (déplacement) sont des fonctions du temps.

Ce type d'équation est a une solution numérique par la méthode de Runge & Kutta, quelles que soient f(t) et $k_1(t)$, pourvu qu'elles soient des fonctions continues.

Pour l'instant, nous avons choisi de garder f constant, suffisamment faible pour ne pas empêcher le mouvement et suffisamment fort pour éviter les oscillations autour des positions d'équilibre dans la gamme de raideur dans laquelle nous travaillons. Mais nous pourrions de même envisager l'amortissement critique, variant donc avec la raideur. Pour les raideurs k_1, nous avons opté pour une variation sinusoïdale entre les valeurs cibles. Ce choix a l'avantage de présenter des dérivées nulles à l'origine et à la cible, ce qui préserve la continuité de la fonction par delà les différents segments de la trajectoire. Elle s'est d'autre part avérée efficace pour la description de différentes trajectoires articulatoires.

Notons que jusqu'à présent seules quelques modélisations ont pris en compte cette variation temporelle de la raideur, et ceci avec des solutions plus ou moins sommaires du type "ramp_and_hold" (cf. par exemple Adamovich & Feldman [14]).

III.2. Les cibles : Nous avons choisi de définir, au niveau des commandes du modèle, une cible comme un état d'équilibre entre les forces agonistes et antagonistes. La position théorique x_c (obtenue sans tenir compte de l'inertie du modèle) correspondant à cet équilibre est donnée par une relation du type :

$$(2) \qquad k_1/k_2 = (x_0-x_c)/(x_0+x_c)$$

La cible-équilibre est donc parfaitement déterminée par la valeur $\alpha = k_1/k_2$. On constate immédiatement qu'une infinité de couples (k_1,k_2) satisfont ce critère. Le passage d'une cible à une autre peut ainsi s'effectuer dans trois conditions différentes :

a. k_1 constant, k_2 varie ;
b. k_1 varie, k_2 constant ;
c. k_1 et k_2 varient ;

Et à l'intérieur du cas (c) on peut encore distinguer deux sous cas, selon que la somme des raideurs (k_1+k_2) varie ou ne varie pas. Cette dernière remarque prend tout son intérêt si on observe l'équation (1) : c'est l'équation du mouvement d'un ressort de raideur (k_1+k_2), sous l'effet d'une force $(k_2-k_1).x_0$. Cette raideur somme correspond donc à la raideur globale du système, la "raideur de co-contraction". Nous verrons par la suite en quoi le fait d'effectuer un mouvement à niveau de co-contraction constant ou variable peut avoir, dans la parole, de l'importance.

III.3. Les commandes : Ceci étant posé, nous décrivons le mouvement comme le résultat d'une succession de commandes au niveau des cibles-équilibres, éxécutées avec des contraintes mécaniques (raideurs) et temporelles données. Considérant en effet sur ce dernier point avec Schmidt [1] qu'un mouvement peut, dans les mêmes conditions biomécaniques être réalisé plus ou moins rapidement, nous introduisons directement le temps au niveau des commandes du système.

Nous avons aussi choisi de privilégier le rôle des muscles agonistes : ce sont eux en effet qui, dans notre modélisation, imposeront la raideur de co-contraction finale (à la cible), compte tenu du coefficient α imposé par la commande. Il s'ensuit que la raideur de co-contraction pourra, si le maintien à la position cible l'exige, être différent de celui programmé par la commande.

Dans ces conditions, la caractérisation d'une cible est faite en donnant :

* le coefficient $\alpha = k_1/k_2$;
* la raideur du ressort agoniste ;
* la raideur de co-contraction .

De plus, au niveau temporel, on imposera :

* le temps t_c de transition entre cibles ;
* le temps t_t de tenue de la cible.

III.4. Quelques résultats généraux : Le premier test du modèle a consisté en l'évaluation de sa capacité à rendre compte des mouvements cycliques du bras décrits par Cooke [7]. A cet égard, les résultats sont encourageants : l'approximation de la trajectoire (figure 1) est bonne, et le profil de vitesse est plus conforme à la réalité que ceux que donnait le modèle masse-ressort non distribué (lumped).

Compte tenu des résultats de Nelson [9], Ostry & Munhall [10] et Ostry & Cooke [15], un second point important est la vérification de la loi Vmax/Amp = c/T. Là encore, même s'ils sont obtenus sur un petit nombre de points, les résultats sont positifs (figure 2).

D'autre part, il est clair que la position spatiale atteinte à la cible est le résultat de l'effet conjoint des commandes cible-équilibre et des commandes biomécaniques et temporelles. Pour une même valeur de α, la position ne sera pas la même selon que le mouvement est fait lentement avec une grande raideur ou vite avec une raideur faible (cas extrêmes). C'est ainsi que l'on pourra ne pas atteindre la position cible-équilibre (qui ne tient pas compte de l'inertie du système) : on parlera alors d'"undershoot" (figure 3). Cette propriété est, pour la parole, tout à fait fondamentale.

IV-APPLICATION AU MOUVEMENT DU DOS DE LA LANGUE.

IV.1. Les données de départ : Nous avons à notre disposition des mesures ultrasoniques du mouvement du dos de la langue, réalisées à

Montreal (pour plus de détails sur l'expérimentation, voir en particulier Keller et al. [16] [17]). Ces mesures portent sur des réalisations du type [kaka], où le geste consonne-voyelle correspond essentiellement au mouvement haut-bas du dos de la langue, sous différentes conditions de contrôle temporel (débit et accent). Pour ce travail, nous nous sommes tout particulièrement intéressés à l'étude du cycle [ka'ka] (Figure 4) répété par un locuteur de langue anglaise. L'intérêt de ce cycle réside dans la présence successive d'une même voyelle non accentuée puis accentuée. Nous pouvons ainsi nous confronter à la fois au problème des plateaux (déja évoqué pour Browman & Goldstein [13]) et à celui des "undershoot".

IV.2. Le cycle [ka'ka] " : Pour générer ce cycle avec notre modèle nous avons utilisé deux démarches, dont les différences portent sur la voyelle accentuée : dans un cas, elle est le résultat d'une tenue ; dans l'autre elle est le résultat d'un mouvement effectué avec une grande raideur sous contrôle des antagonistes à la cible.

Dans les deux hypothèses, nous avons défini deux cible-équilibre, soit deux valeurs de α. L'une correspond à la cible du [k], l'autre à celle du [a]. Lors des réalisations du [k] (figure 5 et 6), à la fois le temps de transition entre cibles, et le contexte de raideur sont suffisants pour que l'on s'approche au mieux la position cible-équilibre. Pour les réalisations du [a], deux cas se présentent :

* il est non accentué : le mouvement est rapide et le niveau de raideur de co-contraction est faible ; on atteint alors à l'instant cible t_c une position plutôt éloignée de la position cible-équilibre;
* il est accentué : temps et raideur sont suffisants pour qu'on atteigne spatialement la position cible-équilibre.

Dans cette dernière éventualité, on l'a dit, deux cas se présentent encore :

* on maintient les commandes à la cible pendant un temps t_1 ; la tenue est donc le résultat d'une programmation (figure 5);
* la raideur de l'agoniste varie vite et elle atteint une valeur plus grande que celle qui correspond au niveau de co-contraction et à α_{cible} à la cible ; le rapport k_1/k_2 atteint ainsi la valeur α_{cible} avant l'instant t_c ; la raideur de l'antagoniste évolue alors en fonction de celle de l'agoniste de façon à ne pas dépasser la cible-équilibre déterminée par α_{cible}, et par conséquent le niveau de co-contraction varie (figure 6).

Quoi qu'il en soit, l'approximation de l'allure de la trajectoire est satisfaisante, et on retiendra les postulats suivants : une voyelle non accentuée correspond à un mouvement rapide et "mou" donnant une position "neutre" - n'est-on pas en cela en accord avec Lindblom [18] ? - ; une voyelle accentuée est obtenue dans des conditions de temps et/ou de raideur telles que la position cible-équilibre soit atteinte.

V-DISCUSSION.

L'évaluation du modèle proposé est loin d'être achevée. Mais elle présente d'ores et déjà des résultats très intéressants. Au niveau des commandes, on limite en effet les spécifications précises pour chaque muscle, grâce à une approche globalisante agonistes-antagonistes (cf. Hogan in Stein [6]). D'autre part, même si elles demandent à être affinées, les approximations dynamiques sont satisfaisantes, et le modèle montre en particulier sa capacité à modéliser correctement les plateaux, sans pour autant faire appel à une introduction "brutale" d'amortissement, comme le proposaient Browman & Goldstein [13]. La modélisation du second ordre et le contrôle par la raideur permettent aussi un feedback permanent par la prise en compte à tout instant de la position réelle, dans la mesure où la force est ainsi constamment fonction de cette position. Cette propriété laisse augurer un bon comportement du modèle face aux perturbations (y compris pour une modélisation du type feedforward (cf Hinton [19]).

Mais on peut dès maintenant mettre l'accent sur les perspectives moins réjouissantes de notre modélisation. Il nous faudra en effet spécifier sinon la distribution des raideurs de tous les muscles protagonistes, au minimum les niveaux de co-contraction (Vincken et al. [20]). De plus, même avec une bonne approximation des trajectoires, le problème de l'inversion dynamique reste posé, qui offre des solutions multiples, et il restera toujours difficile d'évaluer un modèle trop puissant.

REMERCIEMENTS.

Merci à J.P. Blanqui de LABSYS à Grenoble, T. Gay du "Departement of Oral Biology" de Farmington, J. Laver du "Center for Speech Technology Research" d'Edimbourg, B. Lindblom du "Departement of Linguistic" de Stockholm, S. Maeda du CNET de Lannion, et à J.P. Orliaguet du Laboratoire de Psychologie Expérimentale de Grenoble pour les nombreux et fructueux échanges que nous avons eus avec eux, sur tous nos problèmes.

REFERENCES.

1 SCHMIDT R.A. McGOWN C. QUINN J.T. & HAWKINS B. (1986)
Unexpected Inertial Loading in Rapid Reversal Movements : Violations of Equifinality.
Human Movement Science 5, 263-273.

2 KELSO J.A.S. & TULLER B. (1987)
Intrinsic Tome in Speech Production : Theory, Methodology, and Preliminary Observations.
in Motor and Sensory Processes of Language. Ed. by KELLER E. & GOPNIK M., 203-222.
Lawrence Erlbaum Associates, London.

3 MACNEILAGE P.F. (1970)
Motor Control of Serial Ordering of Speech.
Psychol. Rev. 77, 182-196.

4 POLIT A. & BIZZI E. (1978)
Processes Controlling Arm Movements in Monkeys.
Science 201, 1235-1237.

5 BERKINBLIT M.B. FELDMAN A.G. & FUKSON O.I. (1986)
Adaptability of Innate Motor Patterns and Motor Control Mechanism.
Behavioral and Brain Sciences 9, 585-638.

6 STEIN R. B. (1982)
What Muscle Variable(s) does the Nervous System Control in Limb Movements ?
Behavioral and Brain Sciences 5, 535-577.

7 COOKE J.D. (1980)
The Organization of Simple, Skilled Movements.
in Tutorials in Motor Behavior. Ed. STELMACH G.E. & REQUIN J., 199-212.
North-Holland Publishing Company.

8 ASATRYAN D.G. & FELDMAN A.G. (1965)
Functional Tuning of the Nervous System with Control of Movement or Maintenance of a Steady Posture. I. Mechanographic Analysis of the Work of the Limb on Execution of a Postural Task.
Biophysics 10, 925-935.

9 NELSON W.L. (1983)
Physical Principles for Economies of Skilled Movements.
Biol. Cybern. 46, 135-147.

10 OSTRY D.J. & MUNHALL K.G. (1984)
Control of Rate and Duration of Speech Movements.
J. Acoust. Soc. Am. 77, 640-648.

11 MUNHALL K.G. OSTRY D.J. & PARUSH A. (1985)
Characteritics of Velocity Profiles of Speech Movements.
Journal of Experimental Psychology : Human Perception and Performance vol II, 4, 457-474.

12 MUNHALL K.G. & OSTRY D.J. (1986)
On Mass-Spring Systems as Models of Movement Control.
Manuscrit non publié, McGill University.

13 BROWMAN C.P. & GOLDSTEIN L.M. (1984)
Dynamic Modeling of Phonetic Structure.
Haskins Laboratories : Status Report on Speech Research SR-79/80.

14 ADAMOVICH S.V. & FELDMAN A.G. (1984)
Model of the Central Regulation of the Parameters of Motor Trajectories.
Biophysics 29, 130-134.

15 OSTRY D.J. & COOKE J.D. (1987)
Kinematics Pattern in Speech and Limb Movements.
in Motor and Sensory Processes of Language.
Ed. by KELLER E. & GOPNIK M., 223-235.
Lawrence Erlbaum Associates, London.

16 OSTRY D.J. KELLER E. & PARUSH A. (1983)
Similarities in the Control of the Speech Articulators and the Limbs : Kinematics of the Tongue Dorsum Movement in Speech.
Journal of Experimental Psychology : Human Perception and Performance vol 9, 4, 622-636.

17 KELLER E. (1987)
Factors Underlying Tongue Articulation in Speech.
(à paraître)

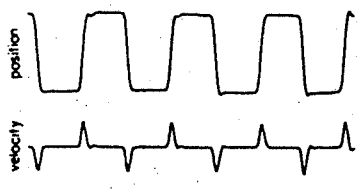
18 LINDBLOM B. (1963)
Spectrographic Study of Vowel Reduction.
J. Acoust. Soc. Am. 35, 1773-1781.

19 HINTON G. (1984)
Parallel Computations for Controlling an Arm.
Journal of Motor Behavior vol 16, 2, 171-194.

20 VINCKEN M.H. GIELEN C.C. & DENIER VAN DER GON (1983)
Intrinsic and Afferent Components in Apparent Muscle Stiffness in Man.
Neuroscience 9, 529-534.

FIGURES

1.a : Chez l'homme (Cooke [7])



1.b : Simulé sur notre modèle.

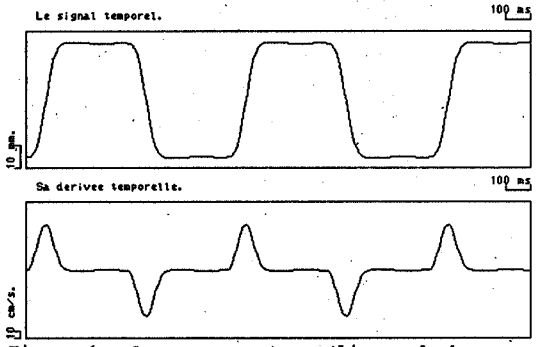


Figure 1 : Les mouvements cycliques du bras.

Figure 2 : Le rapport Vmax/Amp=f(T).

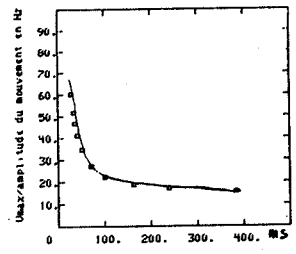


Figure 3 : L'"undershoot".

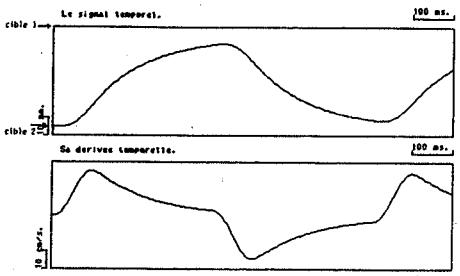


Figure 4 : Le cycle [ka'ka] (échographie)

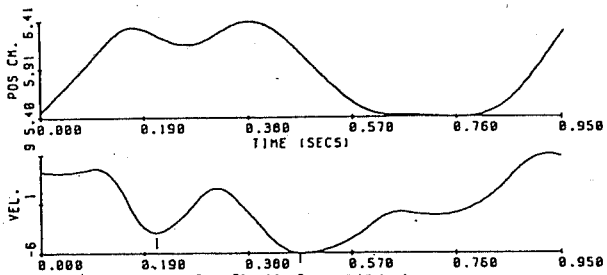


Figure 5 : Le cycle [ka'ka] modélisé avec tenue.

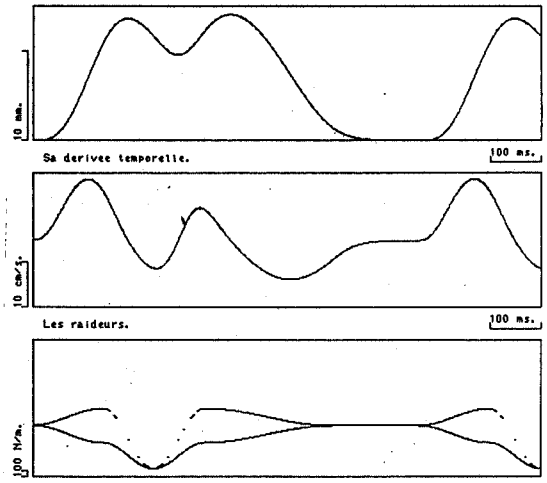
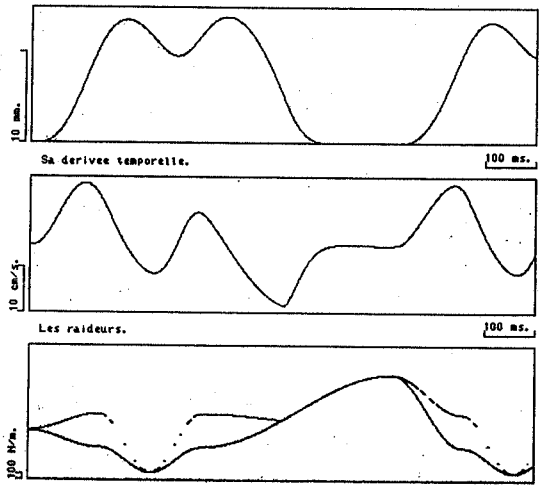


Figure 6 : Le cycle [ka'ka] modélisé sans tenue.



NOMOGRAMMES DU CONDUIT VOCAL PAR MODELISATION ARTICULATOIRE.

Pascal PERRIER[1] Pierre BADIN[1] & Louis Jean BOE[2]

INSTITUT DE LA COMMUNICATION PARLEE, UA CNRS 368.
1 Laboratoire de la Communication Parlée, ENSERG/INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex.
2 Institut de Phonétique de Grenoble, Université III, Domaine Universitaire, 38400 St Martin d'Hères.

ABSTRACT

For the study of vowel production FANT's [1] nomograms is the basic tool. But, even though the vocal tract is fairly well represented by a three - parameter model, it was tempting to compare this representation with more realistic data, namely an articulatory model. With a slightly modified version of MAEDA's articulatory model [3], we generated nomograms with the same conditions as FANT, and we compared the results (1) with FANT's nomograms and (2) LADEFOGED and BLADON's human nomograms [2]. Our simulations show a good agreement with LADEFOGED and BLADON's measures and reconfirm the stability of F2 in the prepalatal region ; however the model leads to a stable F3 in the same region, which is not fairly representative of reality. The effects of bandwith inversion and formant merging in the region of F2/F3 convergence corresponding to [i] are also verified.

I-INTRODUCTION

Dans le cadre de l'étude acoustique de la production des voyelles, les fameux nomogrammes proposés par FANT [1] sont un outil fondamental. Mais s'il est vrai que la caractérisation de la forme du conduit vocal à l'aide de trois paramètres (position et aire de la constriction, aire aux lèvres) s'est révélée extrêmement pertinente, il n'en est pas moins vrai que cette description reste schématique, et qu'avec les progrès de la connaissance des phénomènes de production, la tentation est grande de les confronter à des modélisations plus réalistes ou à des mesures directes sur le conduit vocal. C'est d'ailleurs ce à quoi se sont attachés LADEFOGED et BLADON [2] qui, experts dans le contrôle du geste articulatoire, ont effectué une série de nomogrammes humains, en plaçant leur propre conduit vocal et leurs propres articulateurs dans des conditions autant que possible identiques à celles des nomogrammes de FANT. Même s'ils ont trouvé certains points de désaccord avec celui-ci (sur lesquels nous reviendront dans la suite), la conclusion générale de leur travail est on ne peut plus claire : "These discrepancies, however, are minor in comparison with the major points of agreement between FANT's prediction and the data observed in real speech. FANT's nomograms remain as one of the major achievements of his acoustic theory of speech production."

Si les idées-force tirées des nomogrammes originaux sont donc désormais acquises, l'utilisation de modèles plus réalistes pourra peut-être apporter un point de vue intéressant dans la discussion suscitée par LADEFOGED et BLADON. Mais pour cela, il faut disposer d'un modèle dont la validité ait déjà été largement mise à l'épreuve. C'est le cas du modèle articulatoire de MAEDA [3], [4]. Dans ces conditions, une approche par la modélisation, pourtant réductrice par définition par rapport à l'étude directe de la production humaine, présente l'avantage de la parfaite maîtrise de l'ensemble des paramètres. Rappelons à ce sujet que LADEFOGED et BLADON eux-mêmes ont émis des réserves concernant la précision avec laquelle ils maîtrisaient et évaluaient à la fois la position et l'aire à la constriction.

II-LE MODELE ARTICULATOIRE

Nous avons donc utilisé le modèle de MAEDA qui, rappelons le, a été conçu à partir d'une analyse statistique "intelligente" de 400 radiographies du conduit vocal. Avec 5 paramètres de commande rendant compte des effets de la mâchoire, de la langue et des lèvres, il génère une coupe sagittale du conduit vocal. Le passage à la fonction d'aire est obtenu à l'aide de coefficients résultant de l'étude d'un moulage du conduit vocal [5], et les valeurs formantiques correspondantes sont calculées par un programme de simulation harmonique du conduit vocal incluant les pertes par chaleur, viscosité, vibration des parois, rayonnement, et en supposant la glotte fermée [6]. La position de la constriction est définie comme l'abscisse du milieu de la section d'aire minimale située entre le bas pharynx et les alvéoles.

Nos premières tentatives de génération de nomogrammes à partir de ce modèle ont été infructueues : il était en effet impossible de générer une constriction d'aire proche de 0.65 cm^2 (une des conditions des simulations de FANT) dans les zones uvulaire et alvéo-palatale. Nous avons donc été amenés à modifier la forme des parois fixes du modèle de MAEDA dans ces zones, de façon à pouvoir déplacer la constriction tout au long du conduit vocal. Les coupes sagittales avant et après "chirurgie" sont données à la fig.1. Cette opération de rectification de la voûte palatine peut se justifier compte tenu de la simplification déjà effectuée au niveau de la forme de la langue. Par ailleurs, nous avons vérifié que ces modifications n'affectent pas la répartition des voyelles dans le plan F1-F2.

III-LES NOMOGRAMMES ARTICULATOIRES

III.1. Les conditions d'obtention : Ils ont été générés à l'aide d'un logiciel interactif donnant la coupe sagittale, l'aire de la constriction et son abscisse par rapport à la glotte en fonction des paramètres articulatoires de commande. Les paramètres mâchoire et lèvres sont maintenus constants, et sont choisis de manière à obtenir l'aire aux lèvres désirée et une protrusion des lèvres de 0.5 cm par rapport aux incisives.

Comme dans les nomogrammes de FANT, les valeurs de l'aire aux lèvres ont été choisies égales à 4, 2 ou 0.65 cm² ; l'aire de la constriction est de 0.65 cm² avec un écart n'excédant pas ±5%. Comme chez LADEFOGED et BLADON, la longueur totale du conduit vocal n'est pas maintenue constante (pour nous entre 15.8 et 16.9 cm). Précisons enfin que, la coupe sagittale du modèle de MAEDA étant définie par 26 points (grille linéaire puis semi-polaire), il n'est pas possible de déplacer de manière continue le lieu de constriction. C'est pourquoi les nomogrammes présentés ici sont discrétisés avec un pas moyen de 0.8 cm.

III.2. Les résultats généraux : La figure 2 présente les nomogrammes articulatoires ainsi obtenus. On note d'emblée - et on s'y attendait - la bonne concordance globale avec les nomogrammes de FANT. On remarque en particulier le croisement F2-F3 dans la zone correspondant à des abscisses de la constriction autour de 11 cm [7]. Les variations de F4 sont faibles, ce qui correspond à une stabilité de la longueur équivalente de la constriction. Ceci n'était pas tout à fait le cas pour LADEFOGED et BLADON (p.193), en particulier chez le sujet A.B. Comme LADEFOGED et BLADON nous constatons qu'il n'est pas possible d'obtenir une position de la constriction dans la zone basse du pharynx ou très près des lèvres. Ces auteurs avaient proposé de limiter le domaine de validité des nomogrammes à l'intervalle des abscisses de constriction entre 4 et 13.5 cm. Le modèle de MAEDA n'explore que l'intervalle entre 5.5 et 13.5 cm.

D'autre part, le formant F2 ne décroit pas après avoir atteint son maximum aux environs de 11 cm d'abscisse. Nous confirmons ici les observations faites par LADEFOGED et BLADON qui attribuaient ce phénomène à un applatissement de la langue dans la zone d'articulation entrainant une faible modification de la cavité arrière (à laquelle F2 est dans ce cas affilié). Mais alors que dans leur cas F3 augmentait parallèlement (laissant supposer une variation de la cavité avant), il reste dans notre simulation tout aussi stable que F2. Ceci s'explique assez bien si l'on observe la fig.3, représentant les fonctions d'aire correspondant à l'un de ces tracés : on note en effet une stabilité quasi totale de la fonction d'aire avec seulement un déplacement de la position de l'aire minimale lié à une très faible variation d'aire. Ce léger point de différence avec LADEFOGED et BLADON est très certainement imputable à une modélisation insuffisante de l'apex dans le modèle de MAEDA.

Enfin, nous constatons que F2 converge vers la même valeur, aux alentours de 2200 Hz, quelle que soit l'ouverture aux lèvres. Ce phénomène peut sembler a priori contradictoire avec les résultats de FANT, mais ne l'est que quantitativement : chez ce dernier, les valeurs de F2 pour différentes ouvertures aux lèvres convergent pour des constrictions plus proches des lèvres. Ce phénomène d'indépendance de F2 par rapport aux lèvres peut s'expliquer par le fait que dans cette situation clairement post-focale, F2 est affilié nettement à la cavité arrière, et la fig.4 montre que pour nos nomogrammes, cette cavité varie très peu avec l'aire aux lèvres.

A partir des nomogrammes articulatoires, nous pouvons donc dégager les mêmes observations que LADEFOGED et BLADON à propos des nomogrammes de FANT :

- limitation de la zone de variation du lieu d'articulation,
- non décroissance de F2 dans la zone alvéolaire.

IV-LE CAS DU [i]

Dans le cadre d'une reconsidération des nomogrammes de FANT [8], nous avions étudié les effets de croisement des formants, et en particulier le cas intéressant de la convergence F2/F3 correspondant au [i], et ceci avec le modèle simple à quatre tubes proposé par FANT. Ici nous avons réalisé un nomogramme dans les conditions de constriction et d'ouverture aux lèvres proposées pour le [i] par MAJID et al. [4] (Aire aux lèvres: 4 cm², aire à la constriction : 0.27 cm²). Sur la fig.5, nous retrouvons clairement le croisement F2-F3 caractéristique du [i] focal. De même, la Fig.6 montre le phénomème d'"inversion de bande passante" mis en évidence dans [8]. Dans la région "préfocale" (c'est-à-dire entre la glotte et le point focal) le formant F2 est associé à la cavité avant et donc, à cause de pertes liées au rayonnement aux lèvres, son amplitude est faible et sa bande passante est large, tandis que F3, lié à la cavité arrière (on suppose la glotte fermée, et donc les pertes faibles) présente une amplitude plus forte et une bande passante plus faible (fig.6a). Dans la région post-focale (fig.6b), le phénomème inverse se produit : F2 est lié à la cavité arrière et son amplitude est forte, alors que F3 est lié à la cavité avant et donc amorti par les pertes aux lèvres. Nous pouvons enfin imaginer un [i] focal entre ces deux régions, mais, à cause des limitations dues à la grille de définition de la fonction sagittale pour le modèle, il n'a pas été possible d'atteindre exactement ce point particulier. Rappelons simplement que c'est cette "fusion" entre F2 et F3 dans la région du [i] qui avait suscité la perplexité de LADEFOGED et BLADON ([7],[8]).

Les résultats obtenus pour le [i] sont donc en parfaite concordance avec les résultats prévus sur un modèle simplifé, et se retrouvent également dans la réalité du signal de parole ([8]).

V-CONCLUSIONS.

A l'aide du modèle articulatoire de MAEDA, légèrement modifié, nous avons généré des nomogrammes articulatoires dans des conditions identiques à celles de FANT. L'allure générale des nomogrammes originaux est bien retrouvée, et nous confirmons les remarques de LADEFOGED et BLADON (limitation de la zone possible d'articulation, stabilité de F2 dans la zone post-focale). Ceci peut être interprété soit comme la confirmation de l'excellente maîtrise articulatoire des deux phonéticiens, soit comme une relative insensibilité des données acoustiques à certaines variations géométriques.

Notre travail met une fois encore en évidence l'excellente adéquation entre le modèle articulatoire utilisé et les données physiologiques.

REMERCIEMENTS

A Christian Abry pour ses commentaires et critiques.

BIBLIOGRAPHIE.

1 FANT G. (1960)
Acoustic Theory of Speech Production.
S - Gravenhage : Mouton & Co.

2 LADEFOGED P. & BLADON A. (1982)
Attempts by Human Speakers to Reproduce FANT's Nomograms.
Speech Comm. 1, 185-198.

3 MAEDA S. (1979)
An Articulatory Model of the Tongue based on a Statistical Analysis.
J. Acoust. Soc. Am. 65, S22.

4 MAJID R. ABRY C. BOE L.J. & PERRIER P. (1987)
Contribution à la classification articulatori-acoustique des voyelles : étude des macro-sensibilités à l'aide d'un modèle articulatoire.
11th Int. Congr. Phonetic Sci., A Paraître.

5 SANCHEZ H. & BOE L.J. (1984)
De la coupe sagittale du conduit vocal.
Bull. Inst. Phonétique de Grenoble 13, 1-24.
6 BADIN P. & FANT G. (1984)
Notes on Vocal Tract Computation.
STL-QPSR 2-3/1984, 53-108.
7 ABRY C. & BOE L.J. (1986)
Nomogrammes et systèmes vocaliques.
15 JEP GALF, 303-306.
8 BADIN P. & BOE L.J. (1987)
Vocal Tract Vocalic Considerations.
A Crucial Problem : Formant Convergence.
11th Int. Congr. Phonetic Sci., A Paraître.

FIGURES

Figure 1 : Coupe sagittale type générée par le modèle après modification de la voûte palatine.

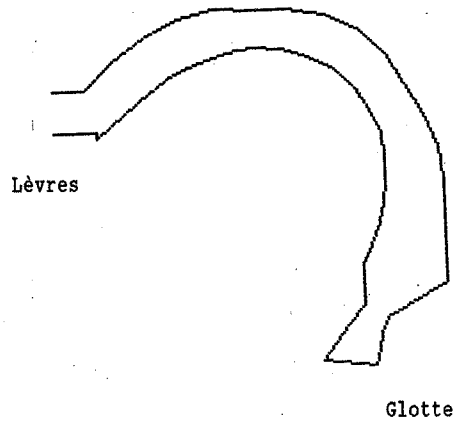


Figure 2 : Nomogrammes articulatoires pour une aire à la constriction égale à 0.65 cm², pour 3 valeurs de l'aire aux lèvres, et pour une protrusion égale à à 0.5 cm.
------ : aire aux lèvres 4 cm²
——— : aire aux lèvres 2 cm²
-•-•- : aire aux lèvres 0.65 cm2

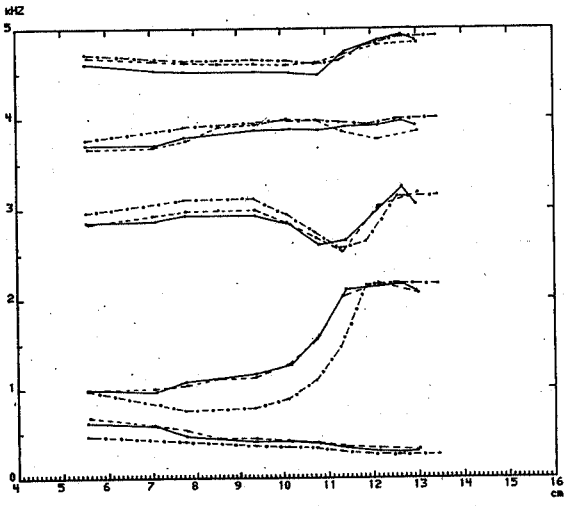


Figure 3 : Différentes fonctions d'aire après le point focal, pour une aire aux lèvres égale à 2 cm².

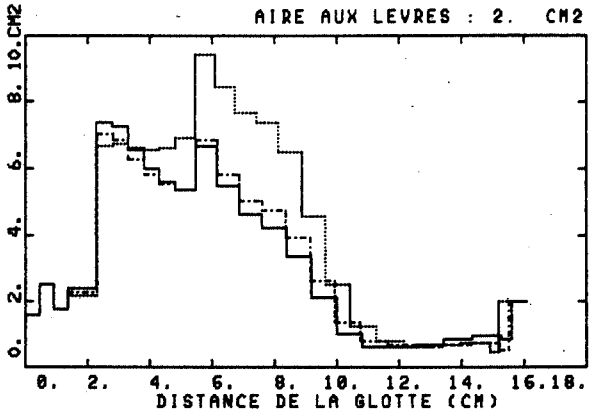


Figure 4 : Fonctions d'aire dans la zone post-focale pour les trois aires aux lèvres (4, 2, 0.65 cm²).

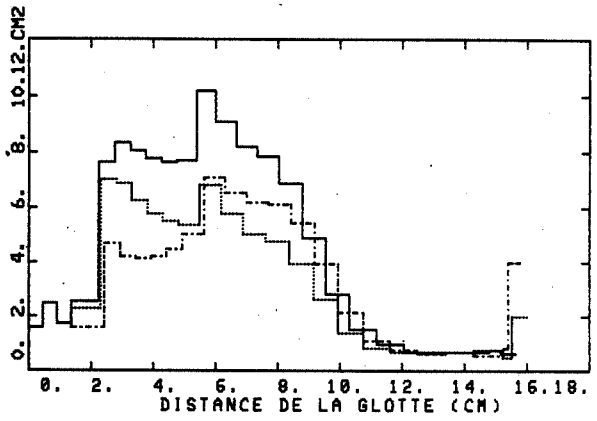


Figure 5 : Nomogrammes autour du [i].

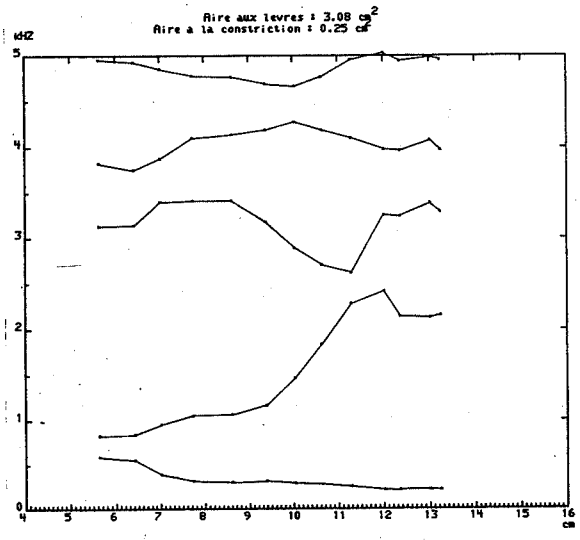
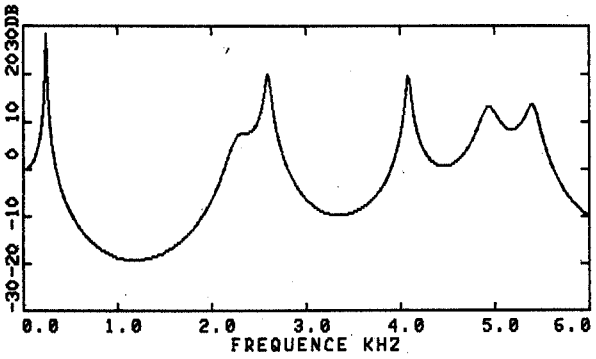
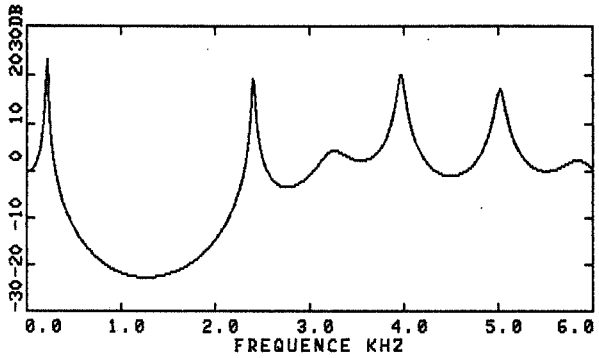


Figure 6 : Fonctions de transfert autour du [i] focal.

6.a : [i] préfocal.



6.b : [i] postfocal



PASSAGE DE LA COUPE SAGITTALE A LA FONCTION D'AIRE.
Les zones de faibles dimensions sagittales.

Pascal PERRIER[1] & Louis Jean BOE[2]

INSTITUT DE LA COMMUNICATION PARLEE, UA CNRS 368.
1 Laboratoire de la Communication Parlée, ENSERG/INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex.
2 Institut de Phonétique de Grenoble, Université III, Domaine Universitaire, 38400 St Martin d'Hères.

ABSTRACT

Starting (1) from a midsagittal section of the vocal tract generated by an articulatory model and (2) from standard vowel area functions, we calculated, for each French vowel, the relationship coefficients between the cross-sectional area and the midsagittal distance. The results reconfirm the validity of average coefficients, as far as the sagittal dimension is not too small. However, the coefficients tends to increase for midsagittal distances lower than 8 mm.

I-INTRODUCTION.

La compréhension des phénomènes articulatoires et acoustiques de la production de la parole a véritablement progressé au cours de la dernière décennie, et ceci principalement grâce aux développements de modèles, articulatoires d'une part, qui génèrent une coupe sagittale du conduit vocal (Lindblom & Sundberg [1], Maeda [2]) et acoustiques d'autre part, qui pour une fonction d'aire donnée génèrent un signal de parole (Maeda [3], Shadle [4]). Mais, par delà l'affinement des connaissances et des techniques, l'association de ces deux types de modèles, c'est à dire le passage de la coupe sagittale à la fonction d'aire, présente encore bien des difficultés.

Et pourtant les travaux sur ce point de passage obligé de la modélisation articulatoire, ne manquent pas, que ce soit au niveau de la recherche de modèles mathématiques (Heinz et Stevens [5], Maeda [6]), ou au niveau des mesures, effectuées soit directement sur le conduit vocal (Fant [7], Sundberg [8], Sanchez & Boë [9]), soit indirectement, à partir des caractéristiques acoustiques du signal (Mermelstein [10]). En outre, tout le monde semble être d'accord sur l'utilisation du modèle de Heinz et Stevens [5] (ou de ses dérivés), $A=\alpha.d^{\beta}$, dont les coefficients varient tout au long du conduit vocal. Malgré cela, indépendamment des différences entre les divers jeux de coefficients proposés par chaque auteur, une question importante reste posée : ces coefficients sont ils valables, quelle que soit la valeur de la dimension sagittale ?

Sur ce point, quelques études apportent déjà des éléments de réponse : Sundberg [8], en étudiant les tomographies du pharynx, remarquait que les parois latérales se rapprochaient lorsque la distance entre la langue et la paroi arrière du pharynx excédait 1.9 cm ; cette tendance a récemment été confirmée par des observations au scanner (Johansson et al. [11]).

Ce travail n'a pas pour objectif de proposer un nouveau jeu de coefficients, ou un découpage plus judicieux du conduit vocal. Il se limite à apporter quelques éléments supplémentaires à la réflexion générale sur ce point difficile.

II-NOTRE POINT DE DEPART.

Nous sommes partis des standard vocaliques du français obtenus à partir du modèle articulatoire de Maeda (Majid et al. [12]). Nous avions, pour cela, appliqué aux coupes sagittales générées par le modèle, des coefficients obtenus par moulage sur un cadavre (Sanchez et Boë [9]). Ces ceux-ci se sont dans l'ensemble révélés corrects (Perrier et al. [13]), puisque les fonctions d'aire correspondantes étaient globalement conformes à celles qui avaient été publiées par ailleurs (Fant [7]). Mais une analyse plus approfondie a mis en évidence certaines difficultés, en particulier pour la voyelle [i], dont la génération acoustique a été délicate, bien que la coupe sagittale ait été apparemment très correcte. Or le [i] présente une faible aperture et sa sensibilité à ce paramètre est très grande. Ceci confirme la nécessité d'affiner tout particulièrement le calcul du passage de la coupe sagittale à la fonction d'aire pour les petites sections.

C'est l'objet de ce travail, qui reste, précisons le, strictement dans le cadre de la modélisation.

II-NOTRE DEMARCHE.

A partir du modèle articulatoire de Maeda, dont la validité a déjà été largement prouvée, nous avons déterminé les coupes sagittales des onze voyelles du français (Majid et al. [12]). Pour cela nous avons respecté une cohérence structurelle, qui intégre les connaissances générales sur les voyelles, à la fois sur les dispositions relatives des principaux articulateurs et sur les coupes sagittales radiographiques.

Parallèlement nous disposons de fonctions d'aire type, qui sont le résultat d'études successives sur le plan articulatoire et acoustique (Mrayati [14], Feng [15]). Elles ont été validées, à la fois par une étude harmonique, et par des vérifications perceptives.

Notre objectif a été alors de faire le lien entre ces deux champs de connaissances, par l'analyse du passage d'une coupe sagittale à la fonction d'aire correspondante. Nous avons, pour cela, supposé valable la formule de Heinz & Stevens [5] ($A=\alpha.d^{\beta}$), β étant choisi égal à 1.5.

La première difficulté à contourner a été la différence entre les longueurs respectives des coupes sagittales et des fonctions d'aire. Notre outil fondamental étant le modèle articulatoire, toutes les longueurs ont été ramenées à celles des coupes sagittales. Mais une simple homothétie présentait l'inconvénient de ne pas faire correspondre les positions respectives du lieu d'articulation, dont l'importance dans la production des voyelles est bien connue. C'est pourquoi, nous avons appliqué une double homothétie, portant respectivement sur les cavités antérieures et les cavités postérieures.

III-LES RESULTATS.

Nous avons ensuite calculé, pour chaque voyelle, les coefficients α de passage des coupes sagittales à ces fonctions d'aires. La figure 1 présente, section par section, l'ensemble des valeurs ainsi obtenues ainsi que la valeur moyenne correspondante.

III.1. La spécificité de la zone du lieu d'articulation : Observons d'abord uniquement les résultats obtenus en dehors des lèvres (qui sont, on le sait, une zone bien particulière (cf. Abry & Boë [16])), c'est à dire les sections 1 à 23. Une première conclusion s'impose : pour la plupart des sections (4, 9,10,11 et 15 à 23), l'écart-type est important. A priori, il semble donc difficile de penser pouvoir appliquer, pour chacune des sections, des coefficients α valables quelles que soient les voyelles. Mais observons maintenant la figure 2. Elle représente l'ensemble des valeurs α obtenues pour des dimensions sagittales supérieures à 0.8 cm. La disparité des valeurs n'a certes pas totalement disparu, mais elle est fortement réduite et est localisée aux sections 4, 15, 16, 17, 22 et 23. Ceci justifie l'hypothèse suivante : la grande valeur de l'écart-type sur les valeurs α est due à la contribution des dimensions sagittales inférieures à 0.8 cm. Pour vérifier cette hypothèse, nous avons pris comme valeurs de référence pour les coefficients α, celles qui correspondent aux moyennes M_i pour les distances sagittales supérieures à 0.8 cm. Nous avons alors calculé pour chaque section et chaque voyelle, les écarts entre les coefficients effectivement trouvés (figure 1) et ces valeurs de référence. La figure 3, qui présente ces écarts, en correspondance avec la distance sagittale, est assez éloquente : de manière générale, les gros écarts se situent dans la zone du lieu d'articulation, pour les distances sagittales inférieures à 0.8 cm, ainsi que pour la section 4, qui présente d'ailleurs l'écart-type le plus important. Si on fait abstraction de ce dernier cas, qui est très spécifique - car il correspond à la jonction des cavités laryngales et pharyngales, qui n'est pas prise en compte dans les fonctions d'aire de référence -, la spécificité de la zone du lieu d'articulation se confirme pour les faibles dimensions sagittales. Ailleurs, même si l'écart n'est pas nul, on peut considérer, que l'utilisation des coefficients moyens M_i est correcte. Cette considération est justifiée par ailleurs, par une étude sur les macro-sensibilités des fonctions d'aire des voyelles du français, qui montre bien la grande sensibilité dans la zone du lieu d'articulation et la relative tolérance en dehors de cette zone (Hassan & Perrier [17]). Par conséquent, il est effectivement justifié, comme cela s'est fait jusqu' à aujourd'hui, d'utiliser des coefficients moyens applicables à toutes les voyelles, à condition cependant que les distances sagittales ne soient pas trop faibles. Par contre il apparaît tout à fait clairement que le cas des dimensions inférieures à 0.8 cm doit être traité de manière spécifique.

III.2. Quels coefficients pour les petites distances sagittales ? : Si on observe les résultats présentés figure 3, on constate qu'il n'est pas simple de dégager une loi pour les distances sagittales inférieures à 0.8 cm. En effet, si pour le pharynx (correspondant aux voyelles [a] et [ɑ]) il semble préférable d'utiliser des coefficients supérieurs aux coefficients moyens M_i (ce qui irait dans le sens des observations de Sundberg [8] et Johansson et al. [11]), ce n'est pas aussi simple dans la zone palatale. Une première approche simplificatrice consiste à prendre, dans ce cas, la moyenne mi des coefficients pour chaque section.

Posons alors les coefficients α, égaux aux valeurs M_i pour les distances sagittales supérieures à 0.8 cm, et aux valeurs m_i dans le cas contraire : pour chaque voyelle, les écarts entre les coefficients calculés en III.1 (figure 1) et ces valeurs-type, sont effectivement réduits par rapport à ceux de la figure 3 (cf. figure 4). Ils ne sont cependant encore, dans certains cas, pas vraiment négligeables, ce qui souligne la nécessité d'une étude plus approfondie de ce type de configurations dans le conduit vocal.

IV. CONCLUSIONS.

Conformément à ce que l'on avait soupçonné, cette étude met clairement en évidence le peu de précision des coefficients moyens de passage de la coupe sagittale à la fonction d'aire, pour les faibles dimensions sagittales (ici inférieures à 0.8 cm). Un double jeu de coefficients prenant en compte cet état de fait a certes permis une amélioration des fonctions d'aire obtenues, mais ces résultats sont encore trop approximatifs.

La nécessité d'acquérir des données par mesures directes en trois dimensions (scanner) sur le conduit vocal de locuteurs, s'impose toujours un peu plus. Mais compte tenu du temps de mesure, une coupe sur l'ensemble du conduit vocal est actuellement impossible. Grâce à cette étude, nous savons qu'un tel travail n'est pas indispensable - les coefficients obtenus par moulage étant satisfaisants pour les dimensions sagittales pas trop petites - et qu'il suffira de se concentrer sur la zone du lieu d'articulation. Nous envisageons une campagne de mesures au scanner faite en liaison avec le Centre Hospitalier Universitaire de Grenoble.

REFERENCES.

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

2 MAEDA S. (1979)
An Articulatory Model of the Tongue Based on a Statistical Analysis.
J. Acoust. Soc. Am. 65, S22.

3 MAEDA S. (1986)
Acoustique du relachement des occlusives : une étude de simulation .
15èmes JEP GALF, 317-320.

4 SHADLE C.H. (1985)
The Acoustics of Fricative Consonnants.
Ph. D. Diss., MIT.

5 HEINZ J.M. & STEVENS K.N. (1965)
On the Relations between Lateral Cineradiographs Area Functions, and Acoustics Spectra of Speech.
5th. Int. Cong. Acoust., Liège.

6 MAEDA S. (1972)
On the Conversion of Vocal Tract X-Ray Data into Formant Frequencies.
Bell Laboratories.

7 FANT G. (1960)
Acoustic Theory of Speech Production.
S - Gravenhage : Mouton & Co.

8 SUNDBERG J. (1972)
On the Problem of Obtaining Area Function from Lateral X-Ray Pictures of the Vocal Tract.
STL-QPSR 1,43-45.

9 SANCHEZ H. & BOE L.J. (1984)
De la coupe sagittale à la fonction d'aire du conduit vocal.
13èmes JEP GALF, 23-25.

10 MERMELSTEIN P. (1967)
Determination of the Vocal Tract Shape Measured Formant Frequencies.
J. Acoust. Soc. Am. 41, 1283-1294.

11 JOHANSSON C. SUNDBERG J. WILBRAND H. & YTTERBERGH C. (1983)
From Sagittal Distance to Area.
STL/QPSR 4, 39-49.

12 MAJID R., ABRY C., BOE L.J. & PERRIER P. (1987)
Contribution à la classification articulatori-acoustique des voyelles : Etude des macro-sensibilités à l'aide d'un modèle articulatoire.
11th Int. Congr. Phonetic Sci., A paraitre.

13 PERRIER P., BOE L.J., MAJID SHIBAB R. & GUERIN B. (1984)
Modélisation articulatoire du conduit vocal. Exploration et exploitation.
14èmes JEP GALF, 55-58.

14 MRAYATI M. (1976)
Contribution aux études sur la production de la parole. Modèle électrique du conduit vocal avec pertes, du conduit nasal et de la source vocale. Etude de leurs interactions. Relations entre disposition articulatoire et caractéristiques acoustiques.
Thèse d'Etat, INP Grenoble.

15 FENG G. (1986)
Apport de la modélisation au traitement du signal de parole, le cas des voyelles nasales et la simulation des pôles et des zéros.
Thèse Sciences, INP Grenoble.

16 ABRY C. & BOE L.J. (1986)
"Laws" for lips.
Speech Comm. 5, 1.

17 HASSAN O. & PERRIER P. (1987)
Etude des macro-sensibilités des fonctions d'aire des standard vocaliques du français.
Bull. Inst. Phonétique Grenoble, A paraître.

Figure 1 : Coefficients calculés pour les onze voyelles, avec la moyenne par section.

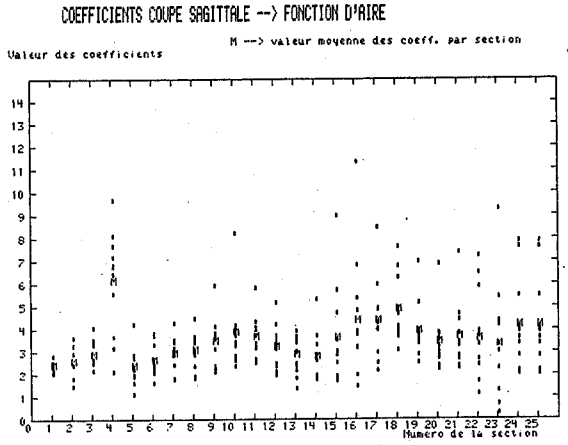
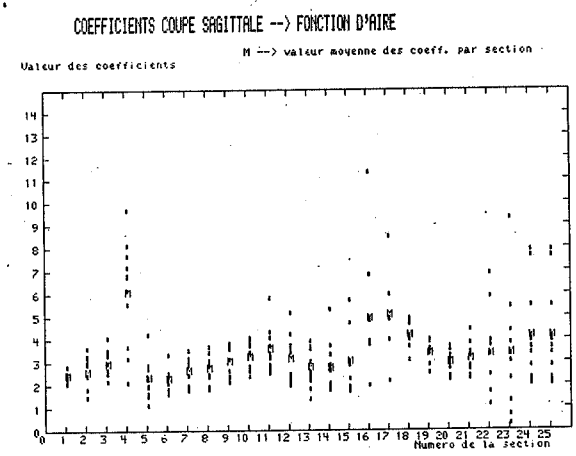
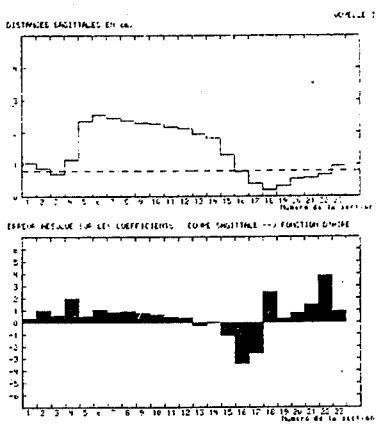


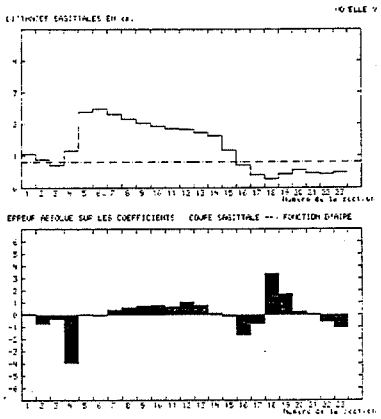
Figure 2 : Coefficients pour les onze voyelles et pour une dist. sagit. > 8 mm.



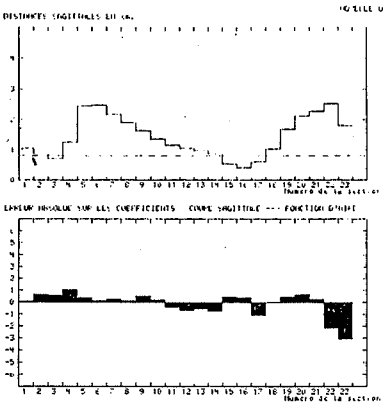
3.a : le [i]



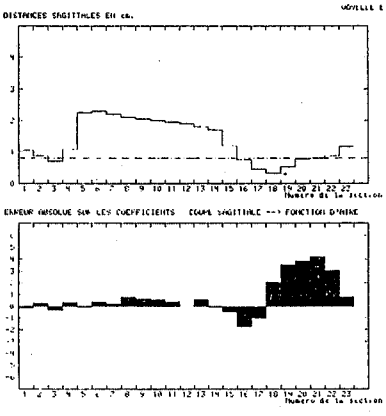
3.b : le [y]



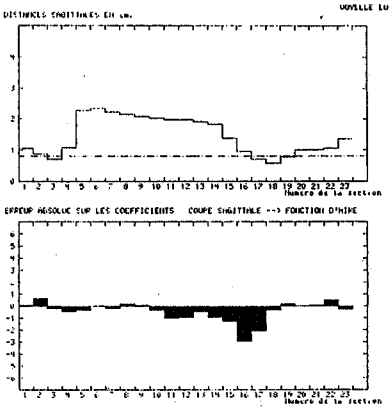
3.c : le [u]



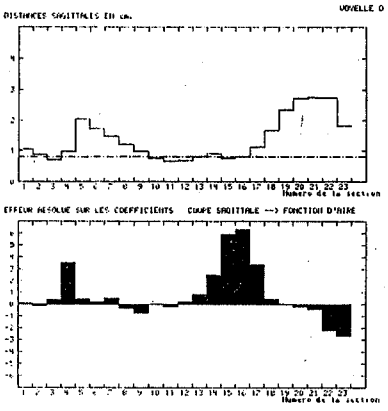
3.d : le [e]



3.e : le [ø]



3.f : le [o]



3.g : le [a]

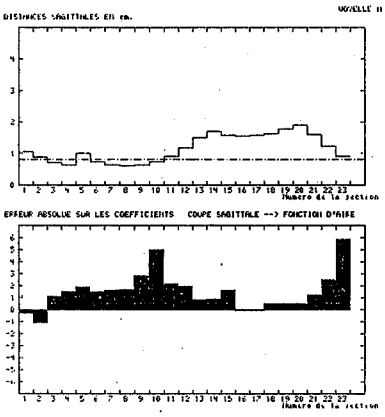
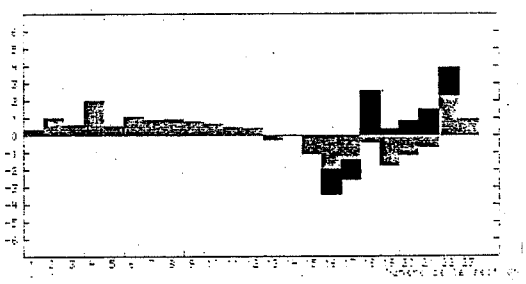
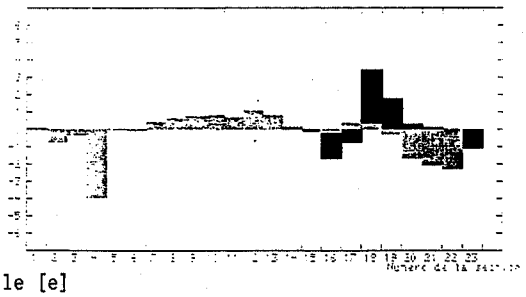


Figure 3 : Distances sagittales et écarts entre les coefficients calculés et les coefficients moyens.

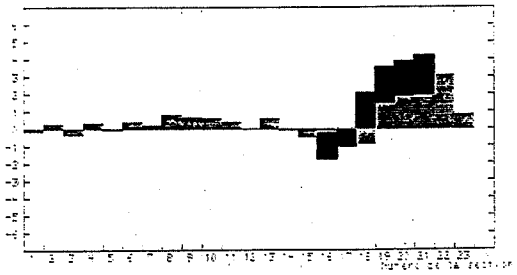
4.a : le [i]



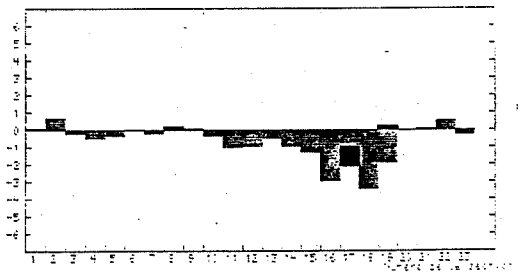
4.b : le [y]



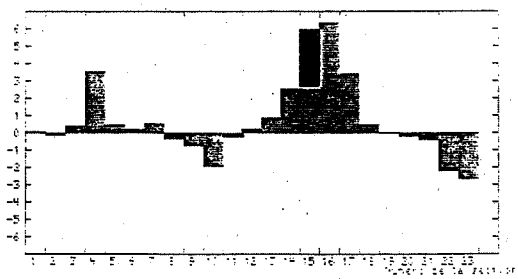
4.c : le [e]



4.d : le [ø]



4.e : le [o]



4.f : le [a]

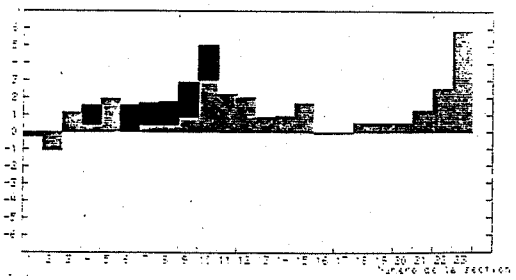


Figure 4 : En noir : écarts par rapport aux moyennes obtenues pour les valeurs de la dist. sagittale supérieures à 8 mm (cf. Fig.3)
En gris : écarts par rapports aux coefficients prenant en compte la spécificité des cas où les distances sagittales sont inférieures à 8 mm.

ETUDE D'UN AEROPHONOMETRE DE GRANDE DYNAMIQUE ET FAIBLE CONSTANTE DE TEMPS

Bernard TESTON

INSTITUT DE PHONETIQUE U.A 261 CNRS AIX-EN-PROVENCE

ABSTRACT.

We describe a system for the measurement of oral and nasal airflow in speech. It is based on the measurement of a differential pressure across a stainless steel mesh.

The device is designed to optimize three criteria: wide dynamic, low response time, and low linearity disturbances obtained by reducing the aerodynamic turbulence effects on the transducers.

INTRODUCTION:

La mesure des paramètres aérodynamiques de la parole, se heurte à deux problémes fondamentaux et contradictoires, qui sont d'une part, l'imprécision des mesures provoquées par les phénomènes de turbulence incontournables, et, d'autre part, par la faible constante de temps de l'instrument nécessaire pour la mise en évidence de transitoires rapides. A ces problèmes fondamentaux, auxquels il faut ajoutter une grande dynamique de mesure, se greffent d'autres problèmes plus spécifiques à la mesure des échanges pendant la phonation ou la respiration, tels que les conditions thermodynamiques du fluide, et, l'adaptation des capteurs de débit aux sujets (TESTON-1980).

Si les mesures des échanges respiratoires peuvent ètre considérées comme satisfaisantes celles des échanges en phonation ne le sont pas du tout, essentiellement au plan des temps de réponse et de la sensibilité des capteurs. Il est à noter que les embouchures utilisées pour le prélèvement des fluides ne doivent pas empécher les sujets dé parler....

C'est dans le cadre de l'amélioration de la mesure des paramètres aérodynamiques de la parole que se situe la présente étude.

LE POLYPHONOMETRE III:

Cet appareil a été réalisé dans le cadre d'un contrat de génie biologique et médical INSERM. Il est constitué par un ensemble de capteurs qui permettent la mesure simultanée:

-du débit d'air buccal, inspiré et expiré (DABI et DABE).
-du débit d'air nasal, inspiré et expiré (DANI et DANE).
-des volumes d'air buccal, inspiré et expiré (VABI et VANE).
-des volumes d'air nasal, inspiré et expiré (VANI et VANE).
-du phonogramme buccal.
-du phonogramme nasal.
-de l'activité vibratoire du larynx.

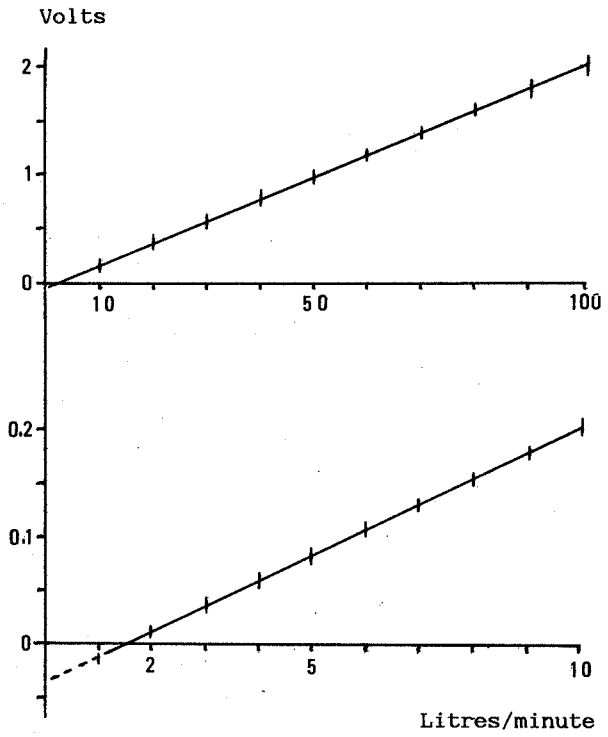
De tous ces paramètres, les débits d'air sont les plus importants, et leurs mesures les plus problèmatiques. Nous utilisons pour cela, le principe bien connu de la mesure de la différence de pression aux bornes d'une résistance à l'écoulement du fluide. Cette résistance est constituée par une grille calibrée en acier inoxydable. Le principe de ce capteur de débit est développé par BERANEK (1955) et il nous semble le mieux adapté à nos problèmes (TESTON-1980). Compte tenu de la rapidité avec laquelle nous devions répondre au contrat, nous avons dessiné un capteur de débit en fonction des capteurs de pression disponibles les mieux adaptés à nos préoccupations. Pour cela, nous nous sommes servi des évaluations que nous avions menées précédemment sur l'influence des volumes morts et les surpressions dans le conduit vocal, provoquées par la résistance à l'écoulement (TESTON-1975).

Les capteurs sont des VALIDYNE homologués pour les mesures de ventilomètrie respira-

piratoire.Ils fonctionnent selon le principe de la variation de mutuelle induction ou transconductance.Leur sensibilité est de 2 mB pleine échelle et leur bande passante est de 1 kH.Associés à une grille de 700 mm2 avec des mailles de 224 micons et un diamètre de fil de 200 microns,ils donnent un signal de 10 Volts pour un débit de 500 litres par minute.Cette valeur représente le débit maximum en respiration forcée rapide.La linéarité de ce capteur est trés satisfaisante,la régression linéaire entre 1 et 100 l/m est de 0.9921 (figure 1).Si l'on considère que le débit normal d'élocution est de l'ordre de 50 l/m,la surpression provoquée par la grille dans le conduit vocal n'est que de 0.2 mB ce qui représente le 1/100 de la PIO moyenne. La dynamique de mesure efficace est de 50 dB (valeur minimale de 1 l/m).

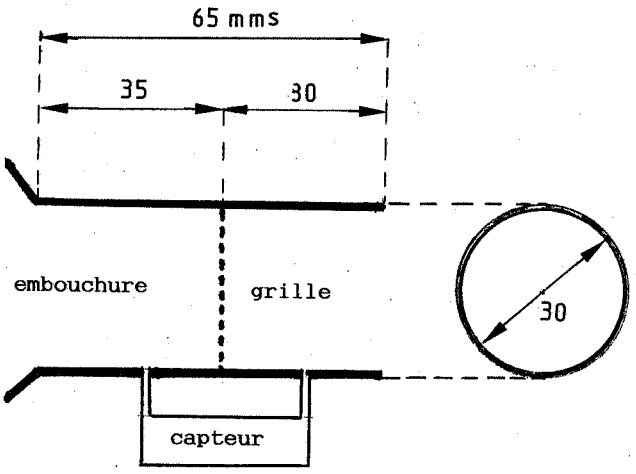
Outre sa grande dynamique de mesure et une bonne symétrie du canal buccal(VABI/VABE ≃1.01),le Polyphonomètre III permet une différenciation totale des débits d'air au nez et à la bouche,et,une trés bonne ergonomie grace à des interfaces capteurs-sujets trés agréables qui laissent à ces derniers toute liberté d'élocution,contrairement aux habituels masques d'anesthésie. Le positionnement du canal nasal au dessous du canal buccal autorise un écoulement naturel du débit nasal,assurance d'une mesure améliorée par rapport aux systèmes précédents.Cependant,la symétrie du canal nasal est perturbée par les faibles sections des conduits aux narines surtout en respiration nasale forcée(VANE/VANI≃1.20).

Le conduit de mesure a un diamètre de 30 mm pour une longueur de 65 mm (figure 2). Ces dimensions représentent un volume mort trés important.L'ensemble embouchure conduit grille,se comporte comme un filtre passe-bas (figure 3).Dés la réalisation de ce système,nous savions qu'il nous faudrait l'améliorer sur ce point.Nous avons également toujours à l'esprit de la linéarité des mesure en milieux turbulent.Ceci a été la principale préoccupation des concepteurs de pneumotachographes respiratoires depuis FLEISCH (1925).Ils ont tous recherché l'écoulement laminaire de l'air.Malheureusement il ne nous est pas possible d'utiliser ces appareils à cause de leur forte constante de temps.ROTHENBERG (1977) ignore élégamment ce problème en considérant que le débit d'air buccal n'est qu'une suite de transitoires.



Signal de sortie en fonction du débit

Figure 1



shéma de principe du capteur de débit d'air buccal

Figure 2

L'AEROPHONOMETRE:

C'est l'appelation que nous donnons à ce nouveau dispositif,car il représente à notre avis la synthèse des meilleurs compromis pour mesurer les paramètres aérodynamiques pendant la phonation.

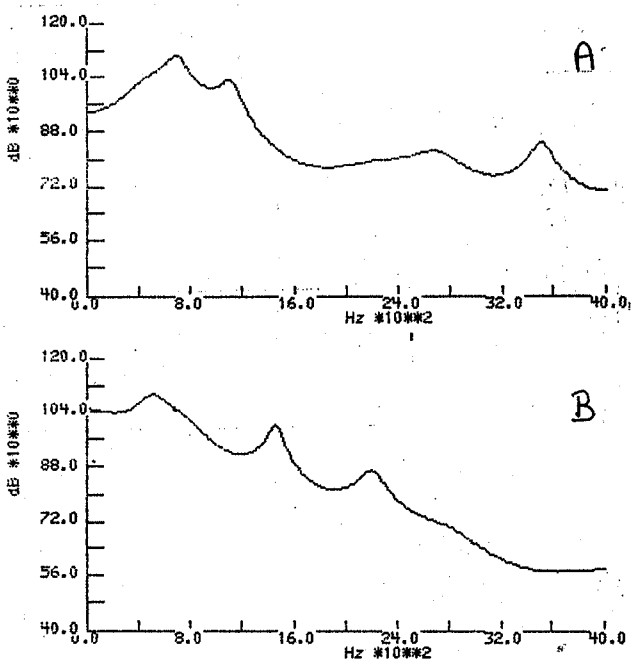
Nouus avons développé ce dispositif à l'apparition sur le marché industriel de capteurs de pression état solide tout silicium rapides et sensibles.Ils ne sont disponibles avec une sensibilité de 20 mB que depuis

deux ans.Nous conservons la mème surface de grille,c'est à dire 700 mm2.La longueur du canal de mesure est réduite à 20 mm,ce qui représente une diminution supérieure à 70%.Les prises de pression sont au nombre de 6 réparties selon une distribution circulaire de 60° (figure 4) à la périphérie du canal de mesure.Chaque prise de pression est équipée d'un capteur différentiel SIEMENS KPY 31 R de 20 mB de sensibilité et de 10 kH de bande passante.Cette disposition des prises de pression permet une grande amélioration de la non linéarité provoquée par les turbulences.En effet la pression est ainsi une valeur moyenne des pressions mesurées à six endroits différents du conduit.Pour affaiblir encore les turbulences aux lèvres,nous avons ajoutté une seconde grille avant les prises de pression.Cette grille ne participe pas à la mesure du débit ,elle "calme" l'écoulement du fluide.Elle est d'une résistance cinq fois plus faible que celle de la grille de mesure,ceci pour ne pas trop surpresser le conduit vocal. Les six capteurs, montés en parallèlle ont une sensibilité de 20/6=3.33 mB.Ce montage est donc moins sensible que les capteurs VALIDYNE utilisés précèdemment.Cependant leur disposition en parallèlle permet une amélioration de la sensibilité par une diminution du bruit total du dispositif et,partant,une augmentation de la dynamique (figure 5).

La dérive des capteurs est parfaitement maitrisée grace à une régulation en température de tout l' ensemble de mesure par un thermostat à 35° qui évitte les condensations dans les conduits.La géométrie du conduit de mesure nasal est améliorée pour diminuer le rapport VANE/VANI de manière sensible en supprimant les trajets en angle droit.

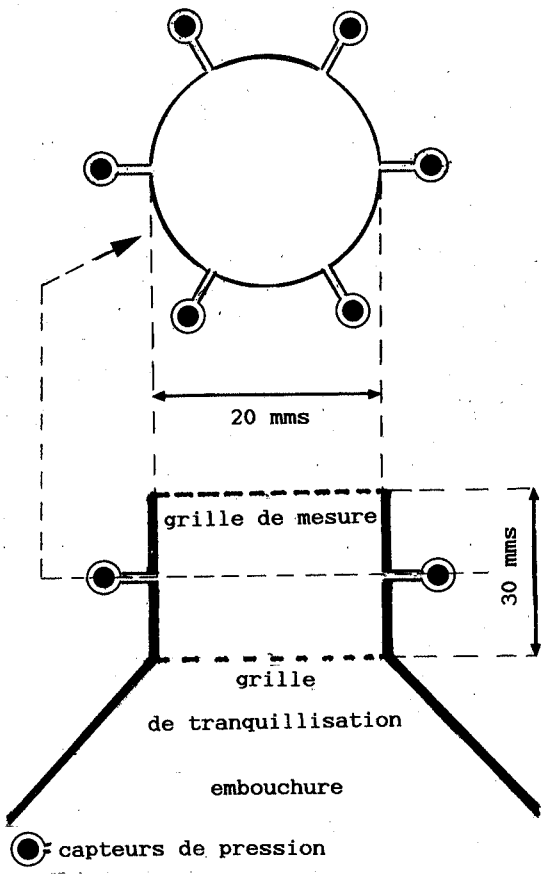
La grande bande passante des capteurs de pression,permet de récupérer les phonogrammes au moyen d'un filtrage passe-haut du signal.On supprime ainsi les microphones de mesure des phonogrammes ce qui simplifie l'ensemble du dispositif de mesure.

Ce dernier peut ètre maintenant reproduit sans difficultés à un prix raisonnable.Toutes ses caractéristiques de mesure ont été améliorées,dont certaines, dans de grandes proportions.En conclusion,nous avons l'impression,pour la première fois,d'utiliser un outil bien adapté pour la mesure des paramètres aérodynamiques de la parole,sans pour autant le considérer comme l'instrument idéal.



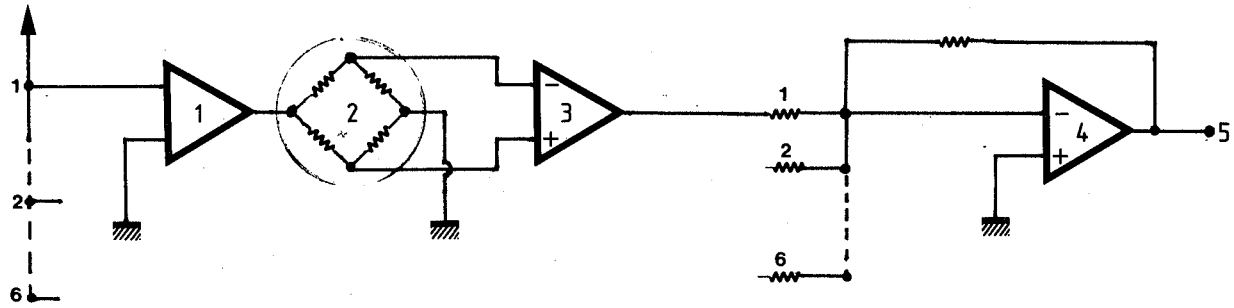
En A,spectre d,une voyelle /a/ émise en champ libre.En B,spectre de la mème voyelle,émise dans l'embouchure du Polyphonomètre.

Figure 3



shéma de principe du capteur de débit de l'aérophonomètre (canal buccal)

Figure 4



-1-Alimentation individuelle des capteurs-2-Capteur à pont de jauges piézo-résistives SIEMENS KPY 31 R-3-Amplificateur différentiel-4-Amplificateur de sommation-5-Signal de sortie dont le bruit total est en 1/V6 du bruit d'un seul capteur.

Shéma de principe de la chaine d'amplification des capteurs.

figure 5

BIBLIOGRAPHIE:

BERANEK, L.L. (1954)
Acoustics, Mc Graw Hill, New-York, 133-136.

FLEISCH, A. (1925)
"Der Pneumotachograph; ein apparat zur bei-schwinddigkeitregstrierung der atemluft."
Arch. ges. Physiol., 209, 713-722.

ROTHENBERG, M. (1977)
"Measurement of airflow in speech."
J. Speech Hearing Res., 20, 146-154.

TESTON, B. (1975)
"Description d'un système d'analyse des paramètres articulatoires."
T.I.P.A., 3, 151-207.

TESTON, B. (1980)
"La mesure des paramètres aérodynamiques du langage."
T.I.P.A., 7, 149-178.

TESTON, B. (1983)
"A system for the analysis of the aerodynamic parameters of speech: The Polyphonometer model III."
10° Int. Cong. Phon. Sc., Utrecht, Sec. 5, 457.

UTILISATION DE L'INFORMATION ACOUSTIQUE CHEZ LE SUJET NORMAL ET CHEZ LE SUJET IMPLANTE

C. Berger-Vachon, B. Djedou

Laboratoire G.B.M. - Université LYON I

ABSTRACT

In this paper ,the authors compare the acoustic performances of a deaf patient fitted with a cochlear prothesis (Chorimac),and thoses obtained by a subject with normal hearing when restricted spectral in--formation is used via the transmitter Chorimac.

The performances deal with the recognition of acoustical features embedded in artificial words (logatomes).They are statistically analysed.

The implanted patient is fitted with the french cochlear implant Chorimac .The acoustical information is given by a set of band-pass filters.

The spectral signal ported by the filters is sent to a loud-speaker;the filters are used ,one by one, one after the other.

The comparative results of these two experiments are analysed and discussed.

INTRODUCTION

Depuis les expériences de A.Djourno et C.Eyries en 1957 {1},on sait qu'il est possible de restituer les sensations auditives à des sujets sourds profond,qui doivent leur surdité à une déficience de l'organe de Corti alors que la capacité fonctionnelle résiduelle du nerf auditif n'est pas nulle.De nombreux travaux ont été faits sur ce thème ces dix dernières années et de nombreuses prothèses visant à réhabiliter ces patients ont été développées{2}.

Le principe de ces prothèses est simple;le signal acoustique est traité par un système électronique;il est ensuite transmis à un récepteur,implanté sous la peau du patient,qui délivre des signaux électrochim--iques qui sont ensuite véhiculés par le nerf auditif jusqu'au cerveau.Jusqu'à présent les principaux travaux ont surtout eu pour but d'évaluer les perfor--mances auditives des sujets implantés {3}.Par contre peu de travaux fondamentaux ont été éffectués pour tenter de mieux cerner la composante informative du signal délivré aux implantés.Ce type d'étude peut être éffectué de deux façons:

* par l'analyse directe de la forme d'onde qui est émise par les prothèses et de juger ainsi son contenu informatif{4},
* en essayant de reproduire chez des sujets ayant une audition normale des sensations se rapprochant de ce qui est proposé aux patients.

Cette dernière méthode semble particulièrement intéressante puisqu'elle permet de préciser certains des phénomènes qui sont observés chez le patient; on peut toutefois critiquer cette approche puisque le sujet bien entendant reçoit normalement le signal par voie aérienne donc l'oreille est impliquée dans le mécanisme de l'audition,ce qui n'est pas le cas pour le patient implanté.Néanmoins l'étude des signaux test par des sujets bien entendants est aussi un moyen pour juger la qualité informative du signal et elle permet de plus de séparer ce qui vient de l'oreille. C'est dans cette optique que se situe ce travail puisqu'on oppose les performances d'un sujet implanté à celles d'un sujet normalement entendant pour la reconnaissance d'un vocabulaire test adapté à ce type d'expérience.

MATERIEL ET METHODES

1-Matériel

La prothèse Chorimac{5} est composée d'une partie externe,qui est l'émetteur et d'une partie implantée qui représente le récepteur .La liaison entre ces deux blocs se fait par voie hertzienne ou par couplage électromagnétique .Seule la partie exter--ne sera utilisée durant l'expérience avec le sujet normal.L'émetteur commence par un banc de douze filtres (composante essentielle de notre étude) dont les bandes passantes sont les suivantes:

F1 : 0 - 240 Hz
F2 : 240 - 350 Hz
F3 : 350 - 450 Hz
F4 : 450 - 600 Hz
F5 : 600 - 800 Hz
F6 : 800 - 1250 Hz
F7 : 1250 - 1490 Hz
F8 : 1490 - 1850 Hz
F9 : 1850 - 2500 Hz
F10 : 2500 - 3500 Hz
F11 : 3500 - 4900 Hz
F12 : 4900 - 7500 Hz

Afin d'extraire le signal acoustique de l'émetteur du Chorimac et de le rendre audible à un sujet bien entendant,nous avons réalisé un amplificateur de puissance (2w).Cet amplificateur a été connecté séquentiellement à la sortie de chaque filtre(figure 1)

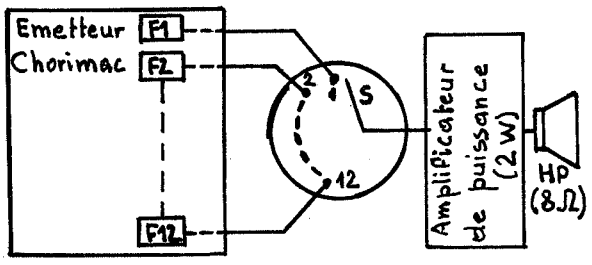


Fig-1-:Distribution de l'énergie des filtres sur un haut-parleur

2-Choix du vocabulaire et méthodes

Le vocabulaire qui a été utilisé est un ensemble de douze mots sans signification (logatomes),construits à partir d'oppositions phonétiques simples:

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

La liste de ces douze logatomes est indiquée ci-après :

1-TATA	4-DADA	7-SASSA	10-ZAZA
2-TOTO	5-DODO	8-SOSSO	11-ZOZO
3-TITI	6-DIDI	9-SISSI	12-ZIZI

Nous avons construit, de façon aléatoire, douze listes de 129 mots (une liste pour chaque canal) con--tenant chacune entre dix et douze occurences de chacun des mots du vocabulaire. Ces listes ont été présentées, sans lecture labiale, à un sujet implanté et à un sujet normal.

Le sujet sourd qui a participé à ce travail est un homme de 58 ans, implanté à l'hôpital Edouard Herriot de Lyon en janvier 1983.

Ce patient était atteint d'une cophose bilatérale tardive depuis dix ans environ au moment de l'impla--ntation. Ses performances ont été comparées avec celles d'un homme d'audition normale de 42 ans.

La lecture des listes au sujet bien entendant a été effectuée selon le protocole suivant :

Le sujet bien entendant était installé dans une chambre sourde complètement isolée phoniquement de la pièce où étaient effectuées les prononciations. Le signal acoustique issu de chaque filtre du Chorimac lui parvenait par l'intermédiaire d'un haut-parleur.

Pour chacune des listes on a construit une matri--ce carrée 12x12; les lignes représentent le stimulus (les mots prononcés par le locuteur à travers le microphone du Chorimac) et les colonnes, la réponse du sujet, la trace de cette matrice représente le nombre de mots reconnus par le sujet.

Ensuite, à partir de ce tableau {6}, nous avons déduit les taux de reconnaissance des :

* consonnes
*voyelles
*trait voisé // non voisé
*trait plosif / sifflant

Le but est de comparer les résultats fournis par les deux sujets (implanté et normal).

RESULTATS

Nous avons indiqué, pour chacune des pronon--ciations d'une liste, sur les tableaux 1 et 2 les taux de reconnaissance, en pourcent, des :

* voyelles,
* trait voisé / non voisé,
* trait plosif / sifflant.

Ensuite on a calculé le classement de chaque canal et le pourcentage moyen attaché à la recon--naissance d'une liste donnée par la formule suivante:

$$\bar{P} = \frac{1}{7} (3 \times P_V + 2 \times P_{V\bar{V}} + 2 \times P_{p/s})$$

$\bar{P}$: moyenne pondérée,

P_V : pourcentage de reconnaissance des voyelles,

$P_{V\bar{V}}$ et $P_{p/s}$ sont les pourcentages de détection, respectivement, du voisement et de l'aspect plosif/sifflant des consonnes.

On peut dire, approximativement, que lorsque deux pourcentages diffèrent de :

* 11,5 % pour les voyelles,
* 12 % pour les traits consonantiques,
* 7 % pour les moyennes pondérées,

la différence est significative.

canal	voyelle	trait V/$\bar{V}$	trait P/S	moy. pond.	classement
1	31.8	62.0	46.5	44.6	12^{e}
2	28.7	70.0	45.0	45.2	10^{e}
3	64.3	56.6	46.5	57.0	7^{e}
4	39.6	52.7	44.2	44.7	11^{e}
5	79.0	51.9	49.6	62.9	3^{e}ex.
6	72.9	53.7	53.4	62.9	3^{e}ex.
7	70.6	63.6	52.7	63.5	2^{e}
8	60.5	59.7	74.4	64.2	1er
9	49.6	58.9	79.8	60.9	5^{e}
10	50.4	55.8	77.5	59.7	6^{e}
11	28.7	49.6	79.8	49.3	9^{e}
12	24.0	59.7	78.3	49.7	8^{e}

Tableau-1-: Pourcentages de reconnaissance obtenus avec le sujet implanté.

canal	voyelle	trait v/$\bar{v}$	trait p/s	moy. pond.	classement
1	28.7	96.1	72.9	60.6	12^{e}
2	39.5	100	76.7	67.4	11^{e}
3	49.6	96.9	76.0	70.6	10^{e}
4	89.9	100	91.5	93.2	9^{e}
5	98.4	100	96.0	98.4	8^{e}
6	100	100	100	100	1erex.
7	100	100	100	100	1erex.
8	100	100	100	100	1erex.
9	100	100	100	100	1erex.
10	100	96.1	100	98.9	7^{e}
11	100	98.4	100	99.5	6^{e}
12	99.2	100	100	99.66	5^{e}

Tableau-2-: Pourcentages de reconnaissance obtenus avec le sujet ayant une audition normale.

DISCUSSION

Pour compléter cette étude comparative, on peut considérer les résultats obtenus par une méthode de reconnaissance automatique basée sur la métrique euclidienne (tableau-3-) et sur l'archétype du plus proche voisin:

canal	voyelle	trait v/$\bar{v}$	trait p/s	moy. pond.	classement
1	44.2	50.3	54.3	49.6	8^{e}
2	62.7	55.0	51.9	56.5	4^{e}
3	62.0	55.8	52.7	56.8	3^{e}
4	64.3	47.3	60.5	57.4	2^{e}
5	65.1	52.7	47.3	55.0	6^{e}ex.
6	61.3	61.3	57.4	60.0	1er
7	62.8	55.8	46.5	55.0	6^{e}ex.
8	58.9	54.3	55.0	56.06	5^{e}
9	45.7	46.5	51.2	47.8	10^{e}ex.
10	34.1	55.0	55.6	48.3	9^{e}
11	41.9	48.8	52.7	47.8	10^{e}ex.
12	35.6	50.4	49.6	45.2	12^{e}

Tableau-3-: Pourcentages de reconnaissance automatique par la métrique euclidienne.

* De l' étude du tableau 2, il ressort que l'information ne passe pas bien à travers les premiers canaux (filtres 1, 2 et 3). Ensuite pour les canaux 10 et 11, le sujet distingue moins bien le voisement.

* La comparaison des tableaux 1, 2 et 3 nous conduit à faire les remarques suivantes:

  -- Les canaux 5, 6, 7, 8, 9 correspondent à la bande fréquentielle 600-2500 Hz passent bien l'information pour le sujet implanté. Pour le sujet normal les meilleurs canaux sont 6, 7, 8, 9 et 12.

  -- Les plus mauvais taux de reconnaissance sont donnés par les filtres 1, 2, 4 pour le sujet implanté et 1, 2, 3 pour le sujet normal.

Globalement, on peut dire que les deux sujets reconnaissent mieux l'information acoustique sur la bande fréquentielle 600-2500 Hz.

* Les résultats donnés par la reconnaissance automatique ne correspondent pas tout à fait à ceux fournis par les deux sujets puisque les premiers canaux qui passent mieux l'information que les derniers (filtres 9, 10, 11, 12). Toutefois le canal 6 vient en tête du classement comme chez le sujet normal; ce canal n'est que troisième chez le sujet implanté, mais la différence n'est pas significative.

Sur le plan des performances absolues, les résultats sont nettement meilleurs chez le sujet à audition normale que pour la machine et le sujet implanté, les pourcentages de reconnaissance moyens étant 90.7 %, 55.4 % et 52.9 %, le seuil de significativité étant alors de 2 %.

CONCLUSION

L'étude de la distinction de logatomes par un patient implanté et par un sujet d'audition normale montre que l'information spectrale partielle est traitée approximativement de la même manière. Ces résultats correspondent en outre à ceux qui sont obtenus en réalisant une reconnaissance automatique du signal.

Par contre sur le plan quantitatif, le sujet bien entendant a obtenu des résultats très nettement

supérieurs à ceux rencontrés dans les deux autres situations, ce qui montre que le découpage spectral de l'information n'est pas un facteur déterminant dans le handicap auditif résiduel observé après implantation

Il est aussi remarquable de constater que les résultats obtenus par le patient et en reconnaissance automatique sont équivalents.

Des expériences complémentaires sont à prévoir pour préciser ces constations expérimentales.

BIBLIOGRAPHIE

{ 1 }- A.DJOURNO,C.EYRIES- "Prothèse auditive par excitation électrique à distance du nerf sensoriel à l'aide d'un bobinage inclus àdemeure ".
Presse Méd.,vol.35,1417-1423 (1957).

{ 2 }- Report of the Ad-Hoc Committee on Cochlear Implants -A.S.H.A 28-4, 29-52 (1986).

{ 3 }- C.CHOUARD- "Deuxième symposium inter--national sur l'implant cochléaire".
Paris 22-24 sept. 1983, Proc: Prof.Chouard, Hôp. Saint-Antoine, 75012 Paris (1983).

{ 4 }- C.BERGER-VACHON,J.GENIN,R.MOUHSSINE- "Etude du codage acoustique effectué par la prothèse cochléaire Chorimac".
Annales des télécommunications,42-314,119-131 (1987).

{ 5 }- M.FARDEAU,P.ORANGE- "L'appareillage" Cahiers d'O.R.L,14-6,609-616 (1979).

{ 6 }- R.MOUHSSINE,C.BERGER-VACHON,J.GENIN- "Reconnaissance de mots artificiels à travers la prothèse cochléaire Chorimac".
14e J.E.P ,13-16 ,Paris (1985).

IMPLANTS COCHLEAIRES ET PERCEPTION : PREMIERS RESULTATS

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

*LABORATOIRE DE LA COMMUNICATION PARLEE
46, AVENUE F. VIALET 38031 GRENOBLE CEDEX
** CHU PURPAN, SERVICE ORL, PLACE BAYLAC 31059 TOULOUSE CEDEX
*** UDSM, 3 RUE DE METZ, 31061 TOULOUSE CEDEX

ABSTRACT

Very recently, the domain of perception has been extended to the phonetic and linguistic investigation of totally deaf patients perception, implanted with intra- or extracochlear electrode(s).

This study concerns with some results about patients bearing a single channel extra-cochlear electrode. The aim of this paper is to present results picked up in the very difficult period when patients have yet to learn how to match informations coming from both visual and (lip reading) and auditory channels. The analysis occurring 3 months later the operation, bears on 620 phonems by patient under 5 evaluation sessions.

The results seem to indicate that central vowels and nasal ones get the worst performances, and the cardinal vowels, the best ones but corresponding to very clear lip movements. Among consonants, voiceless ones present the best results and among the voiced consonants, fricatives.

This analysis comes to light the great importance of training sessions in order that patients can found and then improve their new auditory acoustical perception.

INTRODUCTION

Dans cette présentation des résultats, nous nous proposons d'évaluer les performances perceptives phonétiques de 4 sujets cophotiques implantés à l'aide d'une seule électrode extracochléaire dans le service du Dr. FRAYSSE au CHU Purpan de TOULOUSE. Les sujets ont été implantés en 1984 et le premier test d'évaluation a été pratiqué à 3 mois (suivi d'un 2ème à 5 mois pour le sujet HC) en 1984 et 1985. Depuis s'est élaborée une nouvelle méthodologie incorporant une évaluation sur les plans sémantique et syntaxique. L'évaluation phonétique s'est trouvée de ce fait limitée, mais les principes de constitution des corpus de tests sont demeurés semblables. Il nous a semblé interessant, en attente des suivants, de présenter les premiers résultats portant sur une évaluation phonétique assez large.

La perception auditive du sujet s'appuie sur une perception visuelle, la lecture labiale étant autorisée. Selon l'aptitude du sujet, cette aide s'est montrée plus ou moins effective, et généralement plus efficace pour les pré-linguaux (sujets JC et HC) que pour les post-linguaux (MTD et AS).

Après une description du corpus, nous comparerons les limites de reconnaissance de la parole par la machine (reconnaissance automatique) et les limites théoriques de la discrimination du système phonétique par un sujet cophotique non implanté (n'ayant donc que la lecture labiale à sa disposition), afin de mieux apprécier les résultats des 4 sujets présentés ensuite.

CORPUS ET APPAREILLAGE

Nous avons présenté en liste ouverte à une ou deux reprises aux différents sujets (JC, MTD, AS, HC1 et HC2), 69 paires de mots lexicaux mono- ou disyllabiques, et 30 mots isolés, appartenant tous au français standard. Les deux séances pour le sujet HC ont été espacées de 2 mois. Le test porte environ sur 620 phonèmes par individu. Les mots ont été présentés avec ou sans déterminant.

Pour les deux premiers tests, nous avons défini théoriquement un corpus qui tienne compte du type de déficience et de la bande passante de l'appareillage : les occlusives ont été ainsi privilégiées ainsi que les voyelles ouvertes.

L'appareillage consiste en un stimulateur-transcodeur cochléaire MEDTRONIC SP3062, de type mono-électrode extra-cochléaire. Le stimulateur se caractérise par l'abscence de transmission des niveaux acoustiques inférieurs à 65 dB et sur une limitation de la bande passante de 200hz à 2KHz.

La répartition des 620 phonèmes s'effectue donc ainsi (norme A.P.I) :

- 60% de consonnes dont :
  - 35% d'occlusives (/p, t, k, b, d, g, m, n/)
  - 24% de constrictives
    - 7% de médianes (/f, s, ʃ, v, z, ʒ/)
    - 17% de liquides (/l, R/)
  - 1% de semi-voyelles (/j,w/)
- 40% de voyelles dont :
  - 29% de non-aigües (/a, ɔ, o, ɛ, œ, ə, u, ɑ̃, ɔ̃, ɛ̃, œ̃/)
  - 11% de voyelles aigües (/e,i,y/).

Par ailleurs, les paires de mots sont composées d'items semblables ou phonétiquement très voisins, et dans ce dernier cas, la différence ne repose que sur un seul phonème. Les deux phonèmes en opposition déterminent 3 types de paires, auxquelles s'ajoutent les paires d'homonymes :

- 17% de paires d'homonymes
- 25% de paires quelconques
- 28% de paires métathétiques
- 30% de paires minimales.

Que recouvrent ces dénominations? Les paires métathétiques, création originale, sont des paires de mots phonétiquement identiques mais présentant une inversion dans l'ordre des phonèmes de même catégorie (consonnes ou voyelles) :

ex : /p a n/ ---> /n a p/
1 -> 2 2 -> 1

L'intérêt de ces paires métathétiques est tout d'abord de présenter um moyen terme entre les homonymes et hétéronymes; ensuite, ces paires métathétiques apportent une aide à la précision de l'évaluation en choisissant sélectivement le nombre de traits opposant les phonèmes inversés. Au niveau du diagnostic médical, elles permettent également de localiser la perturbation : niveau périphérique, central ou les deux. Ce sont dans les tests d'évaluation des dyslexies qu'elles trouvent leur meilleur champ d'application.

Les paires minimales [1] sont constituées, comme on le sait, de 2 mots s'opposant par un seul trait en un seul phonème. Lorsque l'opposition entre les deux phonèmes compte plusieurs traits acoustiques, on utilise le terme de paires quelconques. Ainsi dans l'exemple /nɔt/ # /bɔt/, /n/ par rapport à /b/, comporte les traits d'opposition suivants si l'on se réfère au domaine de l'analyse perceptuelle selon la terminologie [1] : nasal, vocalique, continu et aigu.

L'analyse du corpus montre que le test présenté aux sujets cophotiques est difficile en raison 1° de l'occurrence immotivée des mots 2° de la grande proximité acoustico-phonétique des mots 3° de leur présentation en liste ouverte 4° de la longueur du test. Ce sont ces raisons entre autres qui nous ont poussée à définir une nouvelle méthodologie.

Tout aussi difficile est parfois l'interprétation des résultats par le phonéticien, dans la mesure où non seulement la prononciation des sujets pré-linguaux en particulier, mais l'orthographe aussi est parfois mal assurée : c'est le cas en particulier du sujet JC. L'évaluation des performances acoustico-phonétiques de ce sujet s'effectue en prenant comme référence le système du français, que ce soit à l'émission/réception du mot et inversement quand le sujet reproduit le mot : or le système acoustico-phonétique du français n'est pas celui du sujet cophotique prélingual. Pour certains sujets d'ailleurs, on peut se demander si la notion de système reste valide. L'entrainement à la lecture labiale dans tous les cas reste pour le sourd profond la seule voie d'intégration possible du système phonétique.

DISCRIMINATION PHONETIQUE PAR LECTURE LABIALE

LES SOSIES ARTICULATOIRES

Si nous nous plaçons dans les conditions idéales où le sujet cophotique présente une lecture labiale sans faute, les 36 oppositions du système phonétique du français se réduisent à environ 12. En effet, la lecture labiale n'autorise au mieux que la discrimination des modes et lieux articulatoires, en éliminant les traits de voisement ou de nasalité. Pour un sourd profond, la liste des sosies articulatoires pour nous limiter par exemple aux consonnes, est la suivante :

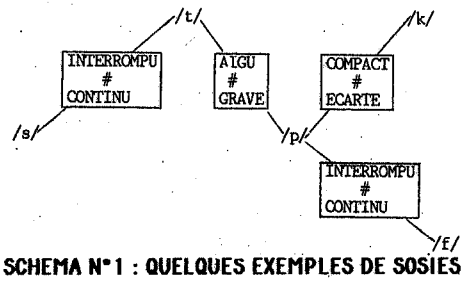
- /p = b = m/
- /t = d = n = ɲ/
- /s = z/
- /ʃ = ʒ /
- /f = v/
- /k = g/

Il faut mentionner que ces deux derniers phonèmes ainsi que /R/, étant des consonnes postérieures, ont toutes les chances de passer inaperçus. A cette liste de sosies, vient s'en ajouter une autre qui provient du domaine acoustique.

LES SOSIES ACOUSTIQUES

Ce nouveau groupe de sosies pour nous en tenir au seul plan des consonnes, concerne les sourds implantés. En effet lorsque l'implantation est réalisée, un certain nombre de traits acoustiques peuvent éventuellement être perçus : l'implantation a ainsi pour effet d'apporter au sujet cophotique une certaine discrimination supplémentaire. Cependant cette dernière ayant pour caractéristique d'être partielle a pour effet d'augmenter le nombre des sosies acoustiques.

Toutes choses égales par ailleurs, sur le plan de la perception acoustique, un seul trait acoustique oppose ces consonnes jumelles : il n'est pas étonnant que des confusions fréquentes se produisent, confusions que l'on rencontre d'ailleurs chez des individus à perception auditive normale. Le schéma ci-dessous présente quelques exemples de sosies acoustiques pour les consonnes :



SCHEMA N°1 : QUELQUES EXEMPLES DE SOSIES ACOUSTIQUES ET LEUR TRAIT D'OPPOSITION

En accord avec les données de la psychologie expérimentale, on peut s'attendre à ce que les performances des sujets dans un premier temps soient même minorées, dans la mesure où les individus ont à intégrer et harmoniser au niveau du décodage, des sources d'informations conflictuelles provenant des canaux visuel et auditif, de leur codage différent, et ce, dans un contexte psychologique délicat, celui de la rupture d'habitude basée jusque là sur le visuel. Les séances de rééducation vont être alors, on le conçoit, d'une importance capitale aux niveaux acoustico-phonétique et psychologique. La répétition des exercices, la perspicacité, la motivation du sujet, et ses ressources psychologiques seront de nature à lui faciliter l'apprentissage et à susciter en lui la perception de la différence aux niveaux acoustique, phonétique et prosodique.

RAPPEL DES PERFORMANCES EN RECONNAISSANCE AUTOMATIQUE DES MOTS ISOLES

Les scores de reconnaissance automatique diffèrent selon qu'il s'agit de parole continue ou de mots isolés, et selon les méthodes utilisées. En parole continue, les résultats les meilleurs avoisinent les 85%. En ce qui concerne les mots isolés, les méthodes globales affichent 95-96% en monolocuteur et 85% en multilocuteur. Les méthodes analytiques qui procèdent par analyse phonétique atteignent des scores de 80%. Il est bon de se souvenir de ces derniers scores en particulier lorsque l'on analyse les performances des sujets cophotiques chez qui la démarche semble se rapprocher prioritairement des méthodes analytiques.

LES RESULTATS

Pour les quelques cas analysés, on note qu'il n'existe pas d'opposition bien nette entre sujets pré- et postlinguaux. Cependant et contre notre attente, c'est un sujet prélingual (HC) qui présente les résultats les meilleurs et un postlingual (AS) qui fournit les moins bons.

PERCEPTION DIFFERENTIELLE DES MOTS

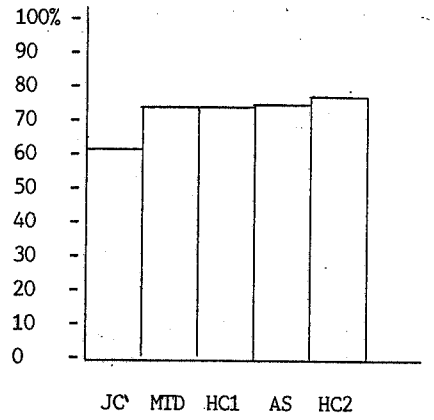


TABLEAU N°1 : POURCENTAGES DE DISCRIMINATION EXACTE DES MOTS SELON LES LOCUTEURS

Les paires d'homonymes sont généralement moins bien reconnues : la moyenne est de 48% pour les 5 tests. On constate par ailleurs une grande dispersion entre les performances, MTD et HC2 fournissant respectivement 8% et 92% de perception exacte des mots semblables. Sans évidemment nier qu'il existe une corrélation entre les deux faits, il est remarquable que chez HC2 une bonne discrimination phonétique ait pu se manifester jusque dans la perception de la similitude qui dans ce contexte de test (discrimination de la différence), peut passer pour un aveu d'échec. Tous les patients savaient qu'il existait des mots semblables, mais la plupart d'entre eux, ayant trop à coeur de prouver leur compétence, ont franchi sans doute inconsciemment les limites de la réponse objective.

RECONNAISSANCE GLOBALE DES MOTS

Les résultats précédents concernaient la perception de l'homophonie ou de l'hétérophonie des mots présentés en paires, indépendamment de l'exactitude phonétique. C'est elle qui nous intéresse désormais. Les pourcentages qui sont donnés ci-dessous (cf tableau 2) correspondent aux scores de reconnaissance exacte des mots dans leur intégralité, mots isolés ou en paire. On remarque les bons scores du sujet HC, malgré la difficulté du test :

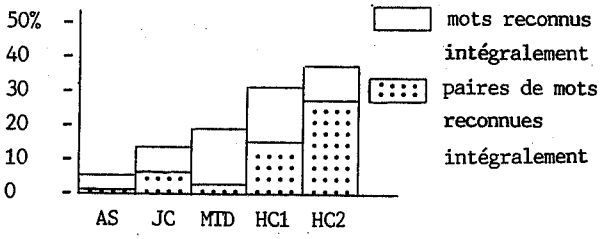


TABLEAU N°2 : POURCENTAGES DE RECONNAISSANCE INTEGRALE DES MOTS ET DES PAIRES SELON LES SUJETS

RECONNAISSANCE DES PHONEMES

Le test propose 620 phonèmes à identifier, mais le nombre réalisé par les différents sujets varie en fonction des erreurs, les élisions ou les épenthèses : les 5 tests comportent de 600 à 730 phonèmes. Les pourcentages bien entendu en tiennent compte.

Les scores présentés ci-dessous (Tableau 3) donnent le taux de reconnaissance moyen des phonèmes. Les performances sont très dépendantes du sujet, les scores pouvant varier parfois dans le rapport du simple au plus du double (respectivement sujets AS et HC2) :

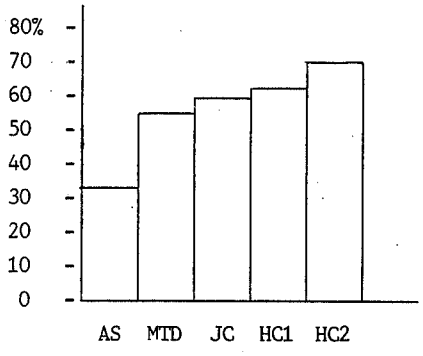


TABLEAU N°3 : POURCENTAGES DE RECONNAISSANCE DES CONSONNES ET DES VOYELLES SELON LES SUJETS

Si l'on ajoute à ces scores ceux des discriminations à un trait près (cf sosies articulatoires et acoustiques), on apporte une amélioration moyenne d'environ 20%.

De manière plus détaillée on constate que comparativement aux consonnes (tableau 4), la perception des voyelles pour l'ensemble des 5 tests (tableau 5) est meilleure :

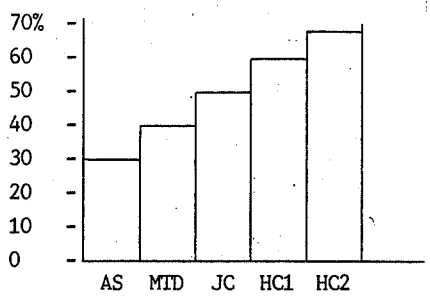


TABLEAU N°4 : CONSONNES

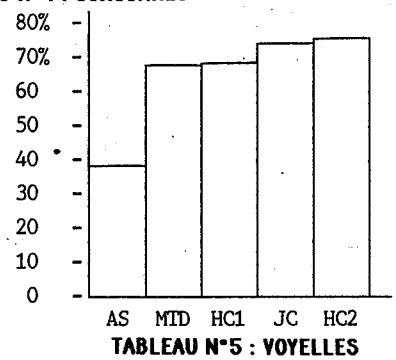
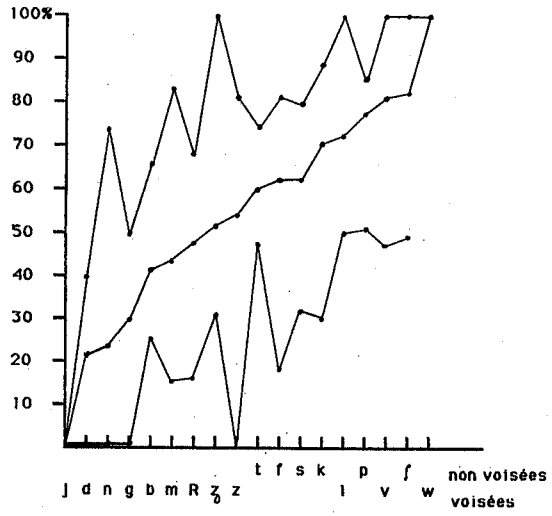


TABLEAU N°5 : VOYELLES

POURCENTAGES DE RECONNAISSANCE SELON LES SUJETS

Les graphiques 1 et 2 présentent les résultats à propos de la discrimination moyenne des phonèmes au travers de 3 courbes superposées. la plus basse propose les scores phonémiques les moins bons, relevés dans l'ensemble des 5 tests, alors que la courbe la plus haute offre les meilleurs, prélevés dans les mêmes conditions. La courbe du milieu représente la moyenne phonémique calculée sur les 5 tests.



GRAPHIQUE N°1 : POURCENTAGES DE DISCRIMINATIONMOYENNE DES CONSONNES ET SCORES MINIMAUX ET MAXIMAUX.

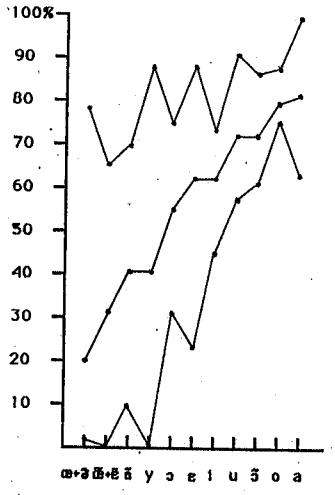
L'examen de ces résultats révèle que d'une manière générale les sourdes sont mieux perçues que les sonores, résultat attendu : comparer par exemple les scores de p/b, t/d, k/g, s/z, ʃ/ʒ, avec une exception f/v.

Les macro-classes et consonnes les moins bien reconnues sont :

- les occlusives sonores : /b, d, g/
- les occlusives nasales : /m, n/ sauf exception
- la liquide /R/
- la semi-consonne /j/.

Contrairement à notre attente, les constrictives sont mieux perçues en général que les occlusives (65% contre 54%). Cette constatation nous a menée dans les tests suivants, à rééquilibrer les corpus en faveur des constrictives en majorant leur nombre, ainsi qu'en faveur des constrictives médianes (/f, s, ʃ, v, z, ʒ/) par rapport aux latérales (/l, R/).

En ce qui concerne les voyelles, comme on le constate graphique n°2 ci-dessous, ce sont les voyelles ouvertes (/a, ɛ, ɔ/) qui sont les mieux reconnues (74% en moyenne), et à égalité ensuite les voyelles fermées (/e,ø,i,y,u), les voyelles fermées (/e, ø, i, y, u) -ou antérieures (/e, i/) et postérieures (/o, u/)-, soit 66%.



GRAPHIQUE N°2 : POURCENTAGES DE DISCRIMINATION MOYENNE DES VOYELLES ET SCORES MINIMAUX ET MAXIMAUX.

APPROXIMATION DU SEUIL MAXIMAL DE REHABILITATION AVEC LECTURE LABIALE

Les scores ayant été calculés, il est tentant de déterminer la discrimination réelle, c'est-à-dire la discrimination obtenue en ôtant du score, le pourcentage correspondant à la probabilité de répartition, ce que l'on pourrait appeler la part du hasard. Du même coup, ces résultats pondérés pourraient fournir une approximation du taux de réhabilitation du sujet par l'implant avec lecture labiale.

Nous avons donc calculé le nombre moyen de phonèmes articulatoirement ou acoustiquement semblables ou très proches (un seul trait acoustique en opposition) : il s'agit encore des sosies articulatoires ou acoustiques. Ce nombre moyen est pondéré par l'effectif réel du phonème dans le corpus. Ce nombre moyen est de 4.84 pour les consonnes et de 4.49 pour les voyelles : cela signifie qu'en moyenne un phonème quelconque, articulatoirement ou acoustiquement, pour un sujet cophotique implanté, peut être confondu, s'il est une consonne, avec 4.84 de ses pairs, et s'il est une voyelle, avec 4.49 d'entre elles. Convertis en pourcentage, ces taux sont égaux tous deux à 22%.

Ce taux de 22% est à ôter des scores de discrimination. Sur l'ensemble des phonèmes, les scores suivants, dans la mesure où la lecture labiale serait excellente, peuvent fournir une approximation du taux moyen (et à la fois seuil maximal) de réhabilitation phonétique du sujet cophotique par l'implant :

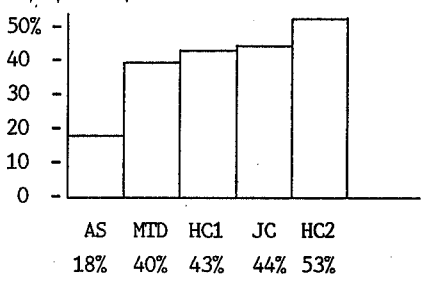


TABLEAU N°6 : SEUIL MAXIMAL MOYEN DE REHABILITATION THEORIQUE DES SUJETS COPHOTIQUES PAR L'IMPLANT ET LECTURE LABIALE

Ce taux de réhabilitation reste très dépendant du sujet; on constate par ailleurs une amélioration sensible du score de HC1 à HC2, respectivement 43% et 53%.

REFERENCES BIBLIOGRAPHIQUES

[1] PECKELS J.P. et ROSSI M.
Le test de diagnostic par paires minimales, REVUE D'ACOUSTIQUE n°27, 1973, 245-262.

[2] FRAYSSE B., URGELL H., CAELEN-HAUMONT G.
Implant mono-cochléaire : technique chirurgicale, Actes de la 1ère Société ORL, ALGER, 1984.

[3] FRAYSSE B., SOULIER M-J., URGELL H., CAELEN-HAUMONT G.
Extra-cochlear implantation : technique and results, AMERICAN JOURNAL OF OTOLOGY (à paraître).

RECONNAISSANCE
DIALOGUE

ADAPTATION AUX LOCUTEURS EN RECONNAISSANCE AUTOMATIQUE DE LA PAROLE PAR ANALYSE DES CORRELATIONS CANONIQUES ET QUANTIFICATION VECTORIELLE

K.CHOUKRI*,** , S.FRAENKEL** , G.CHOLLET**

* Laboratoires de Marcoussis, CRCGE, Route de NOZAY, 91460 Marcoussis.
** ENST-SIGNAL, CNRS UA 820, 46 rue Barrault, 75013 Paris, France.

1. Abstract

In order to obtain high accuracy in speaker independent word recognition, a speaker adaptation approach, based on Canonical Correlation Analysis (CCA), has been introduced. An organization of speaker space is needed to simplify the adaptation vocabulary selection. A dictionary of quantized spectra offers an elegant solution to that problem as it will be shown in this paper. Vector Quantization reduces both time and space requirements. Therefore our isolated words recognizer uses this coding technique. A comparative evaluation of some speaker adaptation algorithms is conducted. It is shown that adaptation, performed using CCA and VQ, permits to improve speaker independence in an isolated word recognizer.

This paper also describes briefly a local canonical correlation analysis (LCCA), an under completion technique, which is introduced to take into account the speech diversity (voiced/unvoiced frames, ...).

A future work will be focused on the use of CCA with stochastic models.

2. Introduction

L'adaptation aux locuteurs d'un système de reconnaissance automatique de la parole (SRAP) semble être une approche prometteuse en vue de la construction de systèmes indépendants du locuteur.

L'adaptation par analyse des corrélations canoniques, tout en donnant de bons résultats [1], a soulevé de nombreux problèmes. Il s'agit notamment du problème du choix du vocabulaire qui *"reflète"* les caractéristiques du nouveau locuteur, ce qu'on désigne par *"vocabulaire d'adaptation"*. Dans ce papier, nous essayerons de concrétiser les suggestions faites par les auteurs dans des publications précédentes. Le concept d'adaptation doit viser au moins deux objectifs. Tout d'abord la méthode d'adaptation doit être un module indépendant susceptible d'être intégrer dans n'importe quel SRAP. Ensuite, elle doit s'imposer une contrainte ergonomique, en réduisant le vocabulaire d'adaptation à un strict minimum, ou bien, et c'est à notre sens plus réaliste, en opérant de manière dynamique.

Cet aspect dynamique est réalisable grâce à une prise en compte de nouvelles occurences, au fur et à mesure de leur arrivée. Cette démarche permet d'affiner la représentation de l'espace paramétrique du nouvel utilisateur. Cette amélioration de la représentation de l'espace paramétrique du nouvel utilisateur sera, vraissemblablement, d'autant plus significative que les occurences en question ne seront pas bien reconnues. C'est à ce dernier mode de fonctionnement que va notre préférence, puisque l'adaptation est vue comme une procédure de convergence vers des SRAP monolocuteurs.

L'adaptation par analyse des corrélations canoniques se prête bien à un fonctionnement en mode dynamique. Cependant le problème du choix des mots d'adaptation demeure, lorsqu'on se contente d'une adaptation statique. On doit trouver le minimum de mots, qui reflètent le mieux la variabilité inter-locuteur. La quantification vectorielle, par le biais de réduction du nombre de spectres à stocker, apporte une solution très élégante à ce problème. Il est lié, pour notre application, à la couverture de l'espace spectral des locuteurs.

3. Recherche de mots d'adaptation et QV

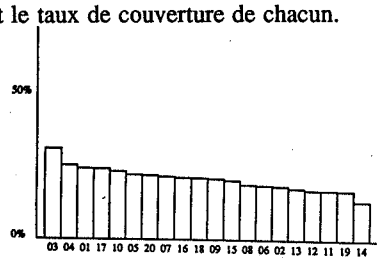
La quantification vectorielle est une technique de codage, qui est de plus en plus utilisée en RAP, car elle permet une mémorisation des distances inter-spectrales. Le principe de base consiste à quantifier les spectres réels dont on dispose par comparaison à un dictionnaire. Dans notre expérience le quantificateur est défini par la méthode statistique des " *Boules optimisées*" [2]. A partir d'un corpus de spectres réels, elle détermine un nombre de classes dont on garde un représentant. Malheureusement, cet algorithme ne gère pas le nombre de classes mais la distorsion maximale admise, ce qui définie le rayon de nos classes. Le dictionnaire de spectres quantifiés serait alors formé par l'ensemble des représentants de chaque classe. Les expériences décrites dans cet article concernent un système de reconnaissance de 20 mots isolés [3]. Pour la génération du dictionnaire, le corpus sera formé par une élocution des 20 mots.

Chaque mot peut être codé sur le dictionnaire de vecteurs ainsi obtenu, et dès lors il sera représenté par une suite d'étiquettes de la façon suivante (m est la longueur du mot et le $C_i(k)$ est l'étiquette du spectre quantifié qui code la k ième fenêtre d'analyse du mot i): $Mot_i = [\ C_i(1),....,C_i(m)]$.

Le choix des mots serait alors dirigé par une recherche d'un nombre minimal de mots qui engendre le dictionnaire de QV, sans se soucier du lien entre acoustique et phonétique sinon en supposant que la quantification, au pire supprime du bruit acoustique et peu d'informations pertinentes. La définition de la couverture de l'espace par un mot i est assez intuitive. Elle consistera

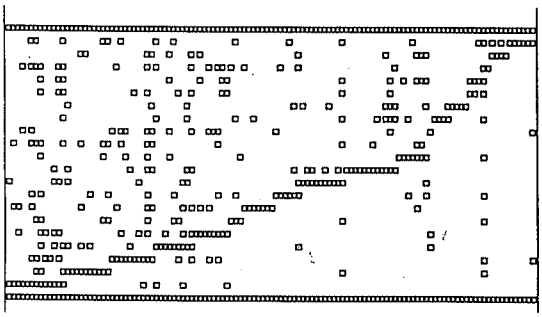
à trouver l'intersection entre l'ensemble des étiquettes qui constitue le dictionnaire et les m étiquettes qui codent le mot i en question.

Cette couverture est dépendante du locuteur. Pour chacun des mots on peut calculer un taux de couverture du dictionnaire. C'est le rapport entre le nombre d'étiquettes différentes, présentes dans le codage du mot et le nombre de vecteurs du dictionnaire. La figure ci-dessous en donne une illustration. On a classé, par ordre de couverture du dictionnaire, les 20 mots du vocabulaire en donnant le taux de couverture de chacun.



Taux de couverture du dictionaire par chacun des 20 mots

Si on veut considérer une chaîne formée de plusieurs mots, on doit aussi calculer un taux de recouvrement entre les divers mots, afin de choisir un minimum de mots qui assure une couverture acceptable du dictionaire. Un exemple graphique est donné ci-dessous. On a porté une représentation du dictionnaire entier en bas et en haut du schéma. Pour chacun des 20 mots, on montre les différentes étiquettes qui figurent dans son codage.



Taux de recouvrement entre les 20 mots

A partir de ce graphique on essaye de chercher cette chaîne d'adaptation optimale. Pour cela on détermine d'abord le mot qui couvre le mieux notre dictionnaire. Ensuite on considère le sous-dictionnaire qui contient seulement les vecteurs qui n'ont pas été touchés par le mot sélectionné à l'étape précédente. On réitère le même procédé avec le sous-dictionnaire et les mots restants. Cette méthode permet d'obtenir un classement des différents mots suivant leur couverture du dictionnaire de QV. La dépendance de cette démarche au locuteur peut être évitée de deux manières. La première en réalisant l'expérience décrite ci-dessus pour plusieurs locuteurs et en déduisant des taux de couverture moyens pour chacun des mots. La seconde consisterait à engendrer un dictionnaire de QV grâce à un corpus enregistré par beaucoup de locuteurs (une sorte de dictionnaire multilocuteur robuste) et de calculer un taux moyen pour chaque mot de tous les locuteurs. Notons que cette procédure est faite en dehors de l'exploitation du SRAP.

Pour notre part, nous privilégions la première méthode car elle se justifie pleinement dans notre approche par corrélations canoniques, dont le fondement est de tenir compte de la variabilité inter-locuteurs, ce qui n'est pas le cas dans la deuxième méthode.

4. Adaptation par analyse des corrélations canoniques

Disposant d'une suite de mots, adéquate pour l'adaptation, nous allons voir s'ils donnent de bons résultats. Une première expérience est menée avec comme paramètres 10 LAR par fenêtre d'analyse, sans QV. On reprend un schéma de SRAP classique, déjà décrit dans [1]. Grâce à une première expérience, on a obtenu des taux de reconnaissance avant adaptation (en monolocuteur et en croisant les références [1]). Ensuite nous avons procédé à une première évaluation en adoptant comme référence une élocution particulière d'un locuteur et en adaptant par corrélations canoniques les élocutions d'un autre locuteur, avec 7 mots d'adaptation. Dans la mesure où on a traité deux locuteurs de la base Texas, ces résultats sont à prendre avec beaucoup de précaution.

Les taux de reconnaisance sont donnés ci-dessous.

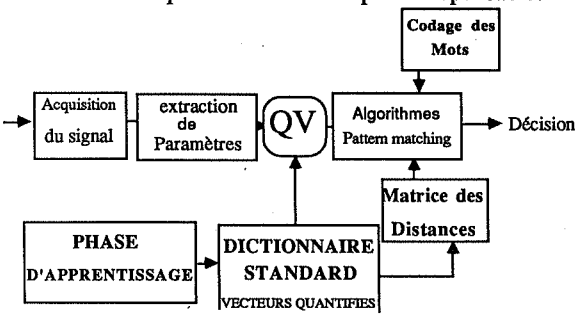
score de reconnaissance		
	première position	deux premières positions
non adapté (monoloc.)	87	93
adapté (monoloc.)	91.2	95.
non adapté (croisé)	28.5	44.5
adapté (croisé)	60.6	67.8

Cela montre fort bien que la variabilité intra-locuteur est assez importante (la base de données a, en effet, été enregistrée sur une longue durée), même si la paramétrisation et le nombre de cœfficients utilisés ne s'avèrent pas optimaux. Quant à la variabilité inter-locuteur on s'y attendait bien.

5. QV et Reconnaissance

Comme nous venons d'utiliser la QV pour simplifier le choix du vocabulaire d'adaptation, et comme elle peut permettre un gain de temps appréciable, nous allons implémenter un SRAP à base de QV. Nous chercherons à montrer que notre approche de l'adaptation par analyse des corrélations canoniques peut être inséré dans ce type de SRAP.

La figure suivante montre un exemple de système de reconnaissance basé sur la Quantification Vectorielle (QV). Il est suffisamment classique pour que le besoin d'en décrire chaque module ne soit pas indispensable.

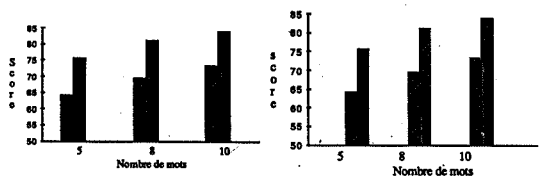


Remarquonscependant que la partie qui le différencie d'une approche classique est dictée par le dictionnaire de quantification, notamment le quantificateur qui devra discrétiser tous les spectres réels des mots tests, et la table de codage des mots. Cette remarque aboutit naturellement à des idées d'adaptation: et si l'on substituait un dictionnaire par un autre, mieux adapté au nouveau locuteur, en bénificiant du codage des mots qui ne sera pas à refaire. C'est l'approche qu'ont longuement étudié Shikano [4,5], Goatcher [6] et Bonneau et al. [7].

6. Evaluation après adaptation par Shikano

Cette approche, dont nous avons décrit le principe dans les dernieres JEP, considère que si la phase de codage des mots est faite grâce à un dictionnaire d'un locuteur standard, il n'est pas nécessaire de le refaire pour un autre locuteur dont on vient d'acquérir le dictionnaire. En effet en cherchant des transformations entre les deux dictionnaires, on peut substituer chaque vecteur du dictionnaire original par un autre du dictionnaire du nouveau locuteur. Dans ce cas, il suffit de remettre à jour la matrice entre spectres quantifiés pour accomplir cette phase d'adaptation. Shikano (mais c'est aussi l'approche de Goatcher et de Bonneau), génère un dictionnaire du second locuteur grâce au vocabulaire d'adaptation. Ensuite, en associant les mots d'adaptation des deux locuteurs par des algorithmes de DTW, il obtient une série d'histogrammes représentant la table de substitution entre les deux dictionnaires. Dans un premier algorithme, il ne garde que le vecteur qui est associé le plus souvent à son vecteur d'origine. Dans un second il considère un spectre fictif, qui est une combinaison linéaire de tous les spectres qui sont associés à un spectre donné. Les pondérations découlent de son histogramme.

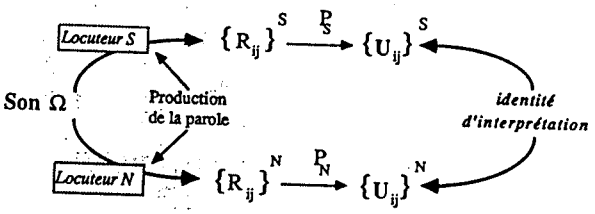
Une première évaluation de cette approche a été entreprise avec la base de donnée décrite dans [3], mais cette fois-ci avec 12 LAR pour améliorer le taux de reconnaissance au départ. Les taux de reconnaissance sont donnés ci-dessous avec 5, 8 et 10 mots d'adaptation.



Taux de reconnaissance par la méthode de Shikano (1er candidat et 2 premiers candidats)

7. Approche par corrélations canoniques

Le principe et l'évaluation de l'analyse des corrélations canoniques ont été longuement décrits dans [1]. A partir d'un ensemble de caractéristiques de deux locuteurs, on cherche à définir un couple de projecteurs vers un espace commun où la variabilité inter-locuteurs serait alors réduite. Un schéma est redonné ci-dessous:



Principe des corrélations canoniques

Cette fois, au lieu de disposer d'une suite de mots réels du locuteur standard, on considère que nous disposons d'un dictionnaire de QV et d'un codage des mots du vocabulaire. Au lieu d'appliquer nos transformations aux mots réels, on va les appliquer directement aux vecteurs du dictionnaire du locuteur standard. En effet, si on dispose des mots $M_1,\ldots,M_n$ du locuteur standard, ceux du nouveau locuteur sont obtenus dans le nouvel espace par l'équation (l'égalité étant admise au sens d'un critère d'erreur):

$$P_S(M_k) = P_N(M_k)$$

Dans l'approche par QV on dispose du dictionnairee du locuteur standard. Si on se replace dans le cadre de l'analyse des corrélations canoniques, on pourra lui appliquer le projecteur P_S au même titre que les mots, puisque ces deux approches sont identiques. En effet, que l'on projette chacune des composantes d'un mot ou que l'on projette les spectres du dictionnaire, la différence est que dans le premier cas on refait la même projection plusieurs fois, les composantes du mot étant les éléments du dictionnaire. Il suffit de projeter le dictionnaire standard pour avoir une équation similaire à celles définissant nos projecteurs, à savoir:

$$P_S(\text{Dictionaire standard}) = P_N(\text{nouveau Dictionnaire})$$

Pour l'acquisition des projecteurs, on utilise les mots d'adaptation avant toute quantification. La méthode de calcul est explicitée dans [1].

Dans l'espace de projection des deux locuteurs, on va disposer d'un dictionnaire adapté au nouveau locuteur. La reconnnaissance se fera de la même manière qu'avant l'adaptation, on a juste changé le dictionnaire. Il faut dire aussi que l'on a changé l'espace de paramétrisation, il faut donc y projeter les mots du nouveau locuteur avant de les coder, le codage des mots de référence n'ayant pas été modifié.

Le nouveau schéma de reconnaissance est alors:

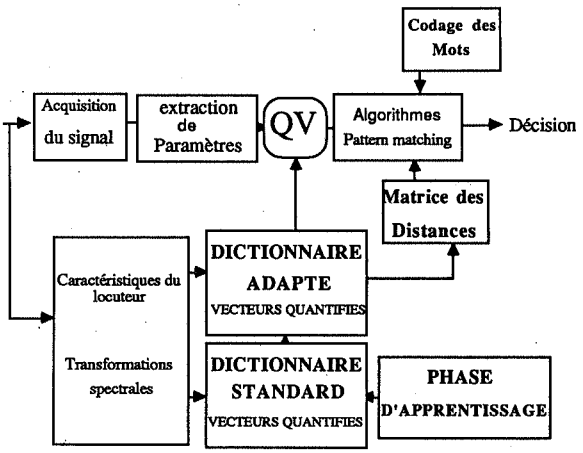


schéma du SRAP adaptatif

Une première expérience a eu lieu, avec une représentation par 10 LAR et a permis de dégager les premiers résultats ci-dessous. Cependant les tests ont porté sur deux locuteurs. Les LAR ont été choisis en raison d'autres expériences en cours.

Taux de reconnaissance		
	première position	deux premières positions
non adapté (monoloc.)	81.5	89.3
adapté (monoloc.)	89.1	96.
non adapté (croisé)	32.8	49.1
adapté (croisé)	56.	63.

Taux de reconnaissance après une adapatation par QV et ACC.

Si on ne se place pas dans l'optique "adaptation dynamique", on peut envisager de construire un dictionnaire, un peu sommaire, grâce au vocabulaire d'adaptation prononcé par le nouveau locuteur. Ceci doit être fait avec beaucoup de précautions afin que le dictionnaire reste, un tant soit peu, robuste. L'approche Boules Optimisées, permet de respecter une structure algébrique de l'espace en gérant le nombre de classes créées (qui n'est pas forcément proportionnel au nombre de mots). Une première phase consistera à coder les mots du nouveau locuteur par son propre dictionnaire. La détermination des projecteurs se fera par l'alignement de mots codés. Cette approche est actuellement testée. On pourra avoir une idée de la dépendance entre la QV et la coarticulation.

8. Conclusion

On vient de montrer quelques expériences d'adaptation, à base de QV et ACC. Les tests préliminaires ont porté sur des paramètres LAR et ont montrés des améliorations minimes par rapport à ceux annocées avec des paramètres cepstraux [1]. Des tests, en cours, portent sur des paramètres MFCC afin des respecter la structure algébrique euclidienne, une des hypothèses émises lors de l'application de l'analyse des Corrélations Canoniques. Cela montrerait aussi que la distance euclidienne n'est pas la mieux adaptée aux LAR.

Aucune conclusion comparative ne peut être tirée étant donné le nombre de tests. Il faut admettre cependant que chacune des deux méthodes présentent un certain nombre d'inconvénients. L'approche de Shikano réalise une transformation point par point et nécessite donc un corpus de départ assez conséquent pour le nouveau locuteur. Cela s'oppose à une approche dynamique et ne respecte pas les contraintes ergonomique que doit s'imposer tout SRAP adaptatif.

Les corrélations canoniques procèdent par le calcul d'une seule transformation pour tout le dictionnaire. Cela ne semble pas très raisonnable vue la diversité des sons (comme par exemple voisé/non voisé). Un compromis entre les deux approches est actuellement à l'étude. Il s'agit d'une analyse des corrélations canoniques locales qui tiendrait compte d'une répartition du dictionnaire standard en sous dictionnaires plus homogènes [8].

Remerciements:

On tient à remercier le professeur **Shikano** pour les documents qu'il nous a communiqué et aussi pour les discussions fructueuses que nous avons eues avec lui. On tient aussi à remercier **Ph. Lookwood** des Ldm, pour les nombreuses suggestions dont il nous a fait part et **P. Fouques**, élève de Mastère à l'ENST pour son expérimentation des corrélations canoniques locales.

9. Références

[1] **Choukri, K. , Chollet, G. (1986),** Unsupervised and Dynamic learning of Spectral transformations using canonical correlation analysis for adaptation of ASR to new speakers. **Computer and Speech Language** (Vol 1, N 2). Academic Press Inc. pp. 95-107.

[2] **G. Flamenbaum, J. Thiery, J.P. Benzecri & C. Mullon (1979)),** Agrégations en Boules de rayon fixe et de centres optimisés. **Les cahiers de l'analyse des données** (Vol 4, N3)

[3] **Montacié, C. , Chollet, G. (1987),** Systèmes de référence pour l'évaluation d'applications et la caractérisation de bases de données en Reconnaissance Automatique de la parole. **16 JEP,** Hammamet , Tunisie.

[4] **Shikano, K., Lee, K.F. & Reddy, R. (1986),** Speaker adaptation through vector quantization. **IEEE-ICASSP,** Tokyo, pp. 2643-2646.

[5] **Shikano, K., (1986),** Speaker adaptation through vector quantization. Communication personnelle. (Draft for Carnegie-Mellon University Tech. Report).

[6] **Goatcher, J.K. & Mason, J.S. (1986),** An adaptive approch to a speaker-independent isolated word system with short training. **International conference on speech input/output techniques and applications,** London, pp.67-70.

[7] **Bonneau, H. & Gauvain, J.L (1987),** Vector quantization for Speaker adaptation. **IEEE-ICASSP,** Dallas.

[8] **Choukri, K. , (1987),** Quelques approches à l'adaptation aux locuteurs dans les systèmes de reconnaissance automatique de la parole. Thèse de L'ENST (en préparation).

RECONNAISSANCE RAPIDE DE MOTS ISOLES PAR QUANTIFICATION VECTORIELLE MULTI-SECTIONS

A. GOURINDA et J.P HATON

CRIN/INRIA B.P 239 54506 VANDOEUVRE-CEDEX
FRANCE

ABSTRACT

In this paper we present a fast word recognition algorithm based on Multisection Vecor Quantization. A separate multisection codebook is designed for each word in the vocabulary by dividing the word into equal-length sections and by designing a codebook for each section. Unknown words are also divided into equal-length sections, each section is averaged and encoded with the Multisection codebooks. For speaker-dependent recognition of the french digits plus the words "oui" and "non" this approach achieved a recognition accuracy greater than 99 percent with only one distortion computation per input section for each vocabulary word. We give a generalization of this algorithm to continuous speech recognition.

I-INTRODUCTION

Parmi les méthodes de compression de données dans un but de stockage ou de transmission , la quantification vectorielle (QV), est celle qui offre actuellement le plus de possibilités. En transmission de la parole, il est possible de transmettre chaque composante d'un vecteur quantifié avec une fraction de bit. Pour la reconnaissance de la parole, on peut utiliser soit un système fondé entièrement sur la quantification vectorielle, soit un système fondé sur une autre technique de reconnaissance, et utilisant la quantification vectorielle comme un premier filtre. Cette méthode a connu également beaucoup de succès aussi bien en codage qu'en reconnaissance de l'image. Le succès de la QV a amené d'autres techniques de reconnaissance à incorporer celle-ci dans leurs approches.

II-METHODE CLASSIQUE

Soit à reconnaître un vocabulaire constitué de V mots. Pour chaque mot du vocabulaire, on construit , par quantification vectorielle, un ensemble de prototypes ("codebook"), obtenu a partir d'un certain nombre de répétitions du même mot. Pour reconnaître un mot inconnu, on calcule sa distorsion moyenne par rapport à tous les ensembles de prototypes, et on identifie ce mot au mot du vocabulaire dont l'ensemble de prototypes correspondant réalise la plus petite distorsion.

Lors de l'apprentissage, on peut soit imposer un ensemble de prototypes à distorsion égale pour chaque mot du vocabulaire, ce qui donne des ensembles de prototypes n'ayant pas le même nombre d'éléments, soit imposer un ensemble de prototypes de taille fixe. Pratiquement, cette dernière solution semble donner de meilleurs résultats.

Cette methode ne nécessite aucun alignement temporel (contrairement aux téchniques fondées sur la programmation dynamique), ni aucune normalisation des mots et donne de bons résultats surtout en reconnaissance mono-locuteur, mais demande beaucoup de calculs, du fait que chaque fenêtre d'analyse ou trame du mot inconnu est comparée à tous les prototypes de chaque mot du vocabulaire. Il est donc nécessaire d'établir un ordre de comparaison tel que nous le proposons dans la méthode exposée ci-après.

III-METHODE UTILISANT UN ORDRE DE COMPARAISON

a-construction de l'ensemble de prototypes pour le mot K du vocabulaire.

Cette construction s'appuie sur la méthode proposée par Burton et al [2], dont le principe est rappelé brièvement:

Soient r_1, r_2,- - - - - - -,r_T, des répétitions du même mot K du vocabulaire. On découpe toutes ces répétitions en sections de longueur fixe, n, qu'on appelle facteur de compression. Chaque section représente donc n fenêtres d'analyse.

En classifiant, avec l'algorithme présenté dans [3,4], toutes les premières sections de chaque répétition, on obtient le premier ensemble de prototypes C_{K1}, et ainsi de suite pour les autres sections. On obtient finalement l'ensemble de prototypes C_K du mot K :

$C_K=\{ C_{K1},C_{K2},- - - - - - - -,C_{KSk} \}$
$C_{kj}=\{ C_{kj1},C_{kj2},- - - - - - - \}$,

S_k etant le nombre de sections du mot K.

En fait, pour la génération de l'ensemble de prototypes, il y a au moins deux approches possibles, soit un alignement à gauche, soit une normalisation linéaire à une même longueur de toutes les répétitions du même mot.

b-reconnaissance d'une forme inconnue.

Comme lors de l'apprentissage, le mot inconnu est découpé en sections. On compare ensuite chaque section à la section correspondante de chaque mot du vocabulaire. Plus précisément:

Soit à reconnaître la forme X décrite par une suite de vecteurs X_i:

$X=X_1X_2X_3$ - - - - - - - $-X_L$

Les X_i correspondent à une certaine paramétrisation des trames de parole(par exemple FFT,MFCC ,LPC,etc.).

Pour simplifier les notations, considérons la comparaison de X avec le mot K du vocabulaire, dont l'ensemble de prototypes associé est C_K; n

est le facteur de compression.

La distorsion correspondante au codage de la trame X_m par la j-ème section C_{kj} du mot K du vocabulaire est:

$$d_{mj} = \min_i d(X_m, C_{kji})$$

d étant la distance inter-trames.

La distorsion pour coder la j-ème section de la forme inconnue par C_{kj} est alors:

$$d_{S_j} = \sum_{m=(j-1)n+1}^{\min[jn,L]} d_{mj}$$

La distorsion moyenne pour coder la forme inconnue par l'ensemble de prototypes C_K est finalement:

$$D_K = \frac{1}{L} \sum_{j=1}^{S_k} d_{S_j}$$

La forme inconnue X sera affectée à la classe du mot r du vocabulaire, si:

$$D_r = \min_k D_k \quad \text{et} \quad D_r < S$$

L'introduction du seuil S permet d'éviter que le système ne prenne des décisions dans des situations ambigues.

Burton et al [2] ont prouvé que, pour l'apprentissage dépendant du locuteur, un seul prototype par section suffit. Ceci correspond, lors de la reconnaissance, à un seul calcul de distance par trame. Nous proposons dans notre algorithme d'effectuer un seul calcul de distance par section, c'est à dire que, pour chaque trame, on ne calcule que ses coefficients d'autocorrélation et on effectue un moyennage pour chaque section.

L'algorithme suivant correspond à une paramétrisation du signal par LPC et une normalisation linéaire de tous les mots à une longueur fixe L. Chaque mot est donc représenté par une suite de L trames.

Algorithme d'apprentissage

Choix du facteur de compression(n)

{Il y aura donc m(L/n) sections. Les trames de 1 à n seront comparées à C_{k1}, les trames n+1 à 2n à C_{k2} et ainsi de suite jusqu'à la fin du mot}

Détection des frontières du mot

Normalisation linéaire du mot

D=0

Pour chaque section i du mot k du vocabulaire representée par C_{ki}

Moyenne = zero

Pour chaque trame j de la section i du mot inconnu

Préaccentuation+fenêtre de Hamming

Calcul des coefficients d'autocorrélation (CA)

Moyenne = moyenne + CA

Fin pour

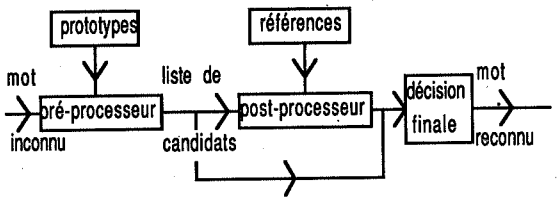
Calcul des coefficients LPC correspondants à moyenne

D = D +d(moyenne,C_{ki})

Fin pour

IV-COORDINATION DE DEUX METHODES DE RECONNAISSANCE

La première méthode est celle que nous venons d'expliquer. Dans cette méthode le mot inconnu est comparé à une suite de prototypes reflètant les caractéristiques spectrales du mot. On introduit ainsi implicitement une erreur de quantification. Pour cette raison la deuxième méthode travaille directement sur des références non quantifiées. Cette méthode permet de raffiner la décision de la première méthode, mais demande le stockage de plusieurs références pour chaque mot du vocabulaire. Par la suite, on appellera pré-processeur la première méthode et post-processeur la deuxième comme l'indique la figure 1.



Architecture du système de reconnaissance

(Le pré-processeur devient inutile quand le pré-processeur transmet un seul candidat)

figure 1

1)-Choix du post-processeur

Le post-processeur peut être soit:

a) Le pré-processeur lui-même, après remplacement des prototypes par de vraies références.

b) Un modèle de Markov (HMM)

c) Un modèle de programmation dynamique(DTW)

2)-Règles gérant le comportement des deux processeurs

a) Pré-processeur

Soit η le mot du vocabulaire réalisant la plus petite distortion:

$$\eta = \min_{i} (D_i)$$

et β le mot réalisant la seconde plus petite:

$$\beta = \min_{i \neq \eta} (D_i)$$

Deux cas peuvent se présenter:

Règle1(un seul candidat η)

$D_\eta < s_1$

$D_\beta - D_\eta > s_2$

Régle2(plusieurs candidats)

La liste contiendra tous les candidats j tel que:

$D_j - D_\eta < s_3$

b) Post-processeur

Le post-processeur, doit reagir en fonction de la règle appliquée par le pré-processeur

Règle1 ----------> Rien

Règle2 ----------> Règle des K-plus proches voisins

c) Choix des seuils s1, s2 et s3

On choisit ces seuils, de manière à ce que quand le pré-processeur ne peut pas transmettre un seul candidat, il en élimine au moins 80%.

3)-Calcul des distorsions par le post-processeur

Pour effectuer ce calcul, il y a trois possibilités, soit:

a)- quantification de la forme de référence et de la forme à reconnaitre

b)- quantification uniquement de la forme de référence par :
-son propre ensemble de prototypes ,
-tous les ensembles de prototypes
(exemple : on peut exprimer la forme de référence "cinq", soit comme une suite de prototypes appartenants tous à l'ensemble de prototypes cinq, soit

appartenant à tous les ensembles de prototypes.)

c)- pas de quantification

Il est évident que la solution qui introduit le moins d'erreurs de quantifications et tient compte des connaissances du pré-processeur est la solution b). Si, par exemple, on choisit la programmation dynamique comme deuxième méthode alors tous les calculs de distances sont déjà effectués par le pré-processeur.

V-RESULTATS DE LA PREMIERE PARTIE

La paramétrisation choisie du signal est la méthode de prédiction linéaire(LPC). La fréquence d'échantillonnage du signal vocal est de 12 Khz. La distance utilisée aussi bien pour l'apprentissage que pour la reconnaissance est celle d'Itakura-Saito à gain normalisé connue sous le nom de d_{gn} [1]. La fenêtre d'analyse est de 21,2 ms sans chevauchement. A chaque trame du signal on applique une préaccentuation de 0,94 puis une fenêtre de Hamming. La base de données utilisée est composée des chiffres zero jusqu'à neuf plus oui et non, prononcés par deux locuteurs, un homme et une femme, chaque locuteur ayant prononcé 30 fois le même mot. Toutes les locutions ont été normalisées linéairement à 24 et le facteur de compression est 4. Les 10 premières répétitions de chaque mot ont servi pour l'apprentissage et les 20 autres pour le test. La détection de début et fin de mot aussi bien pour l'apprentissage que pour la reconnaissance se fait automatiquement. Il est à noter que le succès de la méthode dépend entièrement du résultat du détecteur, la plupart des erreurs commises par le système provenant de mauvaises détections des frontières de mots.

Nous avons obtenu un taux de reconnaissance de 100% pour le locuteur masculin, et de 98,3% pour le locuteur féminin, ce qui donne un taux moyen de 99,1%.

VI-RECONNAISSANCE DE LA PAROLE CONTINUE

Pour pouvoir généraliser la méthode à la reconnaissance de la parole continue, nous utilisons le phonème comme unité de reconnaissance. Nous disposons d'une base de parole continue, toutes les phrases étant étiquetées manuellement en phonèmes. On parcourt toutes les phrases, et pour chaque phonème on regroupe dans un fichier toutes ses occurrences . Après quantification vectorielle de chaque fichier, nous obtenons un certain nombre de prototypes pour chaque phonème(8 prototypes).

Pour l'étiquetage automatique d'une phrase inconnue en phonèmes, on prend une fenêtre de longueur fixe comme unité de reconnaissance. Une phrase est donc découpée en une suite de segments de taille fixe qui recoivent les étiquettes des phonèmes les plus proches.

Pour chaque fenêtre les 4 phonèmes candidats qui réalisent les 4 plus petites distorsions sont conservés. On départage ces candidats en introduisant une information contextuelle. On considère que la fenêtre courante plus les deux fenêtres qui la précèdent, ainsi que que les deux qui la suivent comme formant un phonème. On compare ce phonème formé de ces cinq fenêtres aux quatres phonèmes candidats. Finalement la fenêtre d'analyse sera reconnue comme étant le phonème candidat qui réalise la plus petite distorsion.

Comme pour la première partie, la paramétrisation utilisée est la méthode de prédiction linéaire (LPC), la distorsion est la distance d'Itakura Saito à gain normalisé. La frequence d'échantillonnage est de 16 Khz. En reconnaissance mono-locuteur, on arrive à reconnaitre pratiquement toutes les zones stables . Nous tentons actuellement d'étendre la méthode à la parole multi-locuteur.

VII-CONCLUSION

Pour la reconnaissance de mots isolés, la méthode que nous proposons dans cet article, représente un gain considérable en temps de calcul. Si par exemple, le facteur de compression est 4, ce qui est le cas de notre étude, au lieu de résoudre quatre fois un système d'équations linéaires à p inconnues, on ne le résoud qu'une seule fois, et , de plus, on ne fait qu'un seul calcul de distance, au lieu de quatre, pour chaque section. En ce qui concerne la reconnaissance multi-locuteur, nous n'avons pas encore de résultats. Si la méthode s'avère concluante, elle permettra encore plus de reduction de calcul que pour la reconnaissance mono-locuteur, du fait que l'ensemble de prototypes de chaque section doit avoir plus d'un seul élément.

Pour la reconnaissance de la parole continue, la méthode proposée ne demande pas beaucoup de calculs(8 comparaisons pour chaque phonème, plus les comparaisons pour tenir compte du contexte droit et gauche), et donne de bons resultats. Ce qui serait intéressant dans un deuxième temps est d'utiliser par exemple la programmation dynamique pour départager les candidats selectionnés par la quantification vectorielle.

REFERENCES

[1] J.E.Shore and D.K.Burton,"Discrete utterance speech recognition without time alignment",IEEE Trans.Inf.Theory,vol.IT-29,pp473-491,jul 1983.

[2] D.K.Burton,J.E.Shore and J.T.Buck,"Isolated-word recognition using multisection vector quantization codebook",IEEE Trans.Acoust.Speech,Signal Processing,vol.ASSP-33,pp837-849,aug 1985.

[3] A.Buzo,A.H.Gray,R.M.Gray and J.D.Markel,"Speech coding based upon vector quantization",IEEE Trans.Acoust.Speech,Signal Processing,vol.ASSP-28,pp562-574,oct 1980.

[4] Y.Linde,A.Buzo and R.M.Gray,"An algorithm for vector quantizer design",IEEE Trans.Commun,vol.COM-28,pp84-95,jan 1980.

[5] B.-H.Juang,D.Y.Wong and Gray,Jr.,"Distortion performance of vector quantization for LPC voice coding",IEEE Trans.Acoust.Speech,Signal Processing,vol.ASSP-30,pp294-303,apr 1982.

PRISE EN COMPTE DES VARIATIONS PHONETIQUES EN RECONNAISSANCE DE LA PAROLE

Eric LAPORTE

Laboratoire d'Automatique Documentaire et Linguistique
Université Paris 7 (tour centrale, 9° étage)
2, place Jussieu
75221 PARIS CEDEX 05

ABSTRACT

This paper deals with ways for taking into account phonetic variations in speech recognition systems. Several recognition methods are considered. Particular emphasis is placed on recognition systems, based on pattern-matching, in which the decision unit is the fraction of speech between two adjacent syllabic centers. The phonetic data involved in this method include a list of references, which should contain variants. Such a method underlines the intrinsic interest in describing variants precisely and systematically. As an example of such a description, some phonetic alternations related to hiatuses in French are studied in detail.

INTRODUCTION

Cet article présente les façons de prendre en compte les variations phonétiques en reconnaissance de la parole, en envisageant plusieurs techniques de reconnaissance. Une place plus importante est faite à la reconnaissance globale avec comme unité la portion de parole comprise entre deux centres de syllabes consécutifs. Les différentes données phonétiques nécessaires dans le cadre de cette technique sont évoquées afin de mettre en évidence l'intérêt d'une description fine et systématique des variantes. A titre d'exemple, quelques variations phonétiques liées aux hiatus en français sont étudiées en détail.

I. VARIATIONS PHONETIQUES

Un système de reconnaissance de la parole doit prendre en compte la variabilité des données qu'il analyse. Les variations de prononciation constituent un problème majeur en reconnaissance automatique de la parole. Nous qualifierons de variations phonétiques celles qui affectent les représentations de la prononciation par des symboles phonétiques. Nous les distinguerons des variations acoustico-phonétiques. Les problèmes posés par ces dernières, et en particulier par la coarticulation entre segments contigus, sont reconnus depuis longtemps et des solutions spécifiques ont été proposées pour les résoudre. Les variations phonétiques sont également importantes : en voici quelques exemples. Le verbe *lier* peut se prononcer en une syllabe ou en deux syllabes. Cette variation affecte d'autres mots en [i] ainsi que des mots en [u], comme *louer*, et en [y], comme *suer*. D'autres variations ont une étendue lexicale encore plus importante. Dans des mots comme *illégal*, on prononce une consonne simple ou une consonne double. Dans *arrêt*, *série* ou *illégal*, on trouve un [e] fermé ou un [ɛ] ouvert, et tous les intermédiaires entre ces deux extrêmes sont acceptables. Dans *colis*, on entend un [o] fermé, un [ɔ] ouvert ou un intermédiaire ; dans *mener*, un [ø] fermé, un [œ] ouvert ou un intermédiaire, ou encore le *e* muet ne se prononce pas du tout. L'élément *-endre* admet plusieurs prononciations naturelles dans l'expression *descendre du bus*. On peut les représenter par [ãdrø] (où la voyelle [ø] admet des variations), [ãd] et [ãn]. Elles correspondent aux prononciations [desãdrødybys], [desãddybys] et [desãndybys]. Comme dans les exemples précédents, il s'agit d'une variation libre : les variantes alternent plus ou moins librement dans cette phrase[1]. Mais ici, la variation dépend du contexte : ainsi, les variantes [ãdrø] et [ãn] ne sont pas employées dans le contexte voisin *descendre ici*, puisque cette expression ne peut se prononcer ni [desãndrøisi] ni [desãnisi]. Ces variations, libres ou contextuelles, sont parfois difficiles à séparer des variations acoustico-phonétiques. Elles sont plus faciles à observer à l'oreille. Leur étude systématique est possible, et nous allons voir l'utilité d'une telle étude pour la reconnaissance de la parole.

II. CAS DE LA RECONNAISSANCE GLOBALE

Lors de la constitution d'un système de reconnaissance globale, des références sont soit construites à partir d'échantillons prononcés par le locuteur, soit synthétisées au moyen de connaissances acoustiques générales. Le signal de parole à reconnaître est mis en correspondance avec ces références, cette opération mettant généralement en jeu au moins une fois un alignement temporel dynamique. Dans ce cadre, le problème est qu'en cas de variabilité, le locuteur peut utiliser des variantes distinctes lors de l'apprentissage et pour la reconnaissance, par exemple [illegal] et [ilegal]. Si c'est le cas, il y aura des différences phonétiques entre les références et le signal de parole à reconnaître. Deux approches ont été proposées pour résoudre ce problème, et de nombreux compromis sont d'ailleurs possibles entre elles :

a) La première consiste à prévoir des références distinctes pour des variantes distinctes. Ceci est généralement fait au hasard des diverses prononciations que donne le locuteur lors de l'apprentissage, par exemple en lui faisant prononcer chaque énoncé un certain nombre de fois, pour faire apparaître les variabilités libres, ou en lui faisant prononcer des mots en contexte, pour faire apparaître les variabilités contextuelles. Ainsi, Watanabe [1] utilise comme échantillons pour l'apprentissage des énoncés de plusieurs syllabes, prononcés de manière naturelle, car cela fait apparaître des variations phonétiques (assourdissements, nasalisations, neutralisations, chutes, abrègements de voyelles), alors que dans une expérience antérieure [2], des échantillons d'une ou deux syllabes prononcés avec un souci de clarté ne faisaient pas apparaître ces variantes. Une limite de cette technique est que, même avec des prononciations répétées d'un même énoncé, il est difficile d'inciter le locuteur à utiliser lors de l'apprentissage toutes les variantes phonétiques libres et contextuelles qu'il est susceptible d'employer par la suite. S'il ne les utilise pas toutes, la description des variantes n'est pas exhaustive. En fait, elle ne pourrait s'approcher de l'exhaustivité que si on disposait de données systématiques sur les variations phonétiques, et si on concevait l'apprentissage en tenant compte de ces données. Il en est de même, *a fortiori*, si les références sont synthétisées au lieu d'être construites à partir d'échantillons.

b) La deuxième solution consiste à ne pas prévoir de variantes au niveau des références, et à contrôler les variations phonétiques, comme les variations acoustico-phonétiques, à l'aide de l'algorithme de reconnaissance acoustique. Ce composant du système permet en effet de mettre en correspondance une référence et une portion de parole même si elles présentent une ressemblance seulement approximative. Il fonctionne généralement par alignement temporel dynamique. Or les différences qu'on peut neutraliser par alignement temporel dynamique sont surtout les variations acoustico-phonétiques fines et les distorsions temporelles, alors que de nombreuses variations phonétiques ne consistent pas seulement en une distorsion temporelle : citons les variations entre [e] et [ɛ] des deux voyelles du verbe *serrer*, ou les prononciations de *descendre* évoquées en I. L'algorithme de reconnaissance devrait donc être spécifiquement tolérant à ce type de variations. Ici encore, rien ne peut être entrepris sans données sur les variations phonétiques, et en particulier sur leur extension lexicale.

III. CAS DE LA RECONNAISSANCE ANALYTIQUE

Dans un système de reconnaissance analytique, une analyse acoustique du signal de parole permet de détecter des segments phonétiques, puis les séquences ainsi obtenues sont comparées aux entrées d'un dictionnaire phonétique. Cette dernière opération est une mise en correspondance temporelle et le

problème des variantes phonétiques se retrouve à ce niveau : en cas de variabilité, toutes les variantes rencontrées doivent pouvoir être reconnues. Ici encore, deux moyens sont envisageables pour cela, ainsi que des compromis entre ces deux moyens :

a) Décrire les variantes exhaustivement et prévoir des entrées de dictionnaire distinctes pour des variantes distinctes. Ainsi, dans l'exemple précédent, les variantes en [ādrø], [ād] et [ān] sont connues du système, et subissent une reconnaissance acoustique indépendamment les unes des autres. Lorsque l'une d'elles a été reconnue, il faut ensuite passer de cette forme à l'élément lexical, qui, dans ce cas, est probablement unique et a reçu une forme de base représentant toutes les variantes phonétiques. Le dictionnaire doit mettre chaque variante en correspondance avec une forme de base. Cette méthode suppose donc l'existence d'un système de données phonétiques élaboré, avec notamment plusieurs niveaux de représentation : formes de base et variantes éventuelles.

b) L'autre solution consiste à concevoir un algorithme de mise en correspondance temporelle qui comporterait une description implicite des variantes (Vivès, [3]) et qui associerait directement les séquences phonétiques détectées à des formes de base représentant toutes les variantes. Un tel algorithme, pour être spécifiquement tolérant aux variations phonétiques qui peuvent se produire, inclurait des données systématiques sur ces variations.

En conclusion, aussi bien dans le cadre de l'approche analytique que dans le cadre de l'approche globale, la prise en compte des variations phonétiques suppose une description de ces variations, et en particulier de leur extension lexicale. Plus généralement, et en dehors de tout choix technique, il nous semble incontournable qu'on doive disposer de ces données dès lors qu'on désire reconnaître un vocabulaire étendu (Cohen & Mercer, [4]).

Il est plus difficile de savoir quel degré de précision est souhaitable dans l'établissement de ces données. Etant donné la précision limitée avec laquelle on sait reconnaître automatiquement des prononciations, on peut se demander à quelle finesse doit s'arrêter la description phonétique, et en particulier la description des variantes. Si on considère qu'il est inutile de traiter une variation lorsqu'on ne sait pas distinguer automatiquement les variantes [3], la prise en compte des phénomènes phonologiques se réduit. Toutefois, les algorithmes de reconnaissance acoustique s'affinent sans cesse, et la précision dans la reconnaissance phonétique semble être un objectif d'une importance capitale pour reconnaître la parole continue [1]. C'est pourquoi nous n'hésitons pas à anticiper sur les performances prévisibles de la reconnaissance acoustique en proposant des descriptions relativement détaillées. Notons toutefois que la finesse de la description présentée ici est limitée par le fait qu'aucun instrument n'a été utilisé et qu'elle repose sur des observations à l'oreille et à l'intuition.

IV. RECONNAISSANCE GLOBALE AVEC COMME UNITE LA PORTION DE PAROLE COMPRISE ENTRE DEUX CENTRES DE SYLLABES CONSECUTIFS

Dans cette section, nous envisageons la prise en compte des variations phonétiques dans le cadre d'une technique particulière : la reconnaissance globale avec alignement temporel dynamique, l'unité de reconnaissance étant la portion de parole comprise entre deux centres de syllabes consécutifs [2]. Nous rappelons tout d'abord les grandes lignes de cette approche.

IV. 1) Choix d'une unité de reconnaissance acoustique

Dans les systèmes de reconnaissance globale existants, l'unité de reconnaissance acoustique (URA) ou unité de décision est soit le mot, soit une unité de type syllabique telle que la syllabe, soit une unité de type phonémique telle que le diphonème (Decker et al., [5]). De nombreux auteurs, comme Shoup [6], se sont penchés sur le problème du choix entre ces solutions. Le fait que les variations contextuelles de part et d'autre d'un centre de syllabe soient limitées suggère de prendre comme URA la portion de parole comprise entre deux centres de syllabes[2] consécutifs, ce qui donne, pour le discours

Guy est ici. Sa voiture est en panne.

prononcé sans pause[3], le découpage suivant :

[#gi - ie - eti - isi - isa - avwa - aty - yre - etã - ãpa - an#]

La seule contrainte phonétique à laquelle obéit une suite d'unités de ce type est la contrainte "domino" : la deuxième voyelle d'une unité correspond à la première voyelle de l'unité qui la suit.

IV. 2) Le dictionnaire des références

La raison évoquée pour choisir une URA amène à choisir la même unité pour constituer les références servant à la reconnaissance acoustique. De plus, si ces références correspondent à des unités plus courtes que le mot, par exemple à des unités de type syllabique, elles forment un dictionnaire supplémentaire dont la taille n'est pas proportionnelle à la taille du vocabulaire reconnu : elle atteint un maximum de quelques milliers d'entrées et n'évolue plus guère lorsque le vocabulaire reconnu dépasse un certain seuil. Une liste pratiquement exhaustive des références distinctes peut donc être obtenue à partir de dictionnaires phonétiques couvrant l'ensemble de la langue courante. En d'autres termes, un examen extensif du lexique permet d'élaborer des données moins dépendantes du vocabulaire. Or, si la liste des références est rendue indépendante du vocabulaire, le dictionnaire de références ne dépend plus que de la langue et du locuteur. Cela facilite sa constitution et sa maintenance lors de l'évolution du vocabulaire ; de plus, cela accroît sa portabilité. C'est pourquoi notre étude des variations phonétiques se base sur les dictionnaires du LADL, qui comportent 51 000 formes de base[4], soit 350 000 formes fléchies[5], et correspondent à l'ensemble de la langue courante.

IV. 3) Identification des mots prononcés

Après la reconnaissance acoustique, on dispose d'une suite ou d'un treillis d'URA du type considéré. Ce treillis peut être transformé en un treillis de segments phonétiques qui lui est équivalent. Il reste alors à identifier les mots prononcés, en mettant le treillis en correspondance avec les mots d'un dictionnaire phonétique. Même pour une séquence phonétique donnée, plusieurs solutions sont souvent possibles : par exemple, à la suite [øropeɛ̃] peut correspondre aussi bien *européen* que la séquence *eux rot paix un*. Le choix entre solutions concurrentes met en jeu des informations non phonétiques telles que la structure syntaxique du discours.

IV. 4) Calcul de la liste des références distinctes

Le système comprend autant de références distinctes qu'il existe d'unités du type choisi susceptibles d'apparaître dans les discours à reconnaître. Nous avons vu en IV. 2) que la liste de ces unités peut être établie pour l'ensemble de la langue courante à partir d'un dictionnaire phonétique approprié, et qu'elle est alors pratiquement indépendante du vocabulaire connu par le système. Le calcul de cette liste met en jeu le découpage des mots du dictionnaire. Par exemple, le verbe *participer* fournit les unités [arti], [isi] et [ipe]. Les débuts de mots et les fins de mots, comme [pa] et [e] pour *participer*, requièrent un traitement spécial : *a priori*, toute fin de mot peut être combinée avec un début de mot pour former une unité, comme [ipʂwa] dans *type choisi*. Des combinaisons plus complexes peuvent faire intervenir des mots non syllabiques tels que *te, que, de, se, ce, je, le, me, ne, y* : ainsi [ɛski] figure dans plusieurs des prononciations possibles de *Qu'est-ce qu'il y a ?* Ces opérations fournissent une liste d'unités parmi lesquelles il reste à regrouper les unités identiques.

La principale difficulté consiste à établir le dictionnaire phonétique. Tout d'abord, on doit tenir compte des flexions : conjugaisons, pluriels, féminins. Comme ce sont des mots fléchis qui sont employés dans les textes, la liste des unités doit être calculée à partir d'un dictionnaire de formes fléchies. Par ailleurs, en cas de variations phonétiques, les différentes variantes qui apparaissent dans le discours se répercutent sur la suite des unités à reconnaître. Par exemple, la phrase *Guy repart* sera reconnue par l'intermédiaire de l'unité [irpa] si le *e* muet tombe, mais s'il est prononcé ce seront les unités [irø] et [øpa], ou [irœ] et [œpa]. Or, certaines unités n'apparaissent que par le jeu de variations phonétiques : c'est le cas de [ʃŋdy], observable dans la séquence *longue durée*. Pour établir la liste des unités, on doit donc tenir compte de toutes les formes, y compris les variantes. Cependant, s'il est possible de prévoir des références supplémentaires pour les variantes phonétiques des mots, les variations systématiques telles que celles évoquées en I. ont une extension lexicale trop importante pour qu'on puisse les représenter à l'aide d'entrées supplémentaires dans un dictionnaire phonétique. Un codage plus adéquat consiste à employer des phonèmes supplémentaires et des transformations de chaînes de phonèmes, ou transductions. Avec ce formalisme, la forme variable reçoit une représentation unique dans un dictionnaire phonémique, mais les différentes variables sont calculables à partir de cette représentation. Le calcul de la liste des unités à partir du dictionnaire phonémique comprend donc la production automatique des variantes phonétiques, ce qui demande des connaissances précises sur les variations de ce type.

Cette recherche d'une plus grande précision devrait permettre de cibler mieux la recherche de références, et devrait contribuer

à résoudre le problème du compromis entre le bruit (références reconnues à tort) et le silence (références non reconnues alors qu'elles devraient l'être) dans les performances du système. En effet, à défaut de données précises sur les variations phonétiques, on doit choisir un algorithme de reconnaissance acoustique plus tolérant aux dissemblances. Le problème revient alors à comparer des paroles bruitées ou déformées à des références imprécises à l'aide d'un algorithme approximatif. S'il est inévitable que le signal de parole à reconnaître puisse être bruité ou déformé, les deux autres parties du système (les références et l'algorithme de reconnaissance acoustique) peuvent être rendues plus précises, d'une part en affinant la description phonétique, d'autre part en utilisant des algorithmes de reconnaissance acoustique moins approximatifs, qui sont d'ailleurs plus efficaces.

V. UN EXEMPLE : LES HIATUS

Nombreux sont les auteurs qui ont étudié les variations phonétiques, qu'elles soient isolées ou systématiques (Pérennou & de Calmès, [7] ; Rossi & Lambert-Drache, [8] ; Néel et al., [9]). Leur extension lexicale est un aspect important de la question : il s'agit d'un paramètre théorique et de données d'un intérêt pratique certain. Nous avons étudié de ce point de vue plusieurs types de variations phonétiques liées à la présence de hiatus phonétiques ou orthographiques. Une telle étude, pour atteindre le niveau de finesse requis pour la reconnaissance automatique de la parole, nécessiterait peut-être d'utiliser des instruments et d'observer des spectrogrammes : c'est l'approche de Cohen & Mercer [4] et de Vaissière [10]. L'observation des faits à l'oreille est d'une précision limitée : par exemple, une distinction purement phonétique entre [ø], [ə] et [œ] ne peut être vérifiée à l'oreille d'une manière reproductible. Cependant, cette observation approximative permet déjà de préciser l'extension lexicale des phénomènes.

V. 1) Définition

On appelle hiatus un contact entre deux voyelles. Nous considèrerons aussi bien des hiatus phonétiques que des hiatus orthographiques, mais en raison des contraintes de place, il ne s'agira ici que de hiatus intérieurs. Les voyelles [i], [u], [y] suivies d'une autre voyelle dans le même mot, peuvent soit rester syllabiques soit perdre leur caractère syllabique en devenant respectivement [j], [w], [ɥ]. Ainsi, *cambouis* se prononce obligatoirement avec [w] et s'oppose à *brouette* qui se prononce obligatoirement avec [u]. Ces exemples illustrent le fait que la prononciation de [i], [u] et [y] en hiatus peut varier selon que le hiatus est précédé d'un groupe constitué d'une consonne obstruante *(Obs)* et d'une consonne liquide *(Liq)*, comme dans *brouette* et *plier*, ou non, comme dans *cambouis*.

V. 2) Alternance entre [yi] et [ɥi] après un groupe *Obs-Liq*

Dans les mots admettant la prononciation [ɥi] après un groupe *Obs-Liq*, cette prononciation monosyllabique est généralement obligatoire :

bruyère *[yi] [ɥi]
pluie *[yi] [ɥi]
truite *[yi] [ɥi]

Toutefois, dans une douzaine de mots, la prononciation en une syllabe et la prononciation en deux syllabes sont toutes les deux acceptables :

altruisme truisme [tryi] [trɥi]
fluide -ifier -ifiant -ification -ique -iser -ité superfluide [flyi] [flɥi]
incongruité [gryi] [grɥi]
superfluité [flyi] [flɥi]

Cette situation semble liée, pour certains mots, à des considérations morphologiques : ainsi, dans *incongruité*, le hiatus correspond à la limite entre le radical et un suffixe dérivationnel. Cette corrélation est cependant moins claire dans le cas de *fluide*, car le caractère suffixal de l'élément terminal *-ide* n'est pas établi.

V. 3) Alternance entre [ij] et [j] devant voyelle

Dans la plupart des mots, [j] devant voyelle n'admet aucune variation :

pied *[ie] *[ije] [je]

mais dans certains mots, les deux prononciations [j] et [ij] sont en concurrence :

lier ?*[ie] [ije] [je]

Lorsque le *i* reste syllabique, le hiatus disparaît généralement par la présence d'un [j] entre les voyelles en hiatus. Toutefois, ce [j] ne semble pas aussi nettement obligatoire qu'après un groupe *Obs-Liq* ou *Cons-*[ɥ] : comparer *lier* et *plier*.

Les mots concernés sont, mis à part le mot d'origine étrangère *ria*, une dizaine de verbes et certains de leurs dérivés :

fiancer fier skier lier surlier nier expier rire sourire scier obvier

Les formes conjuguées de ces verbes ont eux aussi les deux prononciations du hiatus. Quant à leurs dérivés, ils se répartissent en deux types. Certains ont les deux prononciations :

lier lieur lieuse liant liage lié liure liaison [ij] [j]

Les autres, bien que morphologiquement liés au verbe, se prononcent obligatoirement avec [j] :

fiable fiabilité méfier méfiance méfiant défier défiant défiance confier confiant confiance *[ij] [j]
indéniable indéniablement *[ij] [j]
sciure *[ij] [j]
lien reliage relieur reliure *[ij] [j]

On constate que cette répartition ne suit aucune règle. En conclusion, la prononciation du hiatus est liée à la question de savoir si l'élément final est un suffixe dérivationnel, mais cette corrélation n'est pas parfaite.

V. 4) Alternance entre [u] et [w] devant voyelle

Dans les mots comportant la voyelle [u] devant une autre voyelle, la prononciation [w] est toujours possible. La réciproque n'est pas vérifiée :

cambouis *[u] [w]
évanouir [u] [w]

Ni la conjugaison ni la dérivation ne modifient l'acceptabilité de la prononciation [u] : par exemple, elle reste acceptable dans toutes les formes conjuguées du verbe *louer*, ainsi que dans *loueur, louable, louablement, louage, louange, louangeur, relouer, surlouer, sous-louer, ...*

Il n'existe aucun exemple de mot comportant un [u] entre deux voyelles, même d'origine étrangère :

cacahuète [awɛ] *[auɛ]

Devant [a], de nombreux mots ont un [w] obligatoire, comme *loi* ; mais on trouve des formes où [u] est acceptable dans la conjugaison des verbes en *-ouer*, comme *jouer*, dans les dérivés en *-age* et *-able* de ces verbes, ainsi que dans le nom *rouage* et ses dérivés.

Devant [e] et [ɛ], la plupart des mots peuvent se prononcer avec [u] :

bouée [u] [w]

On trouve un [w] obligatoire dans une dizaine de mots :

ouais boëesse boette boetter débоetter serfouette douelle douellière couenne couenneux couette marouette

ainsi que dans quelques mots d'origine étrangère tels que *oued*.

Devant [i], certains mots peuvent se prononcer avec [u] :

ouïe ouïr hindouisme hindouiste louis épanouir épanoui épanouissement évanouir évanoui évanouissement inouï rouir rouissage rouissoir jouir jouissant jouissance réjouir réjoui réjouissance réjouissant

En revanche, on ne trouve que [w] dans :

oui ouïghour ouistiti bouif embabouiner ribouis cambouis fouine fouiner fouinard fouineur fouir fouisseur enfouir enfouissement enfouisseur serfouir serfouissage gouine baragouineur baragouinage baragouiner couic couinement couiner malouin mouise shampouineur shampouineuse shampouiner méchoui marsouiner vouivre

parmi lesquels beaucoup sont d'origine étrangère ou dialectale.

Devant [ɛ̃], la prononciation [w] est obligatoire :

babouin [wɛ̃] *[uɛ̃]
soin [wɛ̃] *[uɛ̃]

Il en est de même devant [o] ; les seuls exemples que nous en avons rencontrés sont *linguaux* et *statu quo*.

Devant [ø], [œ], [y] et [ɔ̃], la prononciation [w] n'est pas obligatoire :

boueux [u] [w]
joueur [u] [w]
nouure [uy] [wy]
jouons [u] [w]

Devant [ã], cette prononciation n'est généralement pas obligatoire non plus. Il s'agit essentiellement des formes en *-ant* des verbes en *-ouer*. Il en est de même pour *louange* et ses dérivés. Toutefois, la prononciation [w] semble obligatoire dans les mots *quantum*, d'origine étrangère, et *chouan*, d'origine dialectale.

V. 5) Alternance entre [y] et [ ] devant voyelle

Dans les mots comportant la voyelle [u] devant une autre voyelle, la prononciation [w] est toujours possible. La réciproque n'est pas vérifiée :

lingual *[y] [ɥ]
ruade [y] [ɥ]

Devant [i], la prononciation non syllabique [ɥ] est généralement obligatoire :

puis *[yi] [ɥi]

Toutefois, devant l'un des suffixes dérivationnels *-ir, -iser, isme, -iste* et *-ité*, la prononciation syllabique [y] est possible :

amuïr amuïssement [yi] [ɥi]
désambiguïser désambiguïsation [yi] [ɥi]
euphuisme [yi] [ɥi]
casuiste casuistique revuiste [yi] [ɥi]
absoluité acuité ambiguïté annuité assiduité contiguïté exiguïté ingénuité innocuité nocuité perpétuité promiscuité ténuité vacuité viduité [yi] [ɥi]

Devant [ɛ̃], la prononciation non syllabique [ɥ] est généralement obligatoire :

suint suinter suintant suintement *[ɥɛ̃] [yɛ̃]
chuinter chuintant chuintement *[ɥɛ̃] [yɛ̃]

Dans les mots d'origine étrangère *quindécemvir* et *quinto*, elle l'est aussi, mais dans *diminuendo*, elle est en concurrence avec la prononciation en [y].

Devant une autre voyelle, les deux prononciations sont généralement acceptables :

duel [yɛ] [ɥɛ]

On note toutefois les quelques contre-exemples suivants :

puéril -ement -isme -ité -cultrice -culture *[ye] [ɥe]
puerpéral *[yɛ] [ɥɛ̃]
lingual linguatule *[ya] [ɥa]

auxquels il faut adjoindre des mots d'origine étrangère tels que *huerta, marijuana* ou *quichua*.

V. 6) Conclusion

Les quatre variations phonétiques qu'on vient d'examiner ont un point commun. Elles concernent sélectivement certains mots et ne s'appliquent pas aux autres, sans que cette répartition soit nettement corrélée à une autre propriété phonétique. On constate une certaine corrélation avec des informations morphologiques telles que la présence d'un suffixe. Cependant, cette corrélation n'est pas parfaite et il n'existe pas toujours de critères opératoires permettant de décider si un élément terminal donné est ou non un suffixe. Lors de la description systématique de ces quatre variations phonétiques, leur extension lexicale a donc été déterminée en marquant dans le dictionnaire les mots concernés.

BIBLIOGRAPHIE

[1] Takao WATANABE, 1986, "Syllable Recognition for Continuous Japanese Speech Recognition", *ICASSP 86*, Tokyo, pp. 2295-2298.

[2] Takao WATANABE, 1983, "Segmentation-Free Syllable Recognition in Continuously Spoken Japanese", *ICASSP 83*, Boston, pp. 320-323.

[3] R. VIVES, 1985, "Mise en correspondance temporelle de descriptions phonologique et prosodique de mots dans le système de reconnaissance de la parole KEAL", *14° JEP*, Paris, pp. 253-256.

[4] P. S. COHEN & R. L. MERCER, 1974, "The Phonological Component of an Automatic Speech-Recognition System", in *Proceedings of IEEE Symposium on Speech Recognition*, pp. 177-188.

[5] Martine DECKER, Jean-Luc GAUVAIN, Jean-Joseph MARIANI, 1985, "Reconnaissance de mots isolés par diphonèmes enchaînés", *14° JEP*, Paris, pp. 271-274.

[6] June E. SHOUP, 1980, "Phonological Aspects of Speech Recognition", in *Trends in Speech Recognition*, Wayne A. LEA, Prentice-Hall, pp. 125-138.

[7] Guy PERENNOU & Martine DE CALMES, 1986, "BDLEX : une base de données et de connaissances du français parlé", in *Lexiques et traitement automatique des langages*, université Paul Sabatier, Toulouse, pp. 243-258.

[8] Mario ROSSI & Marilyn LAMBERT-DRACHE, 1986, "Traitement des voyelles à timbres multiples, le cas de /E/", in *Lexiques et traitement automatique des langages*, université Paul Sabatier, Toulouse, pp. 141-161.

[9] Françoise NEEL, Maxine ESKENAZI, Jean-Joseph MARIANI, 1986, "Module de traduction phonétique avec variantes", in *Lexiques et traitement automatique des langages*, université Paul Sabatier, Toulouse, pp. 129-138.

[10] Jacqueline VAISSIERE, 1985, "Etude des variations allophoniques de la voyelle /a/ et ses conséquences pour la reconnaissance automatique de la parole", *14° JEP*, Paris, pp. 304-307.

[1] Suivant les locuteurs, ou même, pour un locuteur donné, d'un moment à l'autre.

[2] Plusieurs définitions de la notion de centre de syllabe sont envisageables, par exemple le centre temporel de la voyelle, le maximum d'énergie, ou le minimum de variabilité par rapport au temps. Par ailleurs, le découpage envisagé peut tenir compte non seulement des centres de syllabes mais aussi des silences, car les variations contextuelles de part et d'autre d'un silence sont limitées.

[3] Plusieurs variations phonétiques peuvent intervenir par rapport à la prononciation donnée ici, en particulier pour ce qui est des deux liaisons facultatives et du timbre des deux [e]. Ces variations se répercutent sur la suite des unités à reconnaître.

[4] Les verbes étant à l'infinitif, et les noms et les adjectifs au masculin singulier. L'extension à 60 000 mots est en cours.

[5] Les verbes étant conjugués, et les noms et les adjectifs mis au féminin et au pluriel.

Hamlet : Un Prototype de Machine à Ecrire à Entrée Vocale.

*J. Mariani**

LIMSI/CNRS
BP30
91406 Orsay Cedex (France)
&
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, New York 10598 (USA)

Abstract

This project integrates different parts of a speaker-dependent, isolated-word Voice Activated Typewriter on a Personal Computer (IBM PC-AT).

In order to build up the language model (for French), several routines have been written : automatic grapheme-to-phoneme conversion, semi-automatic training texts (20 pages) processing (building up the Graphemic (2,500 words) and Phonemic (2,000 words) lexicons, syntactic labelling through inductive inference), computation of the probabilistic language model (bigrams and trigrams), definition of the phonological rules.

The speech signal is analysed by 20 digital band-pass filters. Several types of speech compression techniques have been compared on medium and large difficulty vocabularies. Vector Quantization and Non-Linear Time Compression have been choosen.

Recognition is conducted in 3 steps : i) Fast Match based on word length and gross comparison. ii) Detailed match based on conventional DTW algorithms. iii) Use of the language model to take into account the linguistic constraints, and to achieve the grapheme-to-phoneme conversion.

Overall recognition rates of 95% have been obtained with a mean recognition time of 2 s., the 2,000 templates being stored in 60 KBytes of RAM memory.

Introduction

Depuis les premières réalisations fructueuses, les systèmes de reconnaissance vocale ont subi des améliorations suivant trois axes indépendamment : du monolocuteur au multilocuteur, de la prononciation par mots isolés à la prononciation par mots enchaînés, et, plus récemment (L. Bahl, 1983, J. Baker, 1986, W. Meisel, 1986, R. Kurzweil, 1986, J.L. Gauvain, 1986, B. Merialdo, 1987, W. Drews, 1987, R. Campo, 1987...), de la reconnaissance de vocabulaires de taille réduite (quelques dizaines de mots) à la reconnaissance de grands vocabulaires allant jusqu'à 20,000 mots (Averbuch et al., 1987).

Le but de notre travail a été d'intégrer les différentes composantes d'une Machine à Ecrire à entrée Vocale (MEV) monoluteur, et par mots isolés (reconnaissance lexicale, traduction graphémique) sur un microordinateur personnel autonome, n'utilisant pas de processeur spécialisé pour effectuer l'algorithme de reconnaissance. Ce pour un vocabulaire de quelques milliers de mots.

Un autre point d'intérêt était de mesurer la faculté d'un modèle de langage de type "langue naturelle" aussi bien à effectuer la traduction phonème-graphème, qu'à corriger les erreurs de reconnaissance lexicale.

*Ce travail a été effectué lors de mon année sabbatique au T.J. Watson Research Center d'IBM, de Septembre 1985 à Août 1986.

Le projet s'est déroulé en trois phases : d'abord, un modèle de langage a été construit. Puis, des techniques de compression d'information et de reconnaissance ont été expérimentées dans le but d'obtenir une taille mémoire et un temps de réponse acceptable. Finalement, le modèle de langage a été introduit dans le processus de reconnaissance, et le système complet a été testé.

Pourquoi Hamlet ?

Le nom du système a été choisi en référence au problème de traduction graphémique de "To Be or not To Be" (qui pourrait s'écrire "2B or not 2B" dans une tâche de dictée de rapport) d'une part, et d'autre part, parceque le Livre Des Records Guiness rapporte des essais couronnés de succès pour prononcer les 262 mots du monologue d'Hamlet en moins de 24 secondes, soit un débit de 655 mots/minute (à comparer au record de frappe à la machine à écrire de 147 mots/minute par A. Tangora). A cette vitesse, le texte devient cependant totalement incompréhensible...

Le modèle de langage

Construction du modèle de langage

L'univers sémantique porte sur la dictée d'un rapport de recherche dans le domaine des technologies vocales, en Français. Le corpus d'apprentissage est constitué par 20 pages de texte (15,000 mots environ).

A un instant donné de l'apprentissage linguistique, une page de texte est analysée en utilisant le modèle de langage construit à partir des pages précédentes. Le texte est d'abord segmenté en mot, et chaque mot ainsi détecté est cherché dans le lexique. S'il est trouvé, sa représentation phonémique et sa catégorie grammaticale lui sont affectées. Sinon, sa représentation phonémique est obtenue à l'aide d'un programme de conversion automatique graphème-phonème, et sa catégorie grammaticale est inférée inductivement en utilisant une méthode d'analyse syntaxique probabiliste. Le résultat de cette analyse est un texte "traité" (appelé aussi texte "verticalisé" (L. Boves, 1987), où chaque mot graphémique détecté est accompagné de sa représentation phonémique, de sa catégorie grammaticale, et du type d'inférence (lexicale ou syntaxique) utilisée pour obtenir ces informations. Le texte est alors corrigé manuellement, et est utilisé pour mettre à jour le lexique, et la syntaxe, qui servent à leur tour à analyser la page suivante.

Conversion graphème-phonème

Plusieurs programmes de traduction graphème-phonème ont été écrit pour la synthèse de la parole à partir du texte. Celui que nous avons développé ici fonctionne au niveau morphologique. Il utilise un ensemble de règles déclaratives, les exceptions étant considérées comme des règles plus longues. Ces règles sont alors compilées sous forme de structure arborescente, pour accélérer le processus de conversion. Les règles de traduction ont été adaptées à la tâche particulière de la dictée vocale. Par exemple, les signes

de ponctuation ne sont pas prononcés explicitement en synthèse vocale, mais ils seront prononcés lors de la dictée d'un texte. 520 règles de traduction phonémique ont ainsi été définies, et testées.

Quelques problèmes relatifs à la dictée de textes en Français.

Certains de ces problèmes sont liés au fait que la prononciation se fait par mots isolés. Les choix suivants ont été faits : les nombres sont habituellement prononcés comme des suites de chiffres, sauf lorsqu'ils sont inclus dans un mot. L'apostrophe est prononcée comme un mot. Les liaisons entre les mots ne sont pas prononcées.

Des problèmes plus généraux concernent la traduction phonème-graphème des homophones en Français, qui semble être plus difficile que pour d'autres langages, même pour une prononciation par mots isolés.

L'un des problèmes principaux concerne la conjugaison des verbes , qui donne en moyenne 40 formes différentes pour un verbe, et jusqu'à 3 épellations différentes pour une même prononciation, ce pour tous les verbes.

La marque du pluriel de la plupart des substantifs, de la plupart des adjectifs, et de tous les participes passés (-s à la fin du mot) n'est jamais prononcée en mode isolé. La marque du féminin de certains substantifs, de la plupart des adjectifs et des participes passés (-e à la fin du mot) n'est pas prononcée si l'on parle naturellement sans effort particulier d'élocution, même si la prononciation se fait en mode isolé.

L'adjectif démonstratif "Ces" a la même prononciation /se/, mais une écriture différente que l'adjectif possessif "ses".

Règles phonologiques utilisées dans le système

A partir de ces constatations, des règles phonologiques ont été définies et sont utilisées dans le système. Par exemple, une règle énonce "qu'une apostrophe doit être suivie par un mot commençant par une voyelle". Une autre que "l'adjectif possessif "mon" ne peut être suivi par un mot féminin commençant par une consonne". 12 règles génèrales de ce type ont ainsi été définies.

Le lexique

Les mots contenus dans le lexique sont définis par leur forme graphémique, leur forme phonémique, et leur catégorie grammaticale. Au corpus de textes de 16000 mots, qui inclut un ensemble spécial de mots-outils grammaticaux, correspond un lexique d'environ 2500 mots graphémiques différents, et d'environ 2000 mots phonémiques différents. Pour accélérer le processus de recherche lexicale, le lexique phonémique est représenté par une structure arborescente, répondant à un "modèle de cohorte" (W. Marslen-Wilson, 1980). L'analyse des mots du lexique montre qu'environ 10% des mots sont des noms propres, des acronymes ou des mots étrangers.

Les catégories grammaticales

Nous avons défini 160 catégories grammaticales, proches de catégories utilisées dans d'autres systèmes (A. Andreewski, 1972, A.M. Derouault, 1985). Ces catégories sont obtenues à partir de 55 catégories de base, en ajoutant une information sur le genre ou le nombre, par exemple. Elles sont classées en catégories "fermées" (comme la catégorie "jour de la semaine", par exemple), qui comportent déjà tous les éléments qui les constituent, ceux-ci étant faciles à répertorier, ou en catégories "ouvertes". Cette différenciation est utilisée dans le processus d'inférence, lors de l'apprentissage linguistique.

Le modèle de langage

Le modèle de langage est représenté par une chaîne de Markov qui donne la probabilité d'occurence de deux (bigrammes) ou trois (trigrammes) catégories grammaticales successives (A. Andreewski, 1972, L.R. Bahl, 1978, A.M. Derouault, 1985) . Ces probabilités sont obtenues à partir du comptage de ces occurences dans les données d'apprentissage. Elles sont représentées par une structure arborescente. A chaque noeud de l'arbre de profondeur 3 est présente une catégorie grammaticale avec le comptage du nombre de fois où l'on est passé par ce noeud lors de l'apprentissage.

Analyse syntaxique

L'analyse syntaxique est faite en utilisant l'algorithme de Viterbi (G.D. Forney, 1973) appliqué au treillis obtenu en donnant à chaque mot de la phrase ses différentes catégories grammaticales possibles, que l'on trouve dans le lexique. Ce procédé d'analyse peut être étendu à une analyse syntaxique par défaut, lorsqu'un mot n'est pas présent dans le lexique, ou à la conversion phonème-graphème fondée sur une analyse syntaxique (A. Andreewski, 1979, A.M. Derouault, 1985, L. Boves, 1987).

Reconnaissance lexicale

Acquisition du signal

Le microphone est un microphone omnidirectionnel Crown PZM, qui est posé sur le clavier. Le niveau de bruit ambiant est moyen (ambiance de bureau). Le traitement du signal est effectué sur un module PI (Personal Instruments) d'IBM, basé sur Circuit Intégré spécialisé de traitement numérique du signal. La fréquence d'échantillonnage est de 20 Khz, chaque échantillon étant codé sur 12 bits. Le sonagramme est obtenu, après une FFT en 512 points, par un banc de 20 filtres passe-bande, suivant une échelle de Bandes Critiques. Chaque valeur à la sortie du filtre est codée logarithmiquement, le sonagramme est lissé, et l'on effectue une soustraction spectrale du bruit, et une normalisation d'amplitude. La détection du début et de la fin du message de parole est faite en utilisant différents seuils.

Compression de la parole

Différents algorithmes de compression de la parole ont été testés sur des vocabulaires difficiles (syllabes ou mots en "paire minimale"), de façon à tester la diminution de la qualité de la reconnaissance qu'on supposerait devoir se produire, du fait de la diminution d'information permettant de coder chaque mot. Au contraire, nous avons constaté que les meilleurs résultats de reconnaissance sont obtenus avec le taux de compression le plus important. Tout d'abord, une méthode de compression temporelle non-linéaire ("Variable Lenght Trace Segmentation", J.L. Gauvain, 1982, M.H. Kuhn, 1983) est appliquée, et donne un taux de compression de 2 environ. Le but de cet algorithme est de compresser les parties stables du signal. Puis, on applique un codage par quantification vectorielle (A. Buzo, 1979, J. Mariani, 1981...). Le Codebook est construit à l'aide d'une méthode de "covering", et contient 256 prototypes. Le sonagramme est alors codé en utilisant ces prototypes, et cette opération, pour sa part, permet d'obtenir un taux de compression supplémentaire de 20. Soit un taux de compression total de 40 environ.

Apprentissage

Pendant la phase d'apprentissage, tous les mots du lexique phonémique sont prononcés une fois. Ils sont compressés et stockés en mémoire, avec leur étiquette phonémique. Les 2000 références sont stockées dans 60 KOctets de mémoire vive (31 Octets/mot en moyenne). Nous comptons par la suite travailler sur la façon de faire l'adaptation au locuteur sur une petite quantité

de parole, par l'intermédiaire de la quantification vectorielle (K. Shikano, 1986, H. Bonneau, 1987).

Comparaison rapide

La première étape de la reconnaissance a pour but d'éliminer le plus possible de mots du lexique. Dans la mesure où le vocabulaire est important, cette phase doit se faire rapidement. Deux paramètres sont utilisés : d'abord, la longueur du mot à reconnaître est comparée à la longueur de chacune des références (longueur après compression temporelle dans les deux cas). L'utilisation de ce paramètre ne peut se faire qu'à la fin de la prononciation du mot à reconnaître. Puis une fonction de "similarité étendue" est calculée, de manière synchrone au mot-test. Elle donne la distance entre des zones du mot à reconnaître, et les prototypes du codebook. Cette fonction de similarité est alors utilisée pour calculer une distance de similarité grossière entre le mot à reconnaître et les références. Cette procédure de comparaison rapide permet de sélectionner en moyenne 24 mots (1,2% du lexique). Pour chacune des phases de cette comparaison, des seuils sont utilisés pour opérer la sélection.

Comparaison fine

Un procédé plus fin de reconnaissance est alors appliqué aux mots qui ont été sélectionnés. Cette comparaison se fait en utilisant un algorithme de Programmation Dynamique, asymétrique et synchrone au mot-test, faisant intervenir des contraintes de pente, et des seuils de réjection locale et globale. La note de reconnaissance de chacun des mots-candidats est normalisée dans (0,1). La taille moyenne des cohortes phonémiques est de 4 mots.

Utilisation du modèle de langage

La conversion phonème-graphème pour chacun des mots-candidats dans la cohorte phonémique se fait alors en allant chercher les différents mots graphémiques correspondants dans le lexique phonémique structuré de façon arborescente. La cohorte graphémique qui résulte de cette conversion contient les mots-candidats graphémiques avec leur note de reconnaissance, et leur catégorie grammaticale correspondantes. Le choix de la suite de mots graphémiques se fait en appliquant l'algorithme de Viterbi sur le treillis obtenu en combinant les mesures de distance acoustique des mots phonémiques, et les probabilités de succession des catégories grammaticales données par le modèle de langage.

Résultats

Nous avons effectués des tests, tout d'abord sur la dictée d'un texte de 100 mots. Au niveau acoustique, 9 erreurs ont été faites (91% de reconnaissance). Le modèle de langage complet, par trigrammes, corrige 6 erreurs, et n'en introduit aucune au cours de la conversion phonème-graphème améliorant donc le taux de reconnaissance à 97%. Le modèle de langage incomplet, où le texte à dicter n'est pas inclus dans le corpus d'apprentissage linguistique, en mélangeant bigrammes et trigrammes, corrige 5 erreurs, donnant un taux de reconnaissance de 95%. Cette différence avec les résultats obtenus en utilisant le modèle complet montre que l'apprentissage nécessiterait une quantité de données plus importante.
La figure 1 donne quelques exemples de transcription, dans les cas où la reconnaissance lexicale n'a pas fourni le mot correct en première position.

Des tests complémentaires ont été faits 3 mois plus tard, afin de tester le vieillissement des références, sur les 200 mots suivants du texte à dicter. Le taux de reconnaissance acoustique correcte est de 92,5%. L'introduction du modèle de langage porte ce taux à 95%. Les résultats complets sur les 300 mots du texte dicté sont donnés en figure 2.

Phrase correcte (PC):	Titres	et	Travaux	de		
Cohorte retenue (CR):	**titres**	clé	**travaux**	**de**(3)		
	titre	clés	avons(2)	deux		
		et		2		
		...		te(3)		
				peut		
				peu		
				...		

Figure 1.a. Exemple présentant les différents mots graphémiques.
(Le nombre de catégories grammaticales possibles pour chacun est donnée entre parenthèses.)

PC:	se	sont	portés	vers	les	problèmes	relatifs
CR:	**se**	**sont**	**portés**	air	**les**	programme	**relatifs**
	ce	son	porter	faire	lié	**problèmes**	relatif
	ceux	sons	portées	heure	clé	problème	
	CEE	soit		**vers**	clés		
	seul	sans		...	mes		
	...	ont			...		
		...					

Figure 1.b. Autre exemple avec deux erreurs successives.
(on ne fait pas apparaître ici le nombre de catégories pour chaque mot graphémique).

PC:	dans	le	cadre	de	ma	thèse	de	docteur
CR:	**dans**	**le**	4	te	**ma**,	**thèse**	ou	**docteur**
	tant	me	carte	**de**	bas		**au**	laquelle
	banc	eux	par	plus	pas		...	
	en	peut	**capot**	peut	la		...	
	...	...	car		...		de	
			cadre					

Figure 1.c. Exemple où le modèle de langage ne parvient pas à corriger l'erreur : trois erreurs de reconnaissance ont été faites dans un faible intervalle, et la première erreur porte sur un mot qui a la même catégorie grammaticale que le mot correct.
(Ici, un seul exemplaire graphémique est présenté pour chaque mot phonémique)

Figure 1. Exemples d'erreurs rattrapées par la syntaxe ou non.
(Les mots-candidats sont ordonnés suivant leur note de reconnaissance. Les mots reconnus finalement sont en caractères gras et soulignés)

Mode de reconnaissance :	% reco.
Mots phonémiques	92%
Mots phonémiques avec ML	98%
Mots graphémiques avec ML complet	95,7%
Mots graphémiques avec ML incomplet	
- trigrammes	92,7%
- bigrammes + trigrammes	95%

Figure 2. Résultats de reconnaissance globaux.

Le temps de reconnaissance moyen pour un mot est d'environ 2 secondes (1,8 s pour la reconnaissance acoustique, 0,2 s pour le traitement linguistique).

Discussion des résultats

L'analyse des résultats de reconnaissance par mots dans le premier test portant sur 100 mots montre que si les résultats de reconnaissance du mot correct en première position sont de 91%, ils atteignent 98% si l'on considère les 5 premiers candidats.

Toutes les erreurs sont commises sur des mots d'une ou deux syllabes. Dans la mesure où les mots courts sont aussi les plus fréquents, les taux de reconnaissance sur un texte dicté seront inférieurs à ceux obtenus sur la prononciation d'un lexique.

Le taux de reconnaissance ne varie pas lorsqu'on passe d'un lexique de 1500 mots au lexique de 2000 mots. Là aussi, le fait que les mots les plus susceptibles d'entraîner des erreurs sont les plus courts et les plus fréquents, amène à penser qu'ils seront rapidement présents dans le lexique, et que l'augmentation de la taille de celui-ci se fera par des mots longs, plus faciles à différencier. On peut donc penser que le taux d'erreurs n'augmente pas linéairement avec la taille du vocabulaire.

Enfin, on voit que les meilleurs résultats dans le cas d'un apprentissage "incomplet" sont ceux obtenus avec un modèle "bigrammes + trigrammes", dans la mesure où si le trigramme n'a pas été appris, les bigrammes correspondants peuvent l'avoir été.

Conclusion

Nous avons présenté un système de machine à écrire à entrée vocale, qui inclut la reconnaissance par mots et la traduction phonème-graphème. Nous avons montré que l'utilisation d'un modèle de langage améliore le taux de reconnaissance jusqu'à un niveau acceptable. Un taux de reconnaissance de 95% a été obtenu sur une tâche de dictée de texte. Nous pensons que l'utilisation réelle de tels systèmes dans le futur nécessitera une facile adaptation au locuteur, et la reconnaissance de parole continue.

Références

A. Andreewski, C. Fluhr, "Expériences de constitution d'un programme d'apprentissage pour le traitement automatique du langage.", Note CEA 1606, Décembre 1972

A. Andreewski, J.P. Binquet, F. Debili, C. Fluhr, Y. Hlal, J.S. Liénard, J. Mariani, B. Pouderoux, "Les dictionnaires en forme complète, et leur utilisation dans la transformation lexicale et syntaxique de chaînes phonétiques correctes.", 10èmes JEP du GALF, Grenoble, Mai 1979.

A. Averbuch et Al., "Experiments with the TANGORA 20,000-Word Speech Recognizer", IEEE ICASSP , Dallas, 1987

L.R. Bahl, R. Bakis, P.S. Cohen, F. Jelinek, B.L. Lewis, R.L. Mercer, " Recognition of a Continuously Read Natural Corpus.", IEEE ICASSP'78, Tulsa, Avril 1978

L. Bahl et Al., "Recognition of Isolated-Word sentences from a 5000-Word Vocabulary Office Correspondence Task", IEEE ICASSP, Boston, 1983

J.K. Baker, "Automatic Transcription on a Personal Computer", Speech Tech'86, New York, Avril 1986

H. Bonneau, J.L. Gauvain, ""Vector Quantization for Speaker Adaotation", IEEE ICASSP, Dallas, Avril 1987

L. Boves et Al., "The Linguistic Processor in a Multi-Lingual Text-to-Speech and Speech-to-Text System", European Conference on Speech Technology, Edinburg, Septembre 1987

A. Buzo, A.H. Gray, Jr, R.M. Gray, J.D. Markel, "A Two-Step Speech Compression with Vector Quantizing.", IEEE ICASSP'79, Washington, Avril 1979

A.M. Derouault, "Modélisation d'une langue naturelle pour la désambiguation des chaînes phonétiques.", Thèse de Doctorat d'Etat, Univ. Paris VII, Avril 1985

W. Drews, R. Lanoia, J. Pandel, A. Stoelzle, "A 1000 Word Speech Recognition System using a Special Purpose CMOS Processor", European Conference on Speech Technology, Edinburg, Septembre 1987

G.D. Forney, Jr, "The Viterbi Algorithm.", Proc. IEEE, Vol. 61, pp. 268-278, Mars 1973.

J.L. Gauvain, J. Mariani, "A Method for Connected Word Recognition and Word Spotting on a Microprocessor", IEEE ICASSP, Paris, Mai 1982

M.H. Kuhn, H.H. Tomaschewski, "Improvements in Isolated Word Recognition.", IEEE Trans. on ASSP, Vol. 31, N. 1, Février 1983.

R. Kurzweil, "The Kurzweil VoiceWriter. A Large Vocabulary Voice Activated Word Processor", Speech Tech'86, New York, Avril 1986

J. Mariani, "Reconnaissance de parole continue par diphonèmes.", Séminaire GALF "Processus d'encodage et de décodage phonétique", Toulouse, Septembre 1981

W.D. Marslen-Wilson, "Speech Understanding as a Psychological Process", in Spoken Language Generation and Understanding, NATO/ASI series, D. Reidel, 1980

W.S. Meisel, "Implications of Large Vocabulary Recognition", Speech Tech'86, New York, Avril 1986

B. Mérialdo, "Speech Recognition with very large size Dictionary.", IEEE ICASSP'87, Dallas, Avril 1987

G. Pirani, L. Fissore, A. Martelli, G. Volpi, "Experimental Evaluation of Italian Language Models for Large Dictionary Speech Recognition", European Conference on Speech Technology, Edinburg, Septembre 1987

K. Shikano, K.F. Lee, R. Reddy, "Speaker Adaptation through Vector Quantization", IEEE ICASSP, Tokyo, Avril 1986

RECONNAISSANCE DE PAROLE AVEC UN TRES GRAND VOCABULAIRE

Auteurs: B. Mérialdo, A.M. Derouault, M. Elbeze, S. Soudoplatoff

Centre Scientifique IBM France
36 avenue Raymond Poincaré, 75116 Paris FRANCE

We present our approach for Automatic Dictation in French. Our system is mono-speaker. In the first research phase, the sentences are uttered in Isolated Syllable mode, that is, with short pauses between syllables. This paper relates the state of our project and the recent efforts that have been done in the various domains of the recognition process. First, different types of acoustic processing have been experimented, by changing both the parameter space and the distances that were used on it. Criterion issued from both data analysis and information theory were used to evaluate the a-priori quality of those processings. Next, contextual information has been introduced into the Markov models for phonetic units. The analysis of the errors with standard decoding, as well as knowledge from phoneticians, allowed to classify the coarticulatory contexts and define an extended set of context-dependant phonetic units. The Multi-level decoding strategy has been set up. It allows to access the full French dictionary (200,000 inflected forms). At each time-slice, partial stacks of syllables and words are updated according to the acoustic observation. Finally, a syllabic prefilter has been proposed, to reduce the number of syllables candidates. It is based on classification of syllables according to vocalic kernels.

INTRODUCTION

Nous présentons ici notre approche de la dictee automatique en français. Un prototype acceptant les 10000 mots (fléchis) les plus fréquents a été décrit dans [1,2]. Rappelons que les phrases sont prononcées pour l'instant avec de courtes pauses entre les syllabes, et que le système est monolocuteur. Le point de vue adopté, aussi bien pour le niveau acoustico-phonétique, que le niveau linguistique, est celui de la théorie de l'information: le processus de reconnaissance consiste à trouver la suite de mots W qui maximise la probabilité conditionnelle P(W/A), où A désigne l'observation acoustique. Ceci revient à maximiser le produit de P(A/W)P(W). P(W) est donnée par le modèle de langage triclasses décrit en [4].

Nous exposerons successivement le choix de l'observation acoustique, la modélisation acoustico-phonétique, la stratégie de décodage et d'accès à un très grand dictionnaire, et les procédures de présélection des candidats syllabiques.

PARAMETRAGE ACOUSTIQUE

La paramétrisation acoustique est la première étape du traitement de la parole. Dans la formulation Markovienne que nous utilisons, cela correspond au calcul des paramètres A, qui sont les suites d'éléments observés. L'objectif de cette paramétrisation est multiple: il s'agit à la fois de comprimer le signal de parole, qui est trop riche, de faire des traitements qui sont "proches" de ceux qu'effectue une oreille humaine, et éventuellement de transformer l'information auditive de manière compréhensible pour la suite des opérations.

Après digitalisation du signal (10 kHz), le traitement actuel, très classique, s'effectue en trois étapes.

1. Différenciation et segmentation du signal en fenêtres de 256 échantillons, chaque 12.8 milliseconde.
2. Multiplication du signal par une fenêtre de Hanning, calcul de la transformée de Fourier, et sortie du spectre de puissance selon une double échelle logarithmique (vecteur acoustique).
3. Quantification vectorielle de ces vecteurs, sur 200 centres de classes appris sur un corpus, en utilisant l'algorithme classique du "K-mean".

L'utilisation de la quantification vectorielle s'explique afin que l'entrée acoustique du canal Markovien devienne une suite de symboles, au sein d'un vocabulaire limité. On évite ainsi d'avoir à faire une quelconque hypothèse sur la loi de probabilité des vecteurs acoustiques.

Si la première étape est relativement fixée, il n'en est pas de même des deux suivantes. Chaque modification entraîne un changement du système d'étiquettes utilisé en entrée du système de décodage. Nous avons expérimenté plusieurs types d'étiquetage, en faisant varier soit le mode d'obtention des vecteurs acoustiques, soit le type de quantification. Cependant, cette étape se trouvant la plus en amont du système de reconnaissance, nous avons cherché la possibilité d'existence de critères permettant d'évaluer a priori la qualité d'un système d'étiquettes, sans avoir à passer au travers de tout le processus d'apprentissage et de décodage.

Les différents types d'étiquettes que nous avons utilisés se trouvent résumés ici:

- Au niveau du traitement du signal
  - Fenêtres glissantes, spectre en échelle mel.
  - Fenêtres non glissantes.
  - Signal nul en dehors d'une période de pitch.
  - Signal non différencié.
  - Equalization spectrale.
- Au niveau de la quantification vectorielle
  - distance d1 ($\sum |x_i - y_i|$) utilisée.
  - distance d1 utilisée seulement pour les centres de non-silence dans le K-Mean.

Un décodage phonétique a été effectué pour chacun de ces systèmes d'étiquettes. Les taux de reconnaissance phonétiques obtenus

82.4 82.1 83.1 82.4 75.4 84.2 84.5

Montre un réel avantage pour la distance d1.

Au niveau des critères, une étude a montré que la notion de "vérité terrain", consistant en la donnée simultanée d'un corpus de signal et d'une segmentation en phonèmes correspondante, permettait d'en déduire des tables de contingences, qui peuvent également être assimilés à des estimateurs de probabilité p(e,f), e étant une étiquette et f un phonème ([8]). On peut alors calculer des critères, soit de type probabiliste (estimation du taux d'erreur, information mutuelle), soit de type contingentiel (critère du phi2, de Jordan).

Les courbes obtenues montrent que certain d'entre eux se comportent globalement bien, i.e. sont décroissant avec le taux d'erreur. Cependant, cette propriété n'est plus vraie de manière fine. Ceci est du au fait que d'une part les taux d'erreurs obtenus ne sont qu'une estimation du taux d'erreur absolu de la méthode, ni le corpus de test, ni les prononciations, n'étant illimités, d'autre part que les critères eux-même sont une estimation, et enfin que la "vérité terrain" n'est pas d'une justesse absolue.

MODELISATION ACOUSTIQUE.

Principe.

La modélisation acoustique comprend trois niveaux: un mot est décrit dans le dictionnaire comme une suite de syllabes phonétiques, une syllabe phonétique correspond à une suite d'unités phonétiques, et chaque unité phonétique est associée à une source de Markov, qui émet des labels.

Classiquement ([1]), le système d'unités phonétiques utilisé comprend 40 éléments correspondant aux phonèmes classiques et à quelques éléments supplémentaires comme le silence, les fins de syllabes (vocaliques ou consonantiques). Chaque source de Markov phonétique a 7 états, et produit des labels ([1]). Les paramètres sont appris sur 400 phrases d'apprentissage avec l'algorithme de Baum.

Analyse des erreurs.

Avec une machine par phonème, ces modèles ne rendent aucun compte des effets de coarticulation. Ils représentent un modèle moyen pour tous les contextes où apparaît un phonème donné. Il est clair que ceci est responsable d'un certain nombre d'erreurs de reconnaissance. Grâce à l'analyse systématique de ces erreurs, et des contextes dans lesquels elles se produisent, nous avons cherché à cerner les principaux effets coarticulatoires responsables. D'autre part, pour éviter un nombre de modèles trop grand et donc difficile à entraîner, nous avons utilisé cette analyse pour classifier les contextes d'une manière pertinente en vue d'une meilleure reconnaissance.

Les voyelles sont généralement bien reconnues malgré la coarticulation, grâce à leur partie stable. Ce sont les plosives qui posent le plus de problèmes, puis fricatives et liquides (confusions intra-classe).

L'analyse montre qu'un des effets dominants est l'influence du contexte vocalique sur la consonne précédente. Plus précisément, le lieu d'articulation de la voyelle suivante influe sur le lieu de la consonne (gardant le même mode d'articulation), provoquant ainsi des confusions intra-classe. Ceci est bien connu des phonéticiens sous le nom d'influence anticipatrice de la voyelle ([9]).

Choix d'un nouveau système.

Définition.

Les premières expériences décrites dans [5] ont montré l'intérêt de classifier les contextes droits des plosives sourdes en quatre types: voyelles d'avant, milieu ou arrière, liquides. Un modèle de Markov par type de contexte est associé à chaque plosive sourde. Au total un ensemble de 52 unités phonétiques a été testé du point de vue reconnaissance phonétique et reconnaissance de mots avec un dictionnaire de 10000 mots. La modification des modèles pour ces trois consonnes a déjà sensiblement amélioré ces taux d'erreurs ([5]).

Nous avons généralisé la méthode aux autres consonnes, et aux contextes droits non vocaliques. Une machine par type de contexte droit est définie pour chaque consonne et chaque semi-voyelle. Les contextes consonantiques (dans les groupes de consonnes: br, cl etc...) sont classifiés de la manière suivante: plosives sourdes, plosives voisées, fricatives sourdes, fricatives voisées, liquides, nasales.

Le nouveau système a 130 machines phonétiques contextuelles à 7 états chacune.

Apprentissage et équilibrage du corpus.

Avec ce nombre de machines, les paramètres à estimer sont beaucoup plus nombreux, et le corpus d'apprentissage a du être refait de manière à compter suffisamment d'occurences de chaque phonème et type de contexte. Ceci a été effectué d'une manière semi-automatique, en retenant les phrases les plus riches phonétiquement dans un ensemble de phrases françaises, et en rajoutant des phrases faites manuellement pour les contextes manquants. 250 phrases ont été retenues pour apprendre le système.

Résultats.

Ce système de phonèmes contextuels étendu a été testé en termes de reconnaissance phonétique sur 79 phrases prononcées en syllabes isolées: un alignement de Viterbi choisit la suite de phonèmes la plus probable pour la suite de labels obtenue. La plupart des consonnes sont effectivement mieux reconnues, comme le montre le tableau suivant. La première ligne montre les taux de reconnaissance obtenus avec le système sans contexte, la deuxième les taux avec le système contextuel:

P	T	K	B	D	G	F	S	CH	V	Z	JJ	M	N	L	R
55	70	57	25	64	43	85	89	100	75	89	86	52	88	76	74
64	78	81	42	64	57	91	94	100	80	81	90	67	81	80	87

Le taux de reconnaissance phonétique global (substitutions plus délétions) est passé de 80.5% avec un système de 40 phonèmes indépendants du contexte, à 84.5% avec le système contextuel.

DECODAGE MULTI-NIVEAUX

Le Décodage Multi-Niveaux *(Multi-Level Decoding, MLD)* ([7]) est une organisation du processus de reconnaissance qui permet d'utiliser des Très Grands Vocabulaires *(Very Large Size Dictionary, VLSD)* en reconnaissance de la parole. Il est basé sur l'enchaînement suivant:

- une reconnaissance acoustique d'unités sub-lexicales, dans notre cas des syllabes phonétiques,
- une composition d'unités successives pour former des mots, en utilisant un dictionnaire,
- une composition des mots successifs pour former des phrases partielles, en utilisant les contraintes du modèle linguistique.

Reconnaisseur de syllabes

Le dictionnaire de 200,000 mots déclinés donne lieu à 6.400 syllabes phonétiques (en tenant compte des phénomènes de liaison et d'apostrophe). Le reconnaisseur de syllabe cherche quelles sont les syllabes les plus probables qui correspondent à une portion donnée du signal de parole.

Chaque syllabe est représentée par une machine de Markov obtenue en concaténant les machines de Markov phonétiques correspondantes ([2]). L'ensemble de ces machines syllabiques est organisé en une arborescence, appelée arbre syllabique *(Syllabic tree, ST)*., dont les arcs correspondent aux machines phonétiques et les feuilles aux syllabes. Pour rechercher les syllabes les plus probables, on contruit le modèle de Markov suivant:

- l'arbre syllabique,
- suivi d'une machine de silence (destinée à coïncider avec la pause suivant la syllabe),
- suivi du 'modèle phonétique rebouclé', formé en plaçant en parallèle toutes les machines phonétiques, et en permettant une transition de la fin vers le début. Ce modèle est utilisé pour produire le reste de la phrase, considéré comme une suite de phonémes (en général, on impose une contrainte sur ces suites par une probabilité bi-phonèmes).

Pour trouver les syllabes les plus probables, on calcule la probabilité que ce modèle produise l'observation acoustique correspondant au signal de parole. On s'intéresse alors à la probabilité de passer par l'état final d'une syllabe au cours de cette production. Les syllabes les plus probables sont celles pour lesquelles cette probabilité est la plus forte.

Ensuite, pour une syllabe donnée, on regarde quelle est la probabilité de passer par l'état final de la syllabe à un instant t. L'instant où cette probabilité est maximale donne la fin la plus probable de la prononciation de la syllabe.

Le résultat fourni par le reconnaisseur de syllabes consiste en une liste de syllabes phonétiques candidates, chacune avec sa probabilité acoustique, et sa fin la plus probable.

Reconnaisseur de mots

Le reconnaisseur de mots combine les syllabes phonétiques candidates pour fournir des mots candidats. Le dictionnaire contient 200.000 entrées, chaque entrée étant composée de:

- l'orthographe du mot décliné,
- sa phonétique (avec éventuellement des variantes),
- sa ou ses classes grammaticales et les fréquences correspondantes,

Le reconnaisseur de mots utilise une arborescence mot/syllabe *(word/syllable, W/S)*, qui décrit la phonétique des mots comme suite de syllabes phonétiques. Dans cette arborescence, les arcs sont des syllabes phonétiques, les noeuds correspondent à des mots partiels, et les feuilles correspondent à des mots complets.

A chaque instant de la reconnaissance, le reconnaisseur de mots conserve une liste des mots partiels (noeuds) les plus probables correspondant à la prononciation d'une partie de la phrase. En fonction des syllabes candidates fournies par le reconnaisseur de syllabes pour la suite de la phrase, ces mots partiels sont prolongés dans l'arborescence W/S. Lorsqu'une feuille de l'arborescence est atteinte, un mot candidat est produit.

L'arborescence contient également des 'racines secondaires' pour traiter les cas provenant de la liaison et de l'apostrophe. Pour chaque consonne ou groupe consonnantique donnant lieu à liaison ou apostrophe *(l, z, t...)*, on crée une racine secondaire d'où partent des syllabes formées par la concaténation de cette consonne et d'une syllabe commençant par une voyelle phonétique. Par exemple, la syllabe e donnera lieu aux syllabes **le, ze, te...** partant des racines secondaires correspondantes. Des marqueurs permettent de savoir quels mots donnent lieu à liaison ou apostrophe, et quelle racine secondaire ils peuvent utiliser.

Le reconnaisseur de phrases

Le reconnaisseur de phrases combine les mots candidats pour former des phrases partielles. A chaque instant de la reconnaissance, il conserve une liste des phrases partielles correspondant à une partie de la phrase. Lorsque de nouveaux mots candidats sont produits, ces phrases partielles sont prolongées par ces nouveaux mots. Le modèle de langage tri-classes ([4]) est utilisé pour imposer une contrainte linguistique sur les suites de mots ainsi formées.

Ce processus se poursuit jusqu'à ce que la phrase entière ait été traitée. La phrase complète la plus probable est alors fournie comme transcription de l'enregistrement.

Résultats

Nous avons étudié l'influence de la taille du dictionnaire sur la performance du reconnaisseur. Plus le dictionnaire est grand, plus sa couverture est grande (nombre de mots du texte à reconnaître qui sont dans le dictionnaire), mais aussi plus le nombre d'ambiguïtés (donc d'erreurs potentielles) est grand. Le problème est de savoir lequel de ces deux facteurs l'emporte sur l'autre.

Nous avons réalisé l'expérience suivante:

- un texte à reconnaître de 722 mots (79 phrases),
- 3 dictionnaires différents:
  1. les 10.000 mots les plus fréquents (couverture du texte 94%),
  2. les 10.000 mots les plus fréquents plus les 43 mots du texte qui manquent (couverture 100%),
  3. le dictionnaire de 200.000 mots (couverture 100%).

Les taux d'erreurs (sur les mots) sont les suivants:

10,000 mots	17.3%
10,000 mots + 43	10.6%
200,000 mots	12.7%

Donc, en passant de 10.000 à 200.000 mots, on gagne 6% grâce à la meilleure couverture, et on perd 2% à cause du plus grand nombre d'ambiguïtés. Globalement, on gagne 4% si on utilise le dictionnaire de 200.000 mots par rapport au dictionnaire de 10.000 mots.

Ces résultats confirme l'utilité des grands vocabulaires pour les systèmes de dictée automatique.

PRE-FILTRE SYLLABIQUE

Le choix de l'unité syllabique permet d'envisager l'accès à un très grand dictionnaire. Il serait cependant très souhaitable de faire précéder la phase de reconnaissance acoustique d' une étape de pré-décodage fiable bien que grossier, qui détermine le plus rapidement possible au vu de l'observation acoustique un sous-ensemble de l'ensemble des syllabes évitant une recherche exhaustive. La liste de syllabes candidates ainsi obtenue doit être d'une part suffisamment longue pour contenir la bonne syllabe avec un taux de certitude de l'ordre de 99 %, d'autre part suffisamment courte pour jouer le rôle réducteur qui lui est fixé. Pour remplir ces deux contraintes, la trame de ce premier tamis ne doit être ni trop fine ni trop grossière.

Principe.

Dans ce but, nous avons classifié les syllabes en utilisant diverses classifications phonétiques. Chaque représentant de classe est associé à un modèle markovien, obtenu comme concaténation des machines phonétiques grossières. Le prédécodage des classes syllabiques les plus probables permet de proposer une liste de syllabes candidates.

Choix des classifications.

Le statut linguistique de la syllabe est assez bien défini. Une syllabe contient forcément une voyelle phonétique et une seule, donc d'un point de vue acoustique, une partie stable et durable constituée par le noyau vocalique. Si l'on excepte quelques oppositions ouvertes / fermées, il est en général assez aisé de différencier les voyelles entre elles, ce qui n'est pas le cas des consonnes.

Ces considérations nous permettent de penser que le pré-filtre sera d'autant plus performant que la classification sur laquelle il repose accordera un traitement privilégié aux voyelles par rapport aux consonnes.

Classification G1. Dans cette voie, nous proposons une première classification des syllabes. Les suites de consonnes et semi-consonnes sont toutes regroupées en une seule classe grossière C alors que chacune des voyelles constitue à elle seule une classe, (modulo la distinction ouverte/fermée).

Classification G2: Sans privilégier plus particulièrement les voyelles, Chen et Zue [3] répartissent les phonèmes en (liquide ou semi- voyelles), occlusives, voyelles orales, voyelles et consonnes nasales, fricatives sourdes, fricatives voisées.

Classification G3.: Les voyelles sont ramenées aux 10 archi-phonèmes vocaliques. Il y a 237 combinaisons de consonnes et de semi-consonnes, qui peuvent précéder ou suivre la voyelle d'une syllabe. Ces 237 suites sont réparties en 4 classes consonnantiques: liquide ou semi-voyelle, occlusives, nasales, fricatives.

Classification automatique G4 Mérialdo et al. [6] ont introduit une classification hiérarchique fondée sur un indice de similarité $s(,)$ entre 2 machines de markov phonétiques M_1 et M_2 défini comme suit :: $s(M_1,M_2) = \sum p(A \mid M_1) \times p(A \mid M_2)$ où A désigne une observation acoustique.

Nous l'avons utilisé pour classifier les 237 suites consonnantiques dans le but d'obtenir une répartition plus justifiée que celle opérée "de visu" par la classification G3. Les 19 consonnes sont classifiées par $s(,)$. On obtient les classes suivantes:

- -P -T -K -B -D -G CH -F JJ -V -M -N -L -R
- -U AU -O AN -W
- -A IN UN
- -E AI -I -Y OE EU -J UU
- -S -Z
- ON

Chaque suite de machines consonnantiques $M_i^?$, est affectée à la classes de la machine consonnantique $\tilde{M}$ pour laquelle $s(M_i^?, M)$ est maximal.

Résultats.

Lors de la phase de reconnaissance, le pré-filtre consiste en un alignement de Viterbi entre les représentants de classes syllabiques et 1144 syllabes d'un texte test prononcé par le même locuteur. Les N meilleures classes candidates sont soumises, dans un deuxième temps, à un décodage détaillé sur les syllabes membres des classes retenues. (Nous ne présentons ici que les expériences concernant le pré-filtre). Les syllabes appartenant aux meilleures classes candidates sont ordonnées et l'on peut ainsi déterminer la longueur de liste en fonction d'un pourcentage de certitude avec lequel le pré-filtre peut garantir la présence de la bonne syllabe. Les meilleurs résultats sont obtenus pour les classifications G3 et G4, c'est à dire que le pré-filtre donne les meilleures performances avec les classifications avantageant les voyelles. La comparaison décisive entre les différentes classifications, se fait en additionnant la longueur de liste des syllabes relative à un taux de de confiance de 99%, et le nombre de classes testées. Le tableau suivant montre ainsi que G4 minimise le coût total du décodage syllabique.

classification	G4	G3	G1	G2
coût total	435	448	603	1135

En tolérant une erreur de l'ordre de 1%, on arrive à remplacer 6164 alignements de Viterbi par 274 alignements de cohortes (pré-filtre) et un alignement postérieur (décodage détaillé) de 161 syllabes soit un total de 435 qui équivaut à un facteur de réduction 14.17 des calculs.

Bibliographie

[1] H. Cerf, A.M. Derouault, M. El-Beze, B. Merialdo, S. Soudoplatoff: **"Reconnaissance de la parole par des modèles markoviens: application aux grands vocabulaires"**, 15èmes Journées d'études sur la parole, Aix en Provence, 27-29 mai 1986.

[2] H. Cerf, A-M Derouault, M. El-Beze, B. Merialdo, S. Soudoplatoff: **"Speech Recognition experiment with 10,000 word vocabulary"**, NATO Advanced Institute on Pattern Recognition, 18-20 juin 1986, Bruxelles.

[3] F.R. Chen, V.W. Zue: **"Application of allophonic and lexical constraints in continuous digit recognition"**, ICASSP 1984, San Diego.

[4] A.M. Derouault, B. Merialdo: **Natural Language Modeling for Phoneme-to-Text Transcription**, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 8 , Number 6, November 1986, pp 742-749

[5] A.M. Derouault: **"Context-dependant phonetic Markov Models for large vocabulary speech recognition"**, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, April 6-9 1987, Dallas, USA.

[6] B. Merialdo, A.M. Derouault, S. Soudoplatoff: **"Phoneme classification using Markov models"**, ICASSP 1986, Tokyo.

[7] B. Merialdo: **"Speech Recognition with very large size dictionary"**, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, April 6-9 1987, Dallas, USA.

[8] S. Soudoplatoff: **"Speech Decoding using Markov model: search for a prior criterion of quality"**, European conference on Speech Technology, Edinburgh, September 1987.

[9] J.P. Zerling: **"Articulation et coarticulation dans des groupes occlusives-voyelles en français"**, 3rd cycle Thesis, University Nancy II.

Résultats de tests de reconnaissance en mots isolés sur des signaux bruités.

P. Wacrenier

LIMSI/CNRS - BP.30 - 91400 ORSAY CEDEX - FRANCE

Abstract

This article presents the results of two experiments concerning isolated word recognition in noisy environments. In the first experiment, the training phase is carried out in a noise-free environment ("infinite" Signal/Noise ratio). In the second experiment, the training phase is carried out in noise (Signal/Noise ratio = +9 dB). Tests are conducted in different noise-level environments.

For each experiment, we have used an implicit end-point detection, that is, a "word spotting" technique. We have then compared the results obtained without noise processing, or with marking or masking techniques. Results show great variability in recognition scores according to whitch noise processing technique is chosen.

1 Introduction

Cet article présente les résultats de deux expériences de reconnaissance en mots isolés sur des signaux bruités du groupe RSG10 du NATO. La première expérience consiste à réaliser un apprentissage en "environnement calme" (rapport signal/bruit infini), et d'effectuer les tests de reconnaissance en présence de bruit. Lors de la deuxième expérience, l'apprentissage a lieu en présence de bruit (rapport signal/bruit de +9 dB), tandis que la phase de reconnaissance s'effectue avec des niveaux de bruit variables.

De plus, pour chaque expérience, nous avons comparé les résultats obtenus par une méthode de comparaison dynamique sans contrainte de début et de fin de mot, à ceux obtenus lorsque l'on inclut dans cette méthode des techniques de marquage et masquage de bruit.

2 Description du corpus

Les signaux utilisés dans cette étude sont constitués des enregistrements en langue anglaise des tables SB, 1A, 1B et 1C du corpus du groupe RSG 10 du NATO, sur lesquels on a ajouté artificiellement du bruit stationnaire et non stationnaire, de manière à obtenir des rapports signal/bruit de +9 dB, +3 dB et -3 dB.

3 Analyse du signal

Le signal analogique échantillonné à la frequence de 10 KHz, tout d'abord multiplié par une fenêtre rectangulaire de 25.6 ms (256 echantillons) puis pondéré par une fenêtre de Hamming, et enfin préaccentué par le filtre de fonction de transfert H(z) = 1+0.95/z, est ensuite analysé par un algorithme de transformée de Fourier rapide toutes les 12,8 ms.

Le résultat de l'analyse de Fourier est alors traité par un banc de 16 filtres triangulaires répartis suivant une échelle de Bark. Les filtres ont une largeur de 0.94 Bark et se recouvrent à 3dB. La figure 1 représente le gain en dB de chaque filtre en fonction de la fréquence.

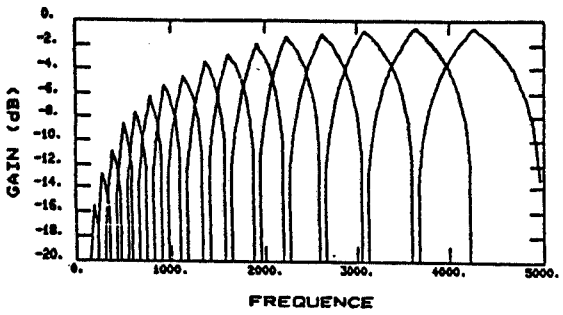


Figure 1: Gain des filtres en fonction de la fréquence.

Le banc de filtres fournit donc l'amplitude des 16 canaux toutes les 12.8 ms.

4 Caractéristiques de l'algorithme de programmation dynamique

L'équation locale utilisée dans cet algorithme de programmation dynamique est représentée par la figure 2. Cette équation est asymetrique le long de la référence afin d'obtenir une distance cumulée indépendante de la longueur du segment de signal à reconnaitre.

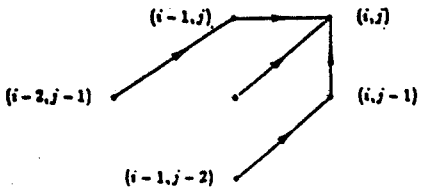


Figure 2: Equation locale de programmation dynamique.

Si $d(i,j)$ représente la distance locale entre le spectre d'indice i du segment à reconnaitre et le spectre d'indice j de la référence, et si $g(i,j)$ représente la distance cumulée le long du chemin optimal allant du point (1,1) au point (i,j) de la matrice de comparaison, alors $g(i,j)$ s'obtient de manière récursive par la relation:

$$g(i,j) = min \begin{cases} g(i-2,j-1) + d(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2 * d(i,j) \\ g(i-1,j-2) + 2 * d(i,j-1) + 2 * d(i,j) \end{cases}$$

La distance locale entre deux spectres est calculée à l'aide de la distance de Minkowsky d'ordre 1, appliquée à une paramétrisation interne d'un spectre. Deux paramétrisations internes ont été testées: la première paramétrisation consiste à calculer la distance filtre à filtre, la deuxième donne l'écart de chaque filtre par rapport à la moyenne.

De plus, comme il est très difficile, en milieu bruité, de connaitre avec précision le début et la fin d'un mot, on n'impose au processus de comparaison dynamique aucune contrainte aux frontières (principe de détection de mots).

5 Methodes de marquage et masquage de bruit

L'algorithme de comparaison dynamique décrit dans le paragraphe précédent ne tient pas compte de la nature bruitée des signaux que l'on traite. C'est pourquoi nous avons inseré dans cet algorithme, les techniques de marquage et masquage de bruit développées par Klatt, Bridle et Holmes [3] dans le but d'améliorer les performances de reconnaissance lorsque le rapport signal/bruit diminue. Le principe de ces méthodes consiste à utiliser une estimation du bruit pendant chaque mot de référence et chaque mot test, dans le calcul de la distance locale du processus de comparaison dynamique. L'estimation du bruit pendant chaque énoncé est déterminée en conservant le dernier spectre avant le début du mot.

5.1 Methode Klatt

La méthode proposée par Klatt utilise un masque de bruit pour calculer la distance entre chaque paire de spectres. Pour chaque canal, le masque est égal à la plus grande valeur des estimations de bruit correspondant au mot de référence et au mot test. Chaque coefficient spectral, de la référence et du mot à reconnaitre, qui est inférieur à la valeur correspondante du masque de bruit, est remplacé par cette derniere. Les spectres ainsi masqués sont alors comparés par la distance locale décrite dans le paragraphe précédent.

5.2 Methode Bridle

Contrairement à la méthode de Klatt, Bridle propose d'utiliser un marquage de chaque coefficient spectral, de la référence et du mot à reconnaitre, qui est inférieur à la valeur correspondante du bruit. Le calcul de la distance entre les spectres s'obtient en testant quels sont les coefficients spectraux qui ont été marqués. Si la plus grande valeur des deux amplitudes spectrales à comparer n'est pas marquée, alors la distance usuelle est utilisée. Dans le cas contraire, la distance entre ces deux amplitudes est fixée à une valeur arbitraire constante (égale à 10 dans notre programme).

5.3 Methode Holmes

La méthode proposée par Holmes utilise le masquage de chaque coefficient spectral des mots de référence, inférieur à la plus grande valeur de bruit correspondante trouvée pendant toute la période d'apprentissage, par cette dernière valeur. De plus, chaque coefficient spectral d'un mot à reconnaitre, inférieur à la valeur correspondante de son spectre de bruit, est masqué par cette valeur. Enfin, un marquage est aussi utilisé lorsqu'une valeur de bruit a masqué l'amplitude d'un canal. Dans ces conditions, la distance usuelle entre chaque paire d'amplitudes spectrales est utilisée, sauf dans le cas où le coefficient spectral de la référence a été marqué, et s'il est supérieur à la valeur correspondante du mot test. Dans ce cas particulier, une distance nulle est alors utilisée.

6 Conditions experimentales

Pour les deux expériences que nous présentons, une segmentation "fine" des mots a été réalisée manuellement. Il faut cependant noter que la segmentation a été impossible à réaliser sur les signaux où du bruit non-stationnaire a été ajouté pour donner un rapport signal/bruit de -3 dB.

Dans le cadre de la première expérience, la première occurrence de chaque chiffre, prononcé sans bruit, a servi de référence. Au cours de la deuxième expérience, les références sont les premières prononciations, en présence de bruit (rapport signal/bruit de +9 dB), de chaque chiffre. La phase de reconnaissance est alors effectuée sur l'ensemble de la base de données.

7 Resultats et conclusion

Les résultats synthétiques des deux expériences figurent dans les courbes suivantes. Chaque courbe donne les taux de reconnaissance en fonction du rapport signal/bruit, suivant:

- le type de représentation interne d'un spectre: différence filtre à filtre ou écart par rapport à la moyenne,
- la méthode utilisée: avec ou sans technique de marquage ou masquage de bruit,
- et le type de bruit ajouté au signal analogique: stationnaire ou non-stationnaire.

Ces courbes soulignent la grande variabilité des performances de reconnaissance en fonction du type de paramétrisation spectrale. Dans le cadre de la première expérience, et en présence de bruit stationnaire, la méthode de masquage de bruit proposée par Klatt, semble donner les meilleurs résultats lorsque l'on utilise une paramétrisation spectrale du type écart par rapport à la moyenne. En présence de bruit non-stationnaire, les meilleurs résultats sont obtenus sans traitement de bruit. Dans le cadre de la deuxième expérience, avec des bruits stationnaires ou non-stationnaires, la technique de marquage de bruit proposée par Bridle donne des performances de reconnaissance équivalentes à celles obtenues sans traitement de bruit lorsqu'une paramétrisation du type distance filtre à filtre est utilisée.

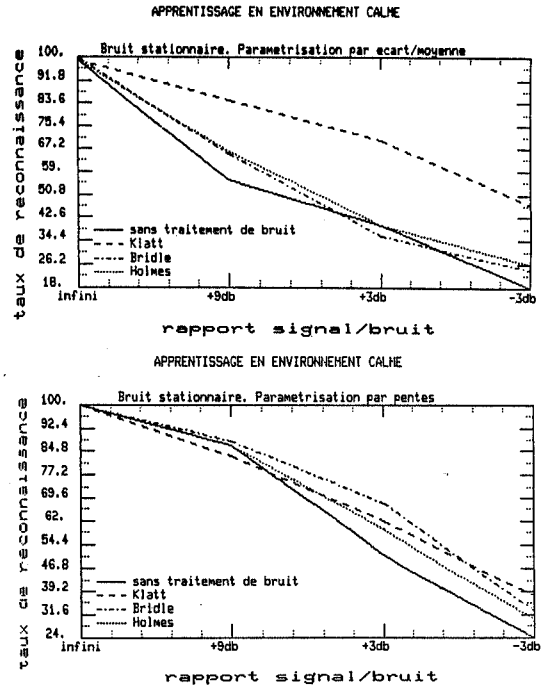


Figure 3: Expérience 1, avec du bruit stationnaire.

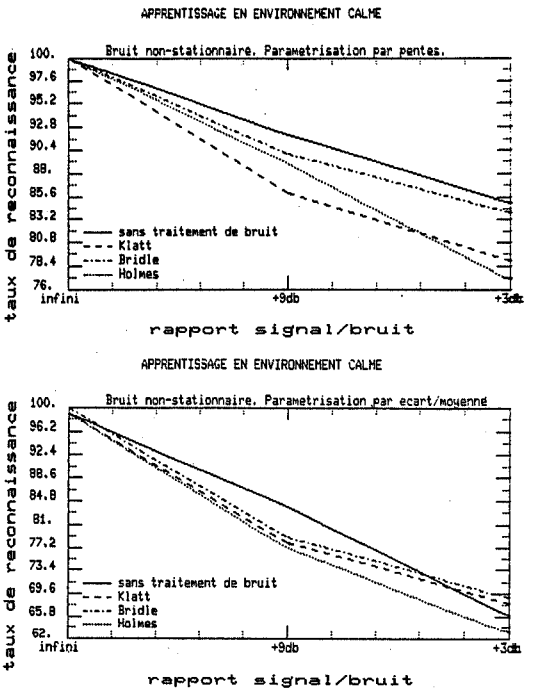


Figure 4: Expérience 1, avec du bruit non-stationnaire.

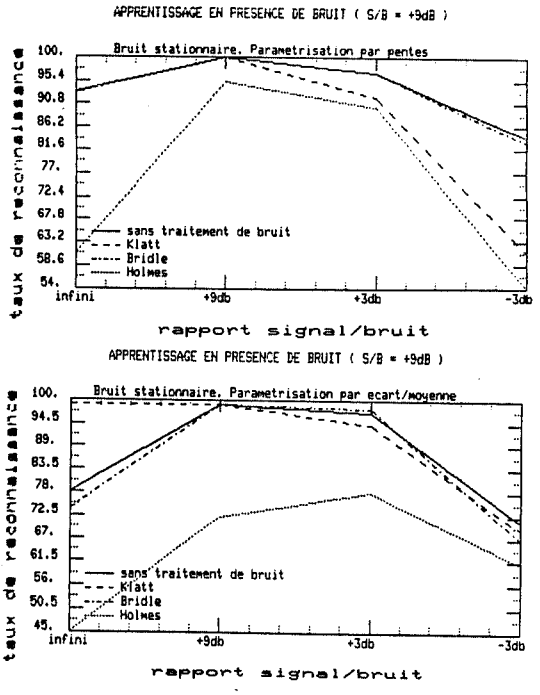


Figure 5: Expérience 2, avec du bruit stationnaire.

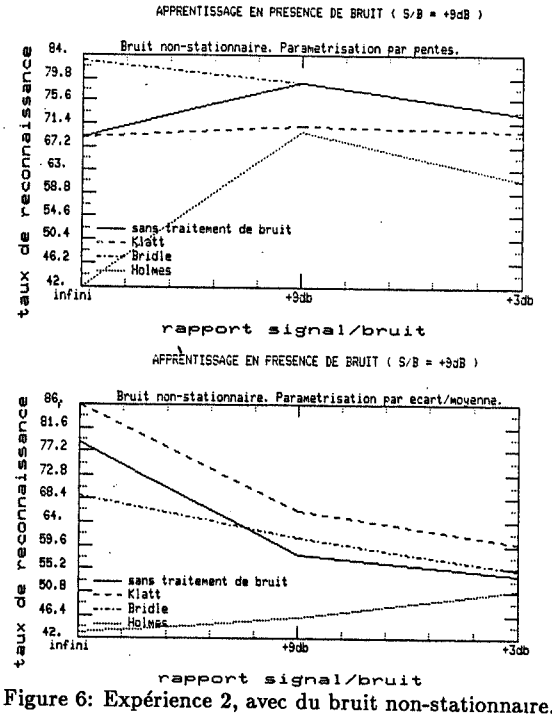


Figure 6: Expérience 2, avec du bruit non-stationnaire.

References

[1] D.H. Klatt: A Digital Filter Bank for Spectral Matching. Proc. IEEE ICASSP,Philadelphia, 1976.

[2] J.S. Bridle, K.M. Ponting, M.D. Brown, A.W. Borret: A Noise Compensating Spectrum Distance Measure Applied to Speech Recognition. Proc. Institute of Acoustics (UK), Windermere, Nov. 1984.

[3] J.N. Holmes, N.C. Sedgwick: Noise Compensation for Speech Recognition Using Probabilistic Model. Proc. IEEE ICASSP, Tokyo, 1986.

UN SYSTEME DE VERIFICATION AUTOMATIQUE DU LOCUTEUR (*)

P. VANUXEEM - M. INVERNIZZI

CENTRE D'ETUDES NUCLEAIRES DE SACLAY

ABSTRACT

This paper presents the results of the studies led at CEN SACLAY (1) in speaker verification. A part of these studies has been realized within the context of the european ESPRIT project number 64. The first part is devoted to the analysis of different parameters on a recorded database of 30 speakers involving 10 sets of 16 isolated words per speaker. An automatic evolutive test program allows the processing of these parameters and the evaluation of the average equal error rate in verification. The second part describes an automatic real time verification system on VAX 785 drifting from works carried out on the database.

INTRODUCTION

Par opposition à la reconnaissance de mots indépendante du locuteur où les systèmes tentent soit d'atténuer l'influence de la variabilité inter-locuteur soit d'effectuer une adaptation aux nouveaux locuteurs, la vérification du locuteur cherche les paramètres les plus discriminants permettant de caractériser un locuteur spécifique tout en essayant de le distinguer des autres.

Bien que la vérification se distingue de l'identification par le fait que le système compare le locuteur entrant dans le système uniquement avec l'identité qu'il fournit, une étude des variations intra-locuteur et inter-locuteur est nécessaire.

Les tests entrepris sur la base de données sont axés dans ce sens et ont permis de rechercher parmi un ensemble de paramètres, ceux donnant un taux d'erreur égal moyen (fausse acceptation de locuteurs imposteurs = faux rejet de bons locuteurs) de vérification le plus faible possible.

Nous aborderons tout d'abord le principe de la méthode employée en test en décrivant le programme de test automatique dont les résultats seront présentés. Puis nous examinerons l'architecture du système de vérification automatique du locuteur ainsi que quelques résultats préliminaires d'évaluation.

LA BASE DE DONNEE

La base de donnée utilisée par le programme de test est constituée par :

- la base JUILLET contenant 30 locuteurs (20 hommes, 10 femmes) ;
- les bases SEPTEMBRE, OCTOBRE, NOVEMBRE contenant chacune les 10 premiers locuteurs (6 hommes, 4 femmes) de la base de JUILLET ;
- la base DECEMBRE contenant les mêmes 30 locuteurs de la base de JUILLET.

Ces différentes bases sont espacées d'environ 1 mois.

(1) Institut de Recherche et de Développement Industriel
Division L E T I
Département d'Electronique et d'Instrumentation Nucléaire

(*) Travaux réalisés dans le cadre du contrat ESPRIT

Un vocabulaire de 16 mots isolés a été choisi de façon à représenter tous les phonèmes et spécialement ceux qui sont supposés être les plus dépendants du locuteur. Tous les 16 mots commencent et finissent par une voyelle afin de rendre plus aisée la détection des débuts et fins de mot. Ces 16 mots sont :

ADIPEUX	EBOULIS	ILLUSION	OKAPI	UNITE
AGUERRI	EFFACE	IMAGE		USAGE
AMENA	ETUDIA	IMMINENT		
AMINCI	EVASE			
ASSECHE				
AZURE				

Chaque locuteur prononce 10 séries de 16 mots durant la même session. L'ordre préfixé des 16 mots est différent selon les séries.

Le matériel d'enregistrement est constitué d'un microphone et d'un magnétophone portable. La numérisation des bandes analogiques est effectuée sur VAX après filtrage antirepliement. Les conditions d'acquisition sont ; fréquence d'échantillonnage : 10KHz, bande passante : 40Hz --> 4KHz, quantification : 12 bits.

Un logiciel d'acquisition permet d'obtenir des fichiers de signaux ILS (*) par locuteur contenant les 160 mots prononcés. Une compression des fichiers en enlevant les silences entre mots permet une réduction importante de la taille de la base de donnée. Les fichiers signaux compressés ILS ainsi que des fichiers contenant des informations sur chaque mot (position, longueur du mot) sont rangés sur un disque mobile du VAX.

LE PROGRAMME DE TEST AUTOMATIQUE

Le logiciel de test automatique se décompose en plusieurs blocs fonctionnels pouvant être lancés séparément :

- l'initialisation ;
- le prétraitement ;
- l'apprentissage ;
- le traitement ;
- la vérification.

A. INITIALISATION

La partie initialisation permet de fixer les conditions d'utilisation du programme de test sur la base de donnée (nombre de locuteurs, de séries, de mots) ainsi que les conditions de prétraitement d'apprentissage, de traitement et de vérification.

B. PRETRAITEMENT

La partie prétraitement permet d'effectuer l'analyse des différents paramètres de prétraitement sur les fichiers signaux compressés. L'analyse est effectuée sur toutes les fenêtres contigues de 256 échantillons du mot pour tous les paramètres (sauf la fréquence fondamentale).

(*) ILS : progiciel de traitement de signal - Signal Technology Inc.

Ces paramètres sont :

- 10 coefficients de prédiction linéaire (LPC) obtenus soit par la méthode d'autoccorélation avec fenêtre de Hamming (ou sans fenêtre de Hamming pour la distance d'Itakura) soit par la méthode de covariance [1] ;
- 10 coefficients cepstraux avec échelle MEL (MFCC) avec fenêtre de Hamming [2] ;
- 32 coefficients de transformée de Fourier rapide (FFT) avec fenêtre de Hamming ;
- 9 coefficients de passage par zéro du signal (PPZ) [3] ;
- 1 coefficient de fréquence fondamentale (FO) sur des fenêtres de 512 échantillons espacées de 128 échantillons de signal [4].

Les coefficients par mot sont rangés dans un fichier prétraitement (un fichier par locuteur, un enregistrement par mot). Un enregistrement contient un nombre de colonnes égal au nombre de coefficients traités par fenêtre et un nombre de lignes égal au nombre de fenêtres dans le mot.

C. APPRENTISSAGE

La partie apprentissage copie dans un fichier apprentissage (un fichier par locuteur) les enregistrements du fichier prétraitement correspondant à la première série.

D. TRAITEMENT

La partie traitement consiste à calculer les distances intra et inter-locuteur à partir des fichiers prétraitement et apprentissage :

- distances intra-locuteur : pour un locuteur donné, nous comparons le mot de référence de la série 1 avec les mots de tests des 9 autres séries ;
- distances inter-locuteur : pour un locuteur donné, nous comparons le mot de référence de la série 1 avec les mots de tests des 9 autres séries de tous les autres locuteurs.

Pour calculer ces distances deux algorithmes de programmation dynamique sur les coefficients du paramètre choisi ont été essayés :

- un algorithme de parcours simple dérivant de la méthode de BROWN et RABINER [5] et de l'algorithme de Sakoe et Chiba [6] étudié au CEN Saclay ;
- un algorithme de parcours complet de Sakoe et Chiba étudié au LDM (*) [7].

Deux distances locales ont été utilisées :

- une distance Euclidienne
- une distance City-block.

D'autre part la distance d'Itakura a été insérée dans le cas de l'analyse LPC [8].

Les distances obtenues sont rangées dans un fichier distance (un fichier par mot, un enregistrement par locuteur de référence).

E. VERIFICATION

L'algorithme de vérification utilise les fichiers distance en considérant pour un locuteur donné, tous les autres locuteurs présent dans le corpus comme des imposteurs potentiels.

Les seuils utilisés pour la décision de vérification sont des seuils par locuteur (de référence) et par mot. Nous obtenons ainsi, en faisant varier le seuil, des courbes de bonne acceptation et de fausse acceptation dont l'intersection donne le taux d'erreur égal.

Les résultats sont édités sous forme de tables de confusion par locuteur et par mot, de résultats statistiques sur le corpus ainsi que de courbes de taux d'erreur.

Notons que pour une session enregistrée nous avons 4320 comparaisons intra-locuteur et 125280 comparaisons inter-locuteur.

(*) LDM : Laboratoires de Marcoussis

Du fait de la taille du corpus et de l'importance des tests, nous rangeons les résultats intermédiaires, les coefficients de prétraitement et les distances, dans des fichiers ILS sur disque mobile.

Pour tous les programmes, nous gardons une entière compatibilité avec le progiciel de traitement de signal ILS.

RESULTATS DU PROGRAMME DE TEST AUTOMATIQUE

L'ensemble du programme de test a été lancé en premier lieu sur la base JUILLET et les différents paramètres ont été étudiés.

Les taux d'erreur égaux moyen sont présentés tableau I.

Tableau I — Taux d'erreur égal moyen par paramètre

		DTW parcours simplifié Brown - Rabiner/ Sakoe- Chiba	DTW parcours complet Sakoe - Chiba
MFCC	Fenêtre de Hamming	2,9 %	2,8 %
	Covariance Fenêtre de Hamming	13,2 %	9,5 %
LPC	Covariance sans Fenêtre	10,8 %	
	Autocorrélation sans Fenêtre Itakura	6,2 %	
PPZ		14,5 %	
FFT	Fenêtre de Hamming	15,8 %	
FO		19 %	

Il apparaît que les coefficients MFCC conduisent aux meilleurs résultats avec des taux d'erreur voisins pour les deux algorithmes de programmation dynamique.

D'autre part l'utilisation de la distance d'Itakura avec les coefficients LPC donne de meilleurs résultats qu'une simple distance Euclidienne.

Une normalisation en énergie des différents mots avant prétraitement suivie d'une pondération spécifique de chaque coefficient MFCC permet de descendre à un taux d'erreur égal moyen de 1,2%.

Une étude des variations des résultats dans le temps sur 10 locuteurs a été réalisé en utilisant les bases JUILLET, SEPTEMBRE, OCTOBRE, NOVEMBRE. Une permutation de la base de référence est effectuée avec un test sur les 4 bases.

Lorsque base de référence et base de test sont identiques, le taux d'erreur moyen est de 2,5%. Lorsqu'elles sont différentes le taux d'erreur moyen est de 6,6%. Les variations du taux d'erreur sont fonction de l'espacement temporel des bases. Cependant d'autres facteurs liés notamment à l'apprentissage interviennent.

La suite de notre étude s'est donc orientée vers l'utilisation des coefficients MFCC.

LE SYSTEME DE VERIFICATION

En test, bien que la série d'apprentissage soit unique, une moyenne des distances intra-locuteur entre la série 1 et les 9 autres séries pour un locuteur donné sert de point de départ à la variation du seuil afin de trouver le taux d'erreur égal.

Pour le système de vérification les contraintes d'un apprentissage souple et rapide nous ont amenés à effectuer un apprentissage sur 2 séries de 8 mots uniquement. Par conséquent la distance intra-locuteur unique par mot n'est plus assez fiable pour envisager d'en déduire un seuil de décision de vérification. Nous avons donc utilisé une moyenne des distances inter-locuteur par mot comme point de départ du calcul du seuil de décision de vérification.

Le système de vérification comporte une base d'environ 30 locuteurs. Il est composé de 2 blocs fonctionnels distincts :

A. l'apprentissage ;
B. la vérification.

A. APPRENTISSAGE

L'apprentissage comprend :

- a. la constitution d'un fichier locuteur. Ce fichier contient le nom et le sexe des locuteurs. Le sexe du locuteur est demandé uniquement à l'apprentissage pour distinguer les distances inter-locuteur des distances inter-locutrices afin de calculer par la suite un seuil d'acceptation différent pour les femmes et les hommes ;
- b. l'acquisition : 2 séries de 8 mots isolés sont prononcées. Ces 8 mots sont : UNITE, IMAGE, AZURE, AMINCI, ETUDIA, USAGE, AMENA, ADIPEUX ;
- c. le prétraitement : pour chaque mot une analyse sur toutes les fenêtres de 256 échantillons du mot acquis fournit 10 coefficients MFCC par fenêtre. Les coefficients MFCC des mots de la première série d'apprentissage sont rangés dans le fichier apprentissage ;
- d. le traitement intra-locuteur : le système calcule la distance intra-locuteur entre les 2 séries d'apprentissage afin de contrôler la bonne prononciation des mots. Le système peut demander ainsi la reprononciation d'un mot si la distance dépasse un seuil S1 fixé déterminé par une valeur proche de la moyenne des distances intra-locuteur ;
- e. le traitement inter-locuteur : le système calcule les distances inter-locuteur entre la première série du locuteur qui effectue l'apprentissage et les premières séries des locuteurs ayant déjà effectué l'apprentissage.

Si le locuteur a déjà effectué auparavant l'apprentissage, l'ancien fichier apprentissage est effacé et un nouveau fichier apprentissage est créé. Dans les enregistrements concernés du fichier distance toutes les anciennes distances inter-locuteur faisant intervenir le locuteur effectuant l'apprentissage sont remplacées par les nouvelles distances inter-locuteur.

B. VERIFICATION

La vérification comprend :

- a. la détermination des seuils de décision de vérification S2.
Ces seuils S2 par mot sont calculés en début de phase de vérification à partir des valeurs moyennes des distances inter-locuteur ou inter-locutrices lues dans le fichier distance ;
- b. l'acquisition : 1 mot proposé aléatoirement par le système parmi les 8 mots d'apprentissage est prononcé par le locuteur ;
- c. le prétraitement : le système calcule 10 coefficients MFCC sur toutes les fenêtres de 256 échantillons du mot acquis.
- d. le traitement : le système calcule la distance entre les coefficients du mot acquis et ceux du mot d'apprentissage correspondant dans le fichier apprentissage du locuteur à vérifier ;
- e. la stratégie de décision de vérification : elle distingue plusieurs cas :

• acceptation sans adaptation: si la distance est inférieure ou égale au seuil S2 et supérieure à un seuil S3 (S3 S2), le locuteur est accepté ;

• acceptation avec adpatation: si la distance est inférieure ou égale au seuil S3 ; une adaptation de l'apprentissage est effectuée. Le seuil S3 est déterminé de façon qu'un imposteur ayant réussi une fausse acceptation ne puisse pas modifier l'apprentissage du locuteur réel. Pour chaque fenêtre, les nouvelles valeurs des coefficients sont fonction des anciens coefficients MFCC du mot d'apprentissage et des nouveaux coefficients MFCC du mot acquis. Les anciens enregistrements du fichier apprentissage sont donc remplacés par des nouveaux enregistrements.
• refus: si la distance est supérieure au seuil S2 le locuteur est rejeté.

Le locuteur peut recommencer une nouvelle vérification ; le système choisit alors aléatoirement un mot.

RESULTATS DU SYSTEME DE VERIFICATION

Quelques résultats préliminaires d'évaluation du système de vérification sont présentés ici.

Les conditions d'évaluation reposent sur 17 locuteurs (hommes) venant d'effectuer un apprentissage.

Chaque locuteur effectue un test de bonne acceptation et un test de bon refus séparé.

Le test de bonne acceptation consiste pour un locuteur X vrai à entrer son nom véritable et à prononcer 30 mots proposés par le système. Un taux de bonne acceptation ainsi que la distance moyenne intra-locuteur sont donnés par le système.

Le test de bon refus consiste pour un locuteur X imposteur à entrer le nom de tous les autres locuteurs Y et à prononcer 2 mots par locuteur Y. Un taux de fausse acceptation ainsi que la distance moyenne inter-locuteur sont donnés par le système.

Le seuil de décision par mot dépendant de la valeur moyenne des distances inter-locuteur peut cependant être réglé pour favoriser soit l'acceptation de locuteurs vrais soit le refus d'imposteurs. Ce dernier cas a été retenu pour ces résultats préliminaires d'évaluation du système.

Le test de bonne acceptation donne en moyenne sur les 17 locuteurs 90% de bonne acceptation.

Les test de bon refus donne en moyenne sur les locuteurs 3% de fausse acceptation.

Le taux d'erreur égal moyen peut être estimé à 6,5% pour des locuteurs venant d'effectuer un apprentissage.

D'autre part les résultats précédents se détériorent dans le temps lorsqu'un locuteur effectue des tests sans nouvel apprentissage, la procédure d'adaptation n'entrant en action que si la distance est suffisamment inférieure au seuil.

CONCLUSION

Ce papier présente succintement une étude concernant la vérification du locuteur. La méthode retenue est très simple et a été validée par une analyse statistique significative.

Actuellement nos travaux visent à intégrer de nouveaux paramètres discriminants.

REFERENCES ET BIBLIOGRAPHIE

[1] MARKEL J.D. GRAY A.H. : Linear prediction of speech - Springer-Verlag 1976

[2] LOCKWOOD Laboratoires de Marcoussis - Décembre 1981

DAVIS S.B., MERMELSTEIN P. : Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences - IEEE Trans on acoustics, speech, and signal processing - Vol ASSP - 28 n - 4 aout 1980

[3] BAUDRY M. : Etude du signal vocal dans sa représentation amplitude-temps - Thèse 1979

[4] SONDHI M.M. : News methods of pitch extraction IEEE Trans on audio and electroacoustics Vol AU-16 n-2 juin 1968

[5] BROWN M.K., RABINER L.R. : An adaptive, ordered, graph search technique for dynamic time warping for isolated word recognition - IEEE Trans. on acoustics speech, and signal processing -Vol ASSP -30 n-4 p535-544 aout 1982

[6] SAKOE H., CHIBA S. : Dynamic programming algorithm optimization for spoken word recognition IEEE Trans. on acoustics speech, and signal processing -Vol ASSP-26 n-1 p43-49 février 1978

[7] FLOCON B. Laboratoires de Marcoussis - 1986

[8] ITAKURA F. : Minimum prediction residual principle applied to speech recognition - IEEE Trans. on acoustics, speech and signal processing - Vol AASP-23 n-1 p67-72 février 1975
CARATY M.J., RODET X. : Distance interspectrale a criteres perceptifs 14ème JEP p87-90 1985

[9] ATAL B.S. : Automatic Recognition of speakers from their voices - Proceedings of the IEEE - Vol 64 n-4 p460-475 avril 1976

[10] ROSENBERG A.E. : Automatic speaker verification: a review - Proceedings of the IEEE - Vol 64 n-4 p475-487 avril 1976

[11] FURUI S. : Cepstral analysis technique for automatic speaker verification - IEEE Trans. on acoustics, speech and signal processing - Vol ASSP-29 n-2 p254-272 avril 1981

Stratégies de dialogue dans le projet DIRA: exemple de scénario en milieu nucléaire

J. Caelen, M.T. Janot-Giorgetti
Laboratoire de la Communication Parlée - ICP, unité associée au CNRS
INPG/ENSERG
46, Av. F. Viallet
38031 Grenoble Cedex

E. Bauer
Institut Laue Langevin
21, Av. des Martyrs
38042 Grenoble Cedex

Abstract

This paper provides some considerations concerning driving dialogs for robots in nuclear environments. The utilization constraints for robots are task dependent. With respect to such constraints, it is necessary to distinguish between several types of events: (a) an active accident, (b) a steady-state accident, (c) a normal site. In the case of (a), a robot is not recommended, because the reliability of such a system is inadequate. On the other hand, in the case of (b) and (c), vocal driving of robots permits a more comfortable human interface, especially in situations where space and time are limited.

After studying this man/machine dialog problem, it appears that a new architecture has to be defined: the pragmatic module in the ASRS (Automatic Speech Recognition System) and the plan generator must be highly interconnected in order to provide a "test-hypothesis" structure for dialog interpretation. During the driving command dialog, various levels of communication must be enabled: (a) command entry, (b) information entry into a data base (the world of the robot), (c) inquiries about the environment, and (d) responses to robot questions to disambiguate commands. This implies a number of different strategies for recognition: bottom-up (in case (a) with limited syntax and vocabulary), top-down (word spotting in case (d), for instance) and both bottom-up and top-down in other cases. Moreover, the ASRS must take into account the synchronism between task planning and dialoging.

A- INTRODUCTION

Actuellement, les débouchés industriels de la RAP n'existent que lorsqu'il est indispensable de communiquer par la voix, (c'est le cas par exemple pour les handicapés moteurs, les communications à distance, les situations où les canaux de perception de l'humain sont mobilisés par des tâches simultanées, dans un environnement hostile, etc.). Il est donc très important pour le développement des marchés futurs de la RAP:

(a) de dégager les situations véritablement intéressantes,

(b) de ne pas isoler le problème de la communication de son contexte, et de la replacer dans une situation de dialogue homme/machine.

Quelques situations commencent clairement à se dégager à l'heure actuelle avec les performances même limitées de la RAP, et des projets d'envergure sont annoncés:

pour la NASA: pilotage de caméras dans l'espace depuis une plateforme habitée, dans le but de repérer et d'assembler des satellites,

dans le projet ATLAS (Renault): assistance à la conduite,

pour la Défense: pilotage simultané de 5 robots tout terrain avec aide à la décision, etc.

A travers le projet EUREKA la communauté européenne développe des actions concernant le pilotage de robots mobiles avancés. Dans ce projet la communication orale occupe encore une place modeste et n'est pas encore suffisamment intégrée, à notre avis, aux autres moyens de communication.

B- LE PROJET DIRA

Les études les plus récentes montrent que la qualité du dialogue conditionne l'utilisation effective d'un système de communication par la voix. Or ce dialogue repose bien évidemment sur la nature de l'application —consultation de base de données, dictée vocale, etc. Du côté des systèmes RAP on a rarement utilisé les contraintes de nature pragmatique (univers de l'application) dans un système de reconnaissance pour en améliorer la fiabilité. Or ceci est particulièrement intéressant dans le cas de la commande de robots qui évoluent dans un monde limité et connu. Ce sont ces constations qui nous ont amenés à réfléchir puis à proposer le projet DIRA (Dialogue Intégré pour un Robot Avancé). Ce projet consiste à réaliser un système de dialogue pour le pilotage vocal de robots dans un environnement hostile, en intégrant l'univers de l'application dans la RAP.

B-1. les moyens concurrentiels pour la communication Homme/Machine

Des moyens de "communication" plus traditionnels concurrencent la parole pour le pilotage des robots: manettes, claviers, systèmes de pointage, etc. Le principal grief qu'on puisse leur faire est qu'ils mobilisent la main et l'oeil pour transmettre des commandes. De plus les actions sont souvent préprogrammées et donc en nombre limité.

La voix ne présente aucun des inconvénients cités ci-dessus mais les systèmes de reconnaissance (RAP) ne sont pas encore assez fiables (si le vocabulaire est volumineux et la syntaxe peu contrainte) et ils nécessitent des matériels sophistiqués. La solution est donc, à l'heure actuelle, d'harmoniser ces moyens de communication entre eux —en effet il n'est pas prouvé par exemple qu'il soit plus facile d'épeler un nombre chiffre par chiffre, que de l'entrer au clavier.

En tout état de cause, il faut augmenter la fiabilité des systèmes de RAP, et pour cela utiliser de nouvelles sources de connaissances en plus de celles qui sont mises en oeuvre habituellement (phonétique, lexique, syntaxe, sémantique): la redondance de l'environnement et la pragmatique propre au domaine d'application. Il est nécessaire d'intégrer le dialogue au comportement de la machine, et, inversement, de contraindre la RAP par la pragmatique.

B-2. Applications envisageables

Parmi les applications envisageables nous considérons surtout celles qui concernent le milieu nucléaire en raison de leur caractère particulier. On y retrouve les grandes catégories de robot (a) manipulateur, (b) d'inspection, (c) de surveillance.

a. Robot manipulateur

Il est capable de se déplacer et d'effectuer certaines actions mécaniques. Il possède un système de vision et divers capteurs.

b. Robot d'inspection

Sa fonction est surtout d'observer. Pour cela il doit être capable de se déplacer pour modifier son angle de

vision ou pour approcher des capteurs du site à examiner, et parfois de bouger des objets qui le gênent dans son action d'observation. En général la complexité des mouvements est moins grande que pour les robots manipulateurs et leur rayon d'action est plus faible.

c. Robot de surveillance

Ce robot est astreint à scruter en permanence un ou plusieurs capteurs, à analyser la situation et à déclancher des actions ou des contrôles en chaine. Il est souvent statique.

B-3. Exemple de scénario en milieu nucléaire

Le domaine nucléaire est caractérisé par la présence de radioactivité qui peut rendre impossible la présence d'êtres humains dans les zones irradiées. Il est possible d'ailleurs que les inconvénients de la radioactivité se cumulent avec des inconvénients dits "classiques". Le milieu nucléaire se caractérise par:

a. un degré élevé de complexité des installations,

b. des exigences de sécurité à respecter,

c. un enjeu important dans les opérations à effectuer et donc une très grande fiabilité dans les interventions à faire. Il est impératif de considérer toutes les mesures permettant une reconnaissance de la situation ou celles qui rendent possible une intervention adaptée.

B-3.1. Domaine et conditions d'application d'un robot à commande vocale

En régime de fonctionnement normal d'une installation nucléaire, l'utilisation d'un robot est envisageable sachant que l'instrumentation installée et la surveillance prévue couvrent l'ensemble des problèmes pris en compte lors de la conception. Il s'agit essentiellement:

a. de compenser les imperfections éventuelles de surveillance et de conception initiale,

b. d'approfondir la surveillance et de recueillir des résultats de mesures en fonction de régimes de fonctionnement particuliers,

c. d'exécuter des manoeuvres prévues manuellement et rendues impossible par la présence de radioactivité ou qui, par leur caractère répétitif, présentent des inconvénients pour les êtres humains,

d. de mettre en place des dispositifs de surveillance ou de protection pour rendre possibles les interventions humaines

e. d'exécuter des missions d'inspection permettant de se rendre compte de l'état du matériel ou des conditions de fonctionnement.

Un robot à commande vocale présente des avantages évidents lorsqu'il est possible de former des équipes entre humains et machines et dans le cas où la commande vocale donne une rapidité d'action plus grande et un meilleur confort d'intervention du fait que l'humain peut se concentrer sur des tâches essentielles sans s'encombrer des commandes souvent compliquées de la machine.

En situation accidentelle --c'est-à-dire une situation caractérisée par un état de l'installation non prévu à la conception et nécessitant (a) des actions de reconnaissance et de sauvetage en phase active de l'accident, (b) des actions d'investigation et de remise en état en phase stabilisée, (c) des actions d'assainissement en phase post-accidentelle-- l'utilisation d'un robot est conditionnée par l'ampleur des conséquences de l'accident. En effet la zone d'intervention du robot doit être suffisamment rapprochée de la zone de présence humaine pour permettre de tirer profit de la commande vocale.

a- Situation accidentelle active

Une telle situation est caractérisée par une évolution rapide des phénomènes qui nécessite en premier lieu l'identification instantanée de la situation, par exemple par reconnaissance d'objets géométriques, l'évaluation de distances, l'énumération des matériels détruits ou en état de marche, la prise en compte de grandeurs physiques pouvant renseigner sur l'ampleur de l'accident. En outre des priorités doivent être accordées à certaines observations ou phénomènes par rapport à d'autres ou dans leur évolution.

Etant donné le caractère imprévisible et imprévu de l'accident et de ses conséquences, l'utilisation d'un robot en phase accidentelle active paraît difficile.

b- Situation accidentelle stabilisée

Dans cette deuxième phase, les évolutions liées à l'accident sont devenues lentes, on dispose de plus de temps pour intervenir. En conséquence les robots deviennent de plus en plus intéressants pour l'inspection guidée dans le but d'identifier plus précisément les dégats éventuels et pour limiter ces dégats. En général le milieu d'intervention est particuliérement hostile et le nombre de personnes à engager sur le terrain doit être minimisé: dans ce cas les commandes vocales sont particuliérement intéressantes pour décharger au maximum l'attention des opérateurs humains exposés. Ces considérations montrent que seuls les robots d'inspection peuvent être utilisés dans l'état actuel de fiabilité des systèmes de RAP.

c- Situation post-accidentelle

C'est la dernière phase avant le retour à une situation normale; on n'observe plus d'évolution significative des phénomènes liés à l'accident. Il est possible alors d'effectuer les remises en état avec des robots manipulateurs commandés par la voix puisque l'urgence de la situation ne nécessite plus une fiabilité en RAP aussi contraignante.

B-3.2. Spécifications pour un robot à commande vocale

Les considérations énumérées ci-dessus montrent qu'un robot évoluant en milieu nucléaire doit être équipé d'un système permettant d'effectuer des déplacements. Il doit être doté d'instruments de mesure spécifiques ainsi que d'un système de vision. En fonction des besoins de l'intervention il doit pouvoir effectuer des opérations mécaniques. L'ensemble des équipements liés au robot doit présenter un haut degré de fiabilité pour éviter que la défaillance ne conduise à une situation difficile. En conséquence, dans la conception du robot, des moyens et des étapes de replis doivent être prévus pour pallier aux défauts de plannification et aux erreurs induites par le dialogue. Dans un premier temps seules les commandes à faible risque seraient confiées au dialogueur.

C- FONCTIONNEMENT GENERAL DU ROBOT A COMMANDE VOCALE

C-1. Architecture logicielle

Le système de contrôle d'un robot comprend généralement trois modules: le superviseur, le plannificateur et le contrôleur d'exécution. Ces modules sont plus ou moins évolués et sont associés dans des architectures parfois très différentes. Ils peuvent communiquent par l'intermédiaire de la "boite aux lettres" à la manière du "blackboard" en RAP.

Le plannificateur a pour rôle de générer la séquence des instructions de déplacement (ou d'observation), de la prise ou dépose d'objets, et en cas d'échec (but non atteint) de générer une séquence de repli ou un nouveau plan. Il dispose (a) d'une base de connaissances (Univers) et (b) d'une base de faits (Actions) contenant:

(a) une modélisation du monde: description géométrique du robot, des objets, etc., description cinétique et dynamique, liaisons, vitesses, inerties, etc.,

(b) les spécifications des tâches: état initial, intermédiaire, final et conditions d'exécution,

et il est capable de raisonner sur cet univers.

Le superviseur permet de mettre en oeuvre une ou plusieurs stratégies. Ces stratégies peuvent être classées en quatre groupes:

1) non hiérarchique: séquence de tâches

2) hiérarchique: décomposition des tâches en arborescence

3) script-based: adaptation sur des plans connus à l'avance

4) opportuniste: proche du raisonnement humain.

Le contrôleur d'exécution peut quant à lui, offrir trois types de comportement: (a) libre: pas de feed-back en cours d'exécution (seulement à la fin), (b) prudent:

contrôle régulier de l'exécution de la tâche, (c) servile: contrôle en pas à pas.

C-2. Système de pilotage

La fig. 1 donne la structure générale du système de pilotage. Le robot possède des capteurs et éventuellemnt un système de pilotage concurrent via une interface de communication. Le dialogueur désigne ici l'ensemble du système de communication vocale et en particulier la reconnaissance (avec éventuellement la synthèse de la parole).

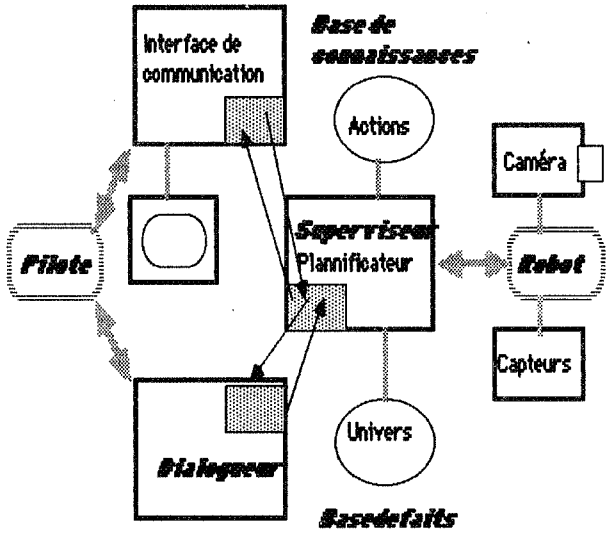


Fig 1: Structure générale du système de pilotage

Dans cette structure le plannificateur et le dialogueur sont en forte interaction --de nature hypothèses-test-- de la manière suivante:

(a) lorsque le plannificateur est en attente de commande (fin d'action précédente) il recherche parmi les hypothèses émises par le dialogueur celles qui sont cohérentes avec son Univers et ses Actions. Il contraint ses hypothèses et en demande le test au dialogueur,

(b) lorsque le plannificateur est en situation d'échec il demande par exemple des précisions au pilote, via le dialogueur, sur la commande inachevée et traite la réponse comme en (a) dans le contexte de l'action en cours,

Après filtrage et test des hypothèses par le dialogueur (1ère passe) plusieurs situations sont encore envisageables:

(1) la pile des hypothèses est vide: le système RAP du dialogueur est en erreur, le plannificateur formule à sa place une hypothèse dont il demande confirmation au dialogueur,

(2) la pile ne contient qu'une hypothèse --c'est le cas le plus simple-- elle est prise en compte par le plannificateur et traitée conformément à son Univers,

(3) la pile contient plusieurs hypothèses, le plannificateur choisit la plus conforme au plan qu'il a prédit.

À cette combinatoire, s'ajoutent les contrôles de tâches qui peuvent correspondre à des interruptions ou reprises d'action selon que l'on se trouve en mode libre, prudent ou servile.

Une hypothèse est généralement une séquence de mots (ou phrases) émise par le dialogueur après reconnaissance et un test est la vérification de séquences de mots sur le signal vocal à travers le dialogueur.

De manière plus détaillée, le graphe PERT de la fig. 2 montre l'ordonnancement des actions du robot dans le cadre du dialogue de pilotage. Chaque action a un "contexte" c'est-à-dire une mémoire des actions passées sous forme d'un historique. Il est évident que ce contexte est plus ou moins contraignant selon le type d'actions exécutées.

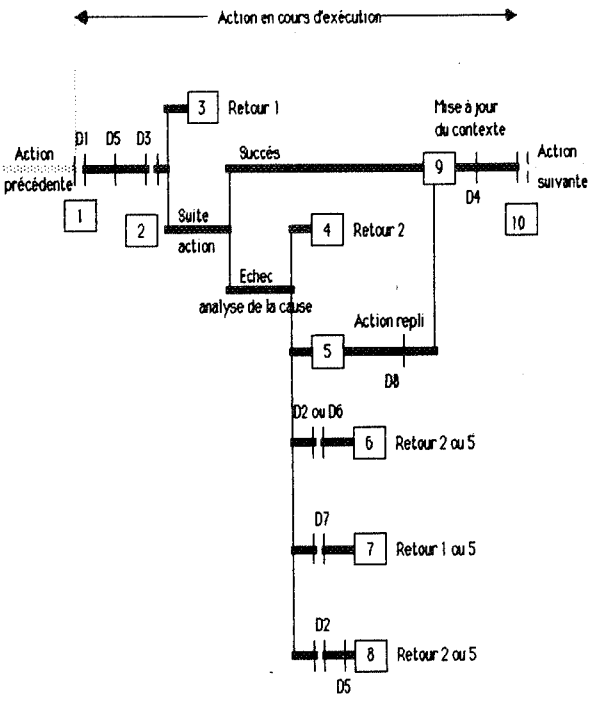


Fig. 2: Ordonnancement des actions du robot

Les événements qui marquent le déroulement d'une action sont: [1] début, entrée d'une commande compatible avec le contexte de l'action précédente, [2] poursuite de l'action après contrôle dans le cas d'une action servile, [3] reprise de l'action après aiguillage sur une fausse solution, [4] échec, but non atteint, cause=erreur de plannification bien que la commande ait été bien interprétée, [5] échec, seule solution possible=repli, [6] échec, cause=commande imprécise ou contradictoire avec l'état de l'univers (apparition d'un fait nouveau, comme un obstacle par ex.), [7] échec, cause=commande non exécutable par manque de connaissances sur les actions ou parce que celles-ci sont mal décrites, [8] échec, cause=commande ambigüe, plusieurs actions sont solutions et le contexte n'est pas suffisamment contraignant pour permettre un choix parmi les solutions multiples, [9] succés de l'action commandée ou de l'action de repli, [10] fin de l'action en cours, mémorisation de son historique.

À ces événements qui provoquent des interruptions de programme, s'ajoutent les contrôles et vérifications diverses qui peuvent s'exécuter au cours de la plannification sans aucune interruption pour le robot. Ces contrôles ne sont pas tous matérialisés sur la fig. 2.

C-3. Les niveaux de langage dans le dialogue

À partir de la fig. 2 il est facile de préciser ce que sont les objectifs des dialogues pour guider le robot:

D0: initialisation du robot, entrée de données sur l'univers du robot par la voix: description du monde, des objets, des actions possibles, des actions de repli, etc., pour constituer la base de faits et la base de connaissances du robot. Notons que ceci peut aussi être fait par des moyens plus traditionnels sauf peut-être pour ajouter des informations de dernière minute dans la base de faits (selon l'état des capteurs par exemple).

D1: entrée d'une commande en langage naturel, la syntaxe et le vocabulaire doivent être réduits pour augmenter la fiabilité du système RAP. Il peut y avoir adaptation au locuteur avant l'expérimentation.

D2: dialogue dirigé par le plannificateur pour obtenir une information complémentaire sur une commande imprécise. La réponse peut être un mot isolé.

D3: contrôle d'exécution, c'est une commande avec attente de réponse (affichée ou orale).

D4: contrôle de type trace pour éventuellement modifier le contexte pour l'action suivante.

D5: questions pour vérifier que la commande a été bien comprise. Elles peuvent être émises indifféremment par le pilote ou le robot.

D6: entrée d'informations pour modifier la base de faits au fur et à mesure de l'apparition d'événements nouveaux. Ce mode de dialogue se rapproche du mode dictée.

D7: entrée d'informations pour modifier la base de connaissances (idem D6).

Il est bien évident que ces dialogues peuvent être menés de façons très différentes les unes des autres, allant du mode libre au mode dirigé. Cependant on peut définir quatre niveaux de langage plus ou moins étendus. Ce sont:

(a) Langage d'entrée de données dans la base de faits:
- pour l'initialiser: en décrivant l'univers visible (positions des objets, tailles, etc.) et l'univers supposé (valeurs peu sûres ou variables accidentelles, etc.)
- pour la modifier en fonction de l'évolution de l'univers (obstacles nouveaux, changement d'univers, etc.)

Le langage adéquat pour cette tâche peut être syntaxiquement contraint mais il doit offrir un vocabulaire suffisamment étendu.

(b) Langage d'entrée de données dans la base de connaissance:
Les objectifs à atteindre sont les mêmes que pour la base de faits mais ici les contraintes pragmatiques sont plus fortes puisque les données sont attendues par le système (degrés de liberté du robot, inertie, vitesse, etc.): cela autorise donc une syntaxe plus lâche que précédemment puisqu'on peut utiliser la redondance sémantique.

(c) Langage de commande:
Pour être ergonomique, il doit être proche du langage naturel et admettre des phrases parfois a-syntaxiques (surtout en situation de stress). Par contre le vocabulaire des mots-clefs de commande doit être assez restreint.

(d) Langage d'interrogation:
Il est du même type que (c). Pour les réponses, un simple mot isolé peut parfois suffire.

En conclusion on peut dire qu'il y a plusieurs niveaux de complexité dans les langages de communication avec le robot. Con peut les classer selon l'étendue du vocabulaire et les contraintes syntaxiques. Notons que la fiabilité maximum demandée au système se situe au niveau (c) et dans les réponses du pilote au niveau (d).

C-3. Implications pour la structure de contrôle du dialogueur

Les différents modes de dialogue et niveaux de langage (de la vérification du mot isolé à la reconnaissance de la parole continue) font qu'il est impossible d'avoir une structure de contrôle figée: elle doit tantôt être entiérement descendante ("word spotting" dans le cas de réponses à des demandes du planificateur), tantôt guidée par la pragmatique lorsque le contexte d'exécution de l'action est très contraignant, tantôt guidée par la syntaxe et le lexique si ce contexte est faiblement contraignant. Dans tout les cas on peut avoir un décodage acoustico-phonétique robuste ascendant avec affinage descendant conditionnel.

CONCLUSION

Ces premières réflexions sur le projet DIRA permettent de faire quelques remarques pour la réalisation du système de dialogue avec le robot.

1. une utilisation en milieu nucléaire est tout à fait envisageable à l'heure actuelle, du moins au vu des spécifications pour le dialogue. Un robot d'inspection à commande vocale apporterait le confort d'intervention indispensable en phase accidentelle stabilisée, et bien sûr dans une situation normale.

2. l'approche utilisée qui consiste à intégrer les contraintes du robot dans les mécanismes de compréhension doit permettre d'augmenter les taux de reconnaissance. En contrepartie le synchronisme dialogueur/générateur de plan nécessite un temps de réponse assez court du système RAP.

3. les divers types de dialogue qui vont du mot isolé à la parole continue et les différents niveaux de langage font qu'une structure de contrôle dynamique doit être choisie pour la RAP. La stratégie de reconnaissance doit dépendre du contexte de l'action en cours et des actions précédentes.

BIBLIOGRAPHIE

[Christiansen, 85] A.D. Christiansen
The history of planning methodology: an annoted bibliography. SIGART Newsletter, Oct. 85, n° 94.

[Farreny, 80] H. Farreny
Un système pour l'expression et la résolution de problèmes orienté vers le contrôle de robots. Thèse d'état, UPS, Toulouse, 1980.

[Guyomard et al, 84] M. Guyomard, J. Siroux, L. Trilling
Le dialogueur: un intermédiaire entre l'utilisateur et l'application. Actes du séminaire "Dialogue homme-machine à composante orale", GRECO-GALF, 1984, pp. 404-416

[Hayes-Roth et al, 79] B. Hayes-Roth, S. Rosenschein, S. Camara
Modeling planning as an incremental opportunistic process. Proc. 6th IJCAI, 1979.

[Lee et al, 83] M.H. Lee, D.P. Barnes, N.W. Hardy
Knowledge based error recovery in industrial robots. Proc. IJCAI, 1983.

[Picardat, 85] J.F. Picardat
Ebauche d'un modèle généralisé pour le contrôle de plan. LSI, rapport interne n° 216, Toulouse, 1985

[Sabah et al, 84] G. Sabah, J.B. Berthelin, A. Vilnat
Le dialogue dans un système de questions réponses. Actes du séminaire "Dialogue homme-machine à composante orale", GRECO-GALF, 1984, pp. 305-330.

[Sacerdoti, 77] E. Sacerdotti
A structure for plans and behavior. Elsevier, 1977.

[Tarnaud, 84] P. Tarnaud
Intégration du dialogue vocal dans un environnement réel complexe. Actes du séminaire "Dialogue homme-machine à composante orale", GRECO-GALF, 1984, pp. 345-363

[Taylor, 83] R.H. Taylor
An integrated robot system architecture manufacturing research department. IBM T.J. Watson Research Center, Proc. IEEE, 1983.

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE DANS L'EXPLORATION DU CHAMP VISUEL A TESTS MULTI-STIMULI

DIAF M.

INGM 25/B2 35 000 - BOUMERDES - ALGERIE

ABSTRACT

The aim of this paper is to show how the Automatic Speech Recognition constitues an effescient solution to the problem of automatic recording of subject's responses submitted to a multi-stimuli visuel field examination. Hospital experiments have particularly permitted us to appreciate the fonctional mode of the system and to analyse the reactions of subjects and staff not accostemed to this type of material.

INTRODUCTION

Si les possibilités d'utilisation de la parole comme moyen de communication entre l'homme et la machine sont trés nombreuses, l'emploi des systèmes de reconnaissance automatique de la parole (RAP) reste jusqu'à nos jours limité. A cause du manque de précision et du prix encore assez élevé de ces systèmes, les utilisateurs ont tendance à attendre ce que l'avenir va apporter. En effet, plusieurs laboratoires spécialisés à travers le monde continuent à investir pour l'amélioration de cette précision, l'indépendance du locuteur et la reconnaissance d'énoncés continus. En ce qui nous concerne, nous nous sommes placés en tant qu'utilisateurs et c'est dans un milieu médical et plus particulièrement en ophtalmologie lors de l'exporation du champ visuel utilisant des tests multi-stimuli que nous avons introduit la RAP/1/. Rappelons que le but d'un tel examen est de détecter les déficits du champ visuel qui sont dûs à la détérioration partielle de la rétine de l'oeil atteint par exemple , d'un glaucome/2/. Le principe de cet examen est de présenter d'une façon pseudo-aléatoire plusieurs combinaisons composées de deux ou trois ou quatre stimuli affichés puis éteints simultanément sur un écran piloté par un micro-calculateur. Ces points doivent être répartis de sorte qu'ils couvrent d'une façon uniforme, la totalité de la rétine. Il est évident que ceux qui " tomberont " dans la zone défectueuse de la rétine seront non perçus par le sujet. Il existe plusieurs appareils basés sur ce principe. Selon leur mode de fonctionnement, on peut les regrouper en deux classes. Dans la première, on trouvera les appareils manuels par l'intermédiaire desquels l'ophtalmologiste propose des tests au sujet. Celui-ci répond verbalement et l'ophtalmologiste porte manuellement les points non perçus sur une feuille préparée à cet effet/3/. Dans la deuxième classe, on trouvera les appareils semi-automatiques dans lesquels se font automatiquement la présentation des tests, le traitement et l'impression des résultats. Quant à l'acquisition des réponses du sujet, elle se fait par l'intermédiaire d'un opérateur spécialiste du champ visuel /4/.

NECESSITE D'UTILISER LA R.A.P.

Dans les deux modes de fonctionnement précédents les résultats de l'examen sont considérablement soumis à l'influence de l'opérateur. Cette influence est d'autant plus remarquable que le nombre de sujets à examiner est élevé. Pour éviter cette tâche monotone à l'opérateur et éviter en même temps son influence sur la qualité des résultats de l'examen, nous proposons de rendre complètement automatique le mode de fonctionnement du système. Dans ce cas, l'échange d'informations ne s'établit plus qu'entre le sujet et le système. Du côté du système, ces informations sont constituées des tests multi-stimuli présentés sur un écran , alors que du côté du sujet, les informations transmises sont ses réponses qui peuvent être zéro, un, deux, trois ou quatre. Il reste cependant à définir le moyen pour le sujet de communiquer automatiquement ses réponses au système et ceci en tenant compte des contraintes liées à l'examen du champ visuel. Ce type d'examen exige que le sujet porte absolument toute son attention sur sa tâche principale qui est de guetter l'apparition des tests. De plus, il doit garder l'oeil à examiner immobile et le regard dirigé vers le centre de l'écran. On peut donc immaginer différents dispositifs permettant au sujet de répondre manuellement. Cependant, pour respecter les exigences citées ci-dessus, nous proposons l'utilisation d'un système de RAP (figure 1).

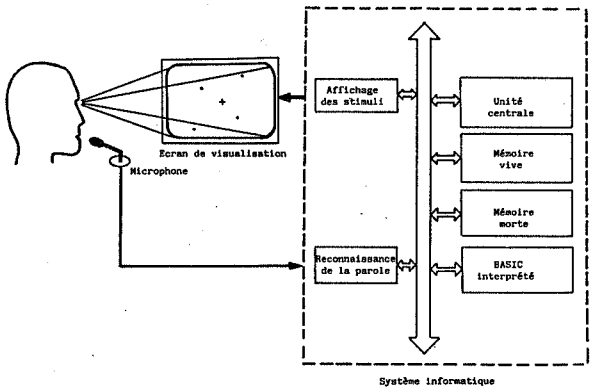


Figure 1. Shéma synoptique du système

CHOIX DU SYSTEME DE R.A.P.

La première opération à effectuer dans une application de la RAP est le choix du système parmi ceux qui existent sur le marché, sachant qu'ils sont nombreux et de performances trés comparables/5/. Nous verrons en fin de ce paragraphe le type de système qui conviendrait, le mieux, pour notre application. Dans notre travail de laboratoire, nous avons effectué les essais à l'aide d'une carte de reconnaissance de bas de gamme (speach-lab 20A-32,Heuristics). Nous avons réévalué le taux de reconnaissance de cette carte pour les chiffres 0,1,2,3 et 4 répétés cent fois chacun par différentes personnes et pour différentes langues (tableaux 1). Au cours de ces expériences, les locuteurs avaient tendance à se relacher à force de répéter plusieurs fois ces chiffres. Ce relachement d'origine psycho-physiologique qui apparait de la même façon chez les sujets lors de l'examen du champ visuel, entraîne des confusions parfois assez importante entre les chiffres. Nous avons remarqué particulièrement que "quatre" a souvent été confondu avec "un" par le système de la

RAP que nous avons utilisé. Ceci est dû au fait que "un" est prononcé un peu comme la lettre "A" et "quatre" comme "kAt". Lorsque les lettres "k" et "t" sont"avalées", le système de RAP comprendra "un". Dans le sud de la France, par exemple, où l'accent est particulier, la confusion entre ces chiffres serait différente. On peut penser à améliorer le taux de reconnaissance avec le choix du vocabulaire. Seulement, le fait d'utiliser un autre mot pour désigner le nombre de stimuli perçus au lieu du chiffre correspondant, compliquerait la tâche du sujet. On peut penser aussi à utiliser un système de RAP plus performant et dôté d'une procédure de rejet permettant une meilleure discrimination entre les chiffres. Ce type de système ne conviendra parfaitement que si, au cours de l'examen, le sujet ne modifie pas trop sa voix et son articulation. Dans le cas contraire, l'examen devient lourd et long puisque le sujet sera appelé à répéter sa réponse à chaque fois que celle-ci est mal comprise. Durant l'examen proprement dit, nous avons relevé plusieurs réponses subjectives telles que " je crois avoir vu deux points" ou " je vois trois points, non, non quatre!" etc... Devant de telles situations, les systèmes de RAP de mots isolés deviennent impuissants. Pour cette application, nous suggérons donc d'étudier et réaliser un système spécialement pour reconnaître les seuls mots zéro,un,deux, trois,quatre, oui et non. L'idéal serait encore, que le système soit multi-locuteur avec un trés haut taux de reconnaissance. De plus, ce système devra présenter un certain temps d'écoute pendant lequel il ne détectera que les mots utiles même si la réponse du sujet est subjective telle que " je crois voir trois points". Dans ce type de réponse, le système devra détecter uniquement le chiffre "trois". Dans le cas de plusieurs chiffres dans la réponse, ce sera le dernier qui sera pris en compte.

ESSAIS CLINIQUES

Dés la mise sous tension du matériel comprenant le système d'exploration du champ visuel et le système de RAP adapté à ce premier, le sujet est invité à subir la phase d'apprentissage dans sa langue d'origine. Par voie d'affichage sur l'écran, on lui demande donc de répéter deux ou trois fois les chiffres de zéro à quatre et les mots oui et non. Aprés cet apprentissage, les directives pour le bon déroulement de l'examen lui sont données par affichage. Au cours des différents examens effectués, un opérateur était là présent pour apporter aide aux personnes d'un niveau socio-culturel relativement bas. Le sujet répondra par "oui" ou par " non " aux questions " avez-vous compris? " et " êtes-vous prêt?" qui lui sont posées par écrit sur l'écran cathodique. Dans le cas positif, commence l'examen proprement dit, c'est à dire la présentation des tests multi-stimuli. Le sujet répond dans le micro par le nombre de points perçus. Le microcalculateur sur lequel est programmée la stratégie d'examen, traite les résultats automatiquement et les sort sur écran et sur imprimante.

Réaction du personnel hospitalier:

Même au stade de développement actuel, le matériel a été apprécié par le personnel puisque son fonctionnement ne nécessite aucune attention soutenue de sa part.

Réaction des sujets:

Les sujets s'installent devant l'appareil avec une bonne volonté car l'exploitation des progrès technologiques et informatiques à des fins médicales les rassure en ce qui concerne la précision des résultats des examens et des diagnostics. En général, l'ensemble des sujets que nous avons examinés étaient coopératifs et attentifs. Cependant, au début de l'examen, lors de la lecture des consignes générales, il semble que le message écrit ne soit pas bien perçu. Souvent l'opérateur doit expliquer oralement ce qui est écrit sur l'écran en reprenant, la plupart du temps, les mêmes termes. Il apparait donc nécessaire d'ajouter à notre configuration un synthétiseur de parole ou un système d'enregistrement de la voix qui permettrait de donner des consignes sous une forme plus assimilable. Ensuite, la phase d'apprentissage trouble certains sujets qui n'en voient pas immédiatement la finalité. A ce niveau, les directives de cette phase transmises sous forme écrite ne sont pas toujours correctement assimilées. Par exemple, lorsque certains sujets lisent " dites un ", ils répondent " dites un ". Un message oral permettrait d'éviter ce type de problème.

D'exellents résultats ont été obtenus de manière entièrement automatique avec des sujets familiers du monde informatique. Ces sujets articulent les réponses de la même façon que lors de la phase d'apprentissage. Cependant, du fait que nous nous sommes adressés à des sujets de défférents âges et issus de différentes classes socio-professionnelles, la présence d'un opérateur s'était avérée nécessaire.

Nous avons examiné des sujets normaux et des sujets pathologiques. Pour tous les sujets saints et coopératifs, la tâche aveugle a été restituée. Les déficits des sujets glaucomateux ont été mis en évidence d'une manière satisfaisante. Nous avons en effet relevé plusieurs scotomes et des tâches aveugles atteintes . Parmi l'ensemble des résultats obtenus, certains restent difficiles à interpréter et peuvent, par conséquent, être considérés comme érronés. Ce type de résultat serait surtout dû à la non coopération du sujet ou à sa prononciation qui introduit des confusions entre les chiffres.

CONCLUSION

L'application de la RAP a permis de faciliter la tâche de l'opérateur et d'éviter son influence sur la qualité des résultats. Le choix du système de RAP à utiliser pour cette application doit être convenablement fait. Les remarques que nous avons faites lors des essais cliniques peuvent aider pour faire ce choix. Rappelons que le produit d'examen du champ visuel équipé de la RAP peut être diffusé en médecine du travail, dans les cabinets ophtalmologiques et surtout dans les centres de santé et les pays en développement. Une version compacte et commercialisable est actuellement en fin de réalisation en collaboration avec le service d'ophtalmologie de l'hôpital Mustapha à Alger.

REFERENCES

1-M. DIAF et al. - " Dispositif automatique adapté au dépistage de masse de glaucome." - 3ème congrès panafricain d'ophtalmologie. Alger-Nov. 1984

2 - ETIENNE R. - " Les glaucomes." - Diffusion Générale de Librairie.1969.

3- C.AARON - "Essai comparatif portant sur 112 yeux de quatre périmètres automatique et deux manuels. " - Thèse de médecine, Université P-M-Curie-Paris VI. 1983.

4- GREVE EL. " Peritest " - Doc.Oph. - Proc. Séries 22-71-74.

5-1.C.S- " 430 voice input/ouput. The state of the art ".

AGE : 45 ans
SEXE : Masculin
LANGUE : Français
ARTICULATION : NORMALE

fréquence d'affichage des chiffres : 97 - 104 - 98 - 103 - 98
fréquence d'assignation de la carte: 111 - 168 - 40 - 93 - 84

I \ J	0	1	2	3	4
0	100	0	0	0	0
1	0	100	0	0	0
2	7	52	40	0	1
3	4	0	1	95	0
4	0	15	0	0	85

% de reconnaissance - 0 - 100 %
% de reconnaissance - 1 - 100 %
% de reconnaissance - 2 - 40 %
% de reconnaissance - 3 - 95 %
% de reconnaissance - 4 - 85 %

La corrélation entre le chiffre I et le chiffre J se lit en % dans le tableau ci-dessus.

Moyenne de reconnaissance de la carte : 84 %

Russe

I \ J	0	1	2	3	4
0	100	0	0	0	0
1	3	95	0	2	0
2	0	0	100	0	0
3	6	0	0	85	9
4	7	1	0	48	40

Moyenne - 86 %

Espagnol

I \ J	0	1	2	3	4
0	100	0	0	0	0
1	0	32	4		64
2	0	0	64	36	0
3	0	0	0	100	
4	0	3	4		93

Précision - 79 %

Marocain

I \ J	0	1	2	3	4
0	100	0	0	0	0
1	0	89	11	0	0
2	0	29	64	7	0
3	0	0	0	100	0
4	4	0	0	22	74

Précision - 85 %

Arabe algérien

I \ J	0	1	2	3	4
0	93	0	7	0	0
1	1	80	0	19	0
2	2	0	98	0	0
3	9	0	0	91	0
4	0	12	0	0	88

Précision - 90 %

Anglais

I \ J	0	1	2	3	4
0	100	0	0	0	0
1	0	100	0	0	0
2	0	0	100	0	0
3	0	0	0	100	0
4	2	3	0	0	95

Précision - 99 %

Kabyle

I \ J	0	1	2	3	4
0	97	0	3	0	0
1	0	100	0	0	0
2	2	0	98	0	0
3	3	0	0	88	9
4	0	0	0	12	88

Précision - 94 %

TABLEAUX I : Réévaluation de la carte Speechlab pour les chiffres 0, 1, 2, 3, 4 répétés 500 fois par des sujets dans différentes langues.

a) Dans ces tableaux, **I** représente le chiffre affiché aléatoirement sur l'écran. **J** est le chiffre répété par le sujet.

b) La valeur indiquée sous le tableau est le taux de reconnaissance moyen dans la langue correspondante.

CONSTITUTION INCREMENTALE D'UN CORPUS DE DIALOGUES ORAUX COOPERATIFS *

GUYOMARD M. SIROUX J.

IRISA-IUT DE LANNION. BP 150. F-22302 LANNION CEDEX
CNET LAA/TSS/RCP. BP 40. F-22301 LANNION CEDEX

ABSTRACT

With the aim of achieving a system for oral interrogation of a telephone information type application, we have carried out experimental work on dialogue between subjects and a simulated machine. The experimental work was carried out in two phases in order to yield a definition of the dialogue making it possible to undertake its implementation. In the first phase we used both a highly constrained dialogue and an unrestricted dialogue. The second phase was based on the conclusions drawn from the previous dialogues and on a strategy for cooperation. In this paper we shall describe and analyse both phases of the experimental work. Particular attention is paid to the dialogue strategy together with an examination of the corpus of the dialogues collected. Some linguistic elements are also given. From a more general point of view, we provide the first indications of what we have learned concerning design of future systems.

0. INTRODUCTION

Dans le but de réaliser un système d'interrogation orale d'une application de type renseignement téléphonique (pages jaunes de l'annuaire), nous avons été amenés à définir de manière fine les composants du système.

Cette définition est complexe: d'une part de nombreux paramètres interviennent, et d'autre part elle ne peut pas se fonder sur des a priori et des hypothèses non vérifiées (SIROUX,86). De plus, s'il existe d'assez nombreux travaux sur le dialogue (homme-homme, homme-machine), les résultats actuels ne sont pas directement exploitables: ou bien ils sont trop éloignés de la problématique de la mise en oeuvre informatique (ROULET,81), ou bien ils sont trop liés à une application particulière dont tous les paramètres n'ont pas été clairement étudiés (GRECO,85), (SIROUX,85).

Ainsi nous avons décidé de faire reposer la réalisation du système sur une expérimentation nous permettant de :

- vérifier que les contraintes issues des limitations imposées par les systèmes de reconnaissance automatique de la parole (RAP) ne constituent pas un obstacle insurmontable pour la viabilité du système. Nous faisons en particulier allusion à la taille du vocabulaire accepté et à la longueur des phrases prononcées.
- répertorier les contraintes ignorées a priori, ayant pour origine le comportement linguistique et stratégique des sujets, en cherchant à distinguer les contraintes incontournables de celles qui, moyennant la fourniture de quelques instructions simples à l'usager, peuvent être surmontées.
- spécifier le futur système, en tenant compte des points précédents afin de produire :
  - le vocabulaire du système de RAP,
  - les constructions syntaxiques,
  - la description du dialogue oral que le système sera capable de gérer.

* Ce travail a été réalisé dans le cadre du projet ESPRIT P1015 "Palabre" partiellement financé par la CEE.

C'est cette expérimentation qui fait l'objet de ce papier. Après une présentation de la méthodologie utilisée (approche en deux phases), nous résumons les principales étapes de l'expérimentation en insistant particulièrement sur l'aspect gestion du dialogue et en donnant les principaux résultats des analyses sur les corpus recueillis. En conclusion nous mettons en évidence l'apport de ce travail sur la conception et la mise en oeuvre du système en cours de réalisation.

1. METHODOLOGIE

Le service de renseignements oral pages jaunes qui aurait pu servir de base (malgré tout incomplète) d'observation n'existant pas, nous avons monté de toutes pièces notre propre expérimentation. Afin d'éviter d'imposer a priori des contraintes que l'on serait dans l'impossibilité de respecter, nous avons décidé de mener l'expérimentation en 2 phases successives, les enseignements tirés de la première devant servir de base à la mise en oeuvre de la seconde. La première phase a elle-même été divisée en 2 parties recouvrant le spectre des types de dialogues possibles :

- d'une part un dialogue très dirigé,
- d'autre part un dialogue complètement libre,

et ceci dans le but de recenser :

- les contraintes qui ne seront pas respectées par la majorité des utilisateurs,
- les fonctions de dialogue qui sont spontanément utilisées par l'usager typique.

Un compromis judicieux peut alors conduire à un dialogue acceptable tant du point de vue de l'utilisateur que du point de vue de l'état de l'art en RAP et en gestion de dialogue. En abrégé, l'expérimentation menée peut se définir comme l'enregistrement et l'analyse de dialogues finalisés contrôlés entre un compère (jouant le rôle de la machine) et des sujets volontaires croyant participer à l'évaluation ergonomique d'un système opérationnel.

Le montage matériel: il est conçu pour donner au sujet l'impression qu'il converse avec une machine. Pour cela la voix du compère est déformée par son passage à travers 2 vocodeurs. D'autre part, afin de s'approcher des conditions de reconnaissance de parole, les conditions d'écoute du compère sont détériorées à l'aide d'un générateur de bruit.

L'application et la documentation: l'application retenue est la consultation des Pages Jaunes de l'annuaire. Elle permet au sujet d'obtenir des informations (nom, adresse, numéro de téléphone, localité) sur les professionnels en fournissant un certain nombre de données (paramètres). Les limitations de la RAP nous ont astreints à réduire l'application visée : au total 98 abonnés professionnels sont référencés sous 24 rubriques (domaine automobile) sur la région de Lannion.

Les sujets: ils sont choisis parmi des étudiants (non familiarisés avec l'informatique) volontaires et rémunérés. Au total 30 sujets ont été retenus, 15 pour chaque phase d'expérimentation.

Les scénarios: le sujet se met (artificiellement) en situation de requérir les coordonnées d'un professionnel en instanciant un scénario parmi 41 possibles. A titre d'exemple voici un des scénarios tels qu'ils sont proposés aux sujets: " Pour votre départ en vacances vous recherchez une baladeuse qui se branche sur l'allume-cigare...".

Le déroulement d'une session: le sujet assimile les consignes expérimentales relatives à la phase en cours puis il instancie son premier scénario. Il se met alors en communication téléphonique avec le "système" de renseignements. La conversation est enregistrée. A l'issue de la session le sujet complète un questionnaire portant sur les informations demandées et obtenues. Chaque sujet déroule 5 scénarios, certains pouvant être enchaînés dans une même session. La séance s'achève par un entretien pendant lequel le sujet évalue le "système". En cas de problème et pour les opérations de service (explications,...) une personne ("l'aide") est toujours présente aux côtés des sujets.

2. PREMIERE PHASE DE L'EXPERIMENTATION

Deux types de dialogues ont été spécifiés, mis en oeuvre et analysés: un dialogue très dirigé et un dialogue libre. Pour chacun d'eux les sujets ont été répartis en deux groupes selon le contenu des scénarios proposés: les sujets familiers de la région (9) et les autres (6).

2.1 Les protocoles.

Le protocole du dialogue dirigé: le dialogue dirigé est formalisé par un automate d'états finis. Dans son principe ce dialogue est assez semblable à celui décrit par Van Katwijk et al. (VAN KATWIJK,79). On peut le caractériser en quatre points:

- Le "système" n'accepte que des réponses strictes de la part du sujet.
- De son côté le système répond strictement à la requête formulée par le sujet.
- Certains énoncés affirmatifs du "système" sont suivis d'une courte pause qui donne au sujet la possibilité d'intervenir.
- Quelques commandes simples ("répétez", "fin", "précisez") permettent au sujet de contrôler partiellement le déroulement du dialogue.

Le protocole du dialogue libre: les consignes sont ici plus floues: le compère doit plutôt subir le dialogue tout en restant coopératif. Notons que le compère employé à cette occasion n'est pas un professionnel des renseignements téléphoniques.

2.2 Dialogues dirigés. Analyses et résultats.

Les 5 sujets retenus ont permis d'enregistrer 23 conversations. A l'exception d'un sujet, maîtrisant d'emblée ce type de dialogue, on n'obtient pas de réponses strictes aux questions oui/non du compère. Le cas est particulièrement frappant dans les réponses aux questions du type: "voulez-vous effectuer une autre demande?".

D'autre part, les sujets ont souvent tendance à prendre l'initiative de la conversation; il s'en suit soit un "dialogue de sourds" soit un sous-dialogue de rattrapage très désagréable pour le sujet. Parmi les fonctions de dialogue spontanément utilisées par les sujets mais non prévues ou envisagées dans un contexte différent, on peut citer:

- la demande d'épellation,
- l'introduction d'une nouvelle requête,
- la demande d'attributs précis (adresse,...).

Notons enfin, de la part des sujets, une faculté d'adaptation que l'on voit se manifester entre le premier et le cinquième dialogue de la session.

Le vocabulaire utilisé par l'ensemble des sujets comprend 131 termes.

2.3 Dialogues libres. Analyses et résultats.

Les 10 sujets retenus ont permis d'obtenir 48 conversations. Nous avons analysé le corpus ainsi recueilli selon trois axes: les fonctions de dialogue, l'aspect linguistique, le comportement des interlocuteurs.

Les faits saillants de ces analyses peuvent se résumer ainsi: pour la gestion du dialogue les enseignements portent surtout sur la phase négociation de réponse d'une requête pour laquelle plusieurs fonctions ont été mises en évidence : l'annonce de généralités, l'annonce d'une réponse élémentaire, la sélection, les demandes de précision et les modifications de requête. L'analyse de la gestion phatique a permis de recenser des demandes de répétition (issues des sujets) qui portent en général sur un élément très informatif, des reprises de propos, des demandes d'épellation.

Sur le plan linguistique, nous avons surtout étudié les énoncés introducteurs de requête. Quatre formes d'énoncés ont été répertoriés (déprécatif, interrogatif, elliptique, impératif). Si certains énoncés sont très longs et complexes, une structure linguistique permettant d' exprimer la demande a cependant pu être dégagée. Le vocabulaire propre à l'usager dans les dialogues libres comprend 353 termes. La plupart des énoncés des sujets sont entachés d'hésitations, de faux départs, de reprises, d'auto-corrections. Ces phénomènes font actuellement l'objet d'une étude en vue d'une modélisation particulière.

L'étude du comportement des interlocuteurs confirme le manque de coopérativité du compère; en outre le ton péremptoire de ses répliques n'engage pas à la poursuite du dialogue. Sa stratégie se révèle par ailleurs très stable par rapport aux pertubations et aux incohérences qui se manifestent pendant le déroulement des dialogues.

2.4 Conclusions sur la première phase.

Un certain nombre de points nous semblent importants d'être soulignés à l'issue de la première phase:

Concernant les dialogues dirigés:

- Le dialogue dirigé tel qu'il est conçu n'est pas viable pour un utilisateur occasionnel.
- A l'instar de Van Katwijk nous tenons pour indispensable la prise en compte des réponses indirectes de l'utilisateur.
- Il est souhaitable d'intégrer pour la phase 2 les déviations fréquemment constatées (épellation, prise d'initiative contrôlée).

Concernant les dialogues libres:

- Les phénomènes linguistiques de l'oral spontané ne peuvent être évités. Il est important de se préoccuper de leur modélisation dès à présent.
- La taille du vocabulaire recueilli est raisonnable. Elle permettrait même d'envisager une extension mesurée de l'application.
- Le comportement du "système" doit impérativement être rendu plus coopératif. Dans la gestion de la succession des énoncés il est indispensable de prévoir les instants d'interventions possibles de l'utilisateur et de traiter de manière adéquate les non-interventions (réponse implicite : "qui ne dit mot consent").
- La structure des dialogues est suffisamment régulière pour entreprendre une modélisation, cependant:
- La dispersion importante (par rapport à la structure du dialogue) de l'occurrence des nouvelles requêtes donne quelques inquiétudes quant à l'utilisation effective des prédictions pour aider la RAP.

3. LA SECONDE PHASE DE L'EXPERIMENTATION.

3.1. Le protocole.

Le cadre matériel de l'expérimentation est identique à celui de la phase 1. L'accent est porté sur la spécification du dialogue visé. Nous nous limitons ici à décrire les aspects liés à la pragmatique du dialogue et à la gestion phatique.

La coopération: les principes coopératifs appliqués dans cette phase de l'expérimentation sont issus des travaux de Kaplan (KAPLAN,82) et sont plus longuement exposés dans (GUYOMARD,86). Ils sont fondés sur la règle pragmatique suivante: lorsqu'une question ensembliste entraîne une réponse stricte vide, il est préférable:

- de démentir les présuppositions fausses contenues dans la question puis,
- de tenter de relancer le dialogue en proposant la réponse à une question affaiblie issue de la question initiale.

On peut décrire le principe de la coopération de la manière suivante: le "système" est dans la situation où un problème P se pose à lui; la solution à ce problème peut être: inexistante (0 solution), unique (1 solution), multiple (n (>1) solutions).

Dans le premier cas le problème P est affaibli en relâchant certaines de ses contraintes. On obtient un problème P' qui, à son tour, possède 0, 1 ou n solutions; le résultat est alors proposé à l'utilisateur en présupposant qu'il peut le refuser, le problème initial ayant été trop déformé pour lui convenir.

Ce cadre général trouve son application tant dans la partie négociation de requête que durant la négociation de réponses.

La situation inverse, caractérisée par un nombre excessif de réponses est également problématique. La solution retenue ici, dans le cas où l'utilisateur ne prend pas l'initiative d'une sélection, consiste à proposer un choix binaire résultant d'une dichotomie sur un critère supposé intéressant.

Les dialogues de continuation: ce sont les dialogues qui se poursuivent à l'initiative de l'usager par l'énoncé d'une nouvelle requête. Au cours d'une session, l'influence d'un dialogue peut se faire ressentir après son achèvement (annonce d'un résultat). Il est nécessaire d'en tenir compte si l'on veut conserver le caractère naturel du dialogue. Nous avons formalisé cette influence du contexte en élaborant 7 règles dont voici un exemple:

Règle d'acquisition de paramètre obligatoire:
Si un paramètre obligatoire de la nouvelle requête n'est pas présent dans l'énoncé initial et si ce paramètre n'a pas été modifié dans le dialogue précédent par une réponse suggestive, on prend pour valeur de ce paramètre celle qui était utilisée dans le dialogue précédent.

Les énoncés implicites: nous avons conservé le principe des énoncés implicites et systématisé son utilisation.

La gestion phatique: nous avons opté pour la stratégie suivante: une mauvaise compréhension du système produit une demande de répétition en bloc ("Pardon"). Le système accepte et interprète les demandes d'arrêt anticipé, les demandes d'épellation et les demandes de répétition partielle. Comme Waterworth (WATERWORTH,86) nous avons basé le processus de confirmation-correction sur un jeu de règles:

- les items particulièrement informatifs sont répétés par le "système".
- le ton employé pour la répétition par le système est soit confirmatif soit interrogatif selon le degré de confiance accordé à l'information par le système. Dans le premier cas (confirmatif) une réponse de l'utilisateur n'est pas obligatoire, alors que dans le second une réponse est attendue.
- un énoncé utilisateur comportant plus d'un item informatif inhibe le processus de confirmation/correction (à l'inverse du choix de Waterworth).

Les consignes données aux interlocuteurs: les consignes proposées au sujet lui demandent de formuler des phrases simples et courtes. Lors de chaque session, la première intervention est construite mentalement avant d'être présentée oralement à l'aide.

3.2 Phase 2. Analyses et résultats.

Les 15 sujets retenus ont permis de recueillir 70 conversations. Nous ne donnons ici que les principales tendances que nous avons pu observer lors des analyses.

L'impression générale qui se dégage de la phase 2 est que les options retenues pour la spécification sont bonnes: la structure de dialogue envisagée n'a pas été mise en défaut, les règles de coopération élaborées ont joué leur rôle, le principe des interventions facultatives est un facteur important de naturel et d'allégement.

L'analyse de la requête initiale montre que les sujets ont tenu compte des consignes de concision. On retrouve les formes déprécatives (55 cas) et interrogatives (23 cas), mais la forme impérative disparaît au profit de la forme elliptique (2 cas).

Les nouvelles requêtes (c'est à dire les requêtes qui amorcent un dialogue de continuation) ont été classées en trois catégories totalisant 32 occurrences:

- les glissements (22 cas),
- les mutations (7 cas),
- les reproductions (3 cas).

Un glissement représente une modification de la requête précédente. Cette catégorie regroupe 22 énoncés dont 13 qui surviennent malgré un succès apparent du dialogue précédent. Une mutation est caractérisée par l'emploi d'une rubrique non évoquée dans le dialogue précédent; on y retrouve les 5 cas de scénarios enchaînés prévus par le protocole. Les reproductions constituent une forme de répétition de la requête précédente; ce phénomène s'explique par une incompréhension de la part du compère (confusion entre "moto" et "auto" par exemple) et par un certain scrupule de la part du sujet à interrompre un dialogue mal engagé. 27 des 32 nouvelles requêtes font suite à la sollicitation du système ("voulez-vous un autre renseignement?"). Les 5 cas résiduels apparaissent plus en amont dans le dialogue et sont à porter à l'initiative du sujet. La forme d'une nouvelle requête est identique à celle d'une requête initiale dans 14 cas; elle est elliptique ou plus rarement anaphorique dans les 18 autres cas.

A 39 occasions le compère a dû réclamer un paramètre jugé obligatoire, 38 fois pour la localité, une seule fois pour la rubrique. La réponse à une telle demande est le plus souvent elliptique et 2 cas de réponses sur-informatives sont à signaler.

Le comportement coopératif du compère peut conduire à négocier les paramètres. Cette situation s'est produite 39 fois. Un seul refus de négocier de la part du sujet a été observé. La forme de la réponse de l'utilisateur est en général elliptique et la désignation de l'item retenu se fait le plus souvent par reprise, de préférence à une description ordinale.

Durant la phase de négociation de réponse ont été observées:

- 10 demandes de précision (dont 7 situées immédiatement après l'énoncé de relance "voulez-vous un autre renseignement?").
- 27 demandes de sélection (dont 21 à l'initiative du compère).
- 15 demandes d'arrêt pendant l'annonce par le système de la liste des réponses élémentaires.

La forme linguistique des demandes d'arrêt est très stable ("ça suffit", "vous pouvez arrêter"). Dans 3 des 15 cas, la demande d'arrêt a été suivie d'un dialogue de continuation.

Les principaux comptages se rapportant à la gestion phatique sont représentés à la figure 1.

De même que dans la phase 1, les demandes de répétition de la part du sujet portent le plus souvent sur le contenu d'un énoncé très informatif. Dans 22 cas, la demande ne porte que sur une portion de l'énoncé (nom, numéro de téléphone). Dans 28 cas

la demande vise l'énoncé immédiatement antérieur. La forme linguistique est très stéréotypée.

fonction phatique	initiative (S/U)	nombre d'occurences
demande de	S	36
répétition	U	36
demande de confirmation	U	1
demande d'épellation	U	6
prise de parole de U pendant les silences de S	U	21
recouvrement	S	3
de parole	U	23

S: système ; U: utilisateur

Figure 1. La gestion phatique - Comptages.

Le vocabulaire utilisé par les sujets durant cette phase comprend 324 termes. Les phénomènes "velléitaires" (inachèvement de mots, hésitations,...) de l'oral spontané sont ici aussi très nombreux.

Les appréciations des sujets sur le système mentionnent le caractère réceptif de la "machine", le dialogue bien mené et la non-possibilité d'obtenir des précisions sur les réponses obtenues.

4. CONCLUSION.

Nous avons présenté les protocoles et les principaux résulats d'une expérimentation de dialogue oral; les corpus recueillis sont disponibles sous forme de transcription écrite et sur cassettes (5,5 heures d'enregistrement, qualité non audio). Des résultats plus détaillés peuvent être trouvés dans (GUYOMARD,87) et (PALABRE,87).

Le travail réalisé a permis d'expérimenter et de valider un mécanisme de coopération et une structure de dialogue qui, à l'évidence, peut se généraliser à une large classe d'applications avec dialogues finalisés. La structure de dialogue a été formalisée par une grammaire de dialogue (PALABRE,87). Cette grammaire est écrite à l'aide de 153 non-terminaux et est constituée de 244 règles. La sémantique sous-jacente inclut les notions de coopération et de gestion phatique décrites dans ce papier.

Signalons également notre inquiétude vis-à-vis du réalisme d'une application dans le contexte d'usagers occasionnels; deux difficultés viennent l'alimenter. Il s'agit d'une part du problème des phénomènes d'hésitations et de reprise dans l'oral spontané, la seconde phase confirmant qu'il est difficile d'y échapper; et d'autre part du problème des prédictions. Les chercheurs en RAP ont fondé des espoirs sur la capacité prédictive des dialogues pour éviter les problèmes d'explosion combinatoire auxquels ils sont confrontés. Or la puissance prédictive d'un dialogue est une fonction inverse du degré de liberté envisagé pour l'utilisateur dans la définition du dialogue. Il est à craindre que les contingences de la mise en oeuvre exigent beaucoup des prédictions et nous obligent donc à restreindre la richesse du dialogue.

Enfin, cette expérimentation met en évidence l'existence d'une expertise liée au dialogue, indépendante de l'"expertise" linguistique et de l'expertise portant sur l'application (TAYLOR,86). Elle possède ses propres règles et ses propres finalités parmi lesquelles on peut citer:

- sa faculté à structurer une conversation,
- ses objectifs de convivialité,
- sa capacité à faciliter la tâche de la RAP, en diminuant le nombre des énoncés utilisateur et leurs longueurs sans entamer la qualité du dialogue.

Remerciements.

Nous tenons à remercier Mademoiselle J. Damay, Messieurs J. Bluteau, A. Cozannet, E. Dalila et R. Descout pour leur collaboration à ce travail.

BIBLIOGRAPHIE.

(GRECO,85) Groupe "Dialogue Homme-Machine" du GRECO "Communication Parlée". Dialogue Oral Homme-Machine en Situation Orientée par l'Action. 5ème congrès AFCET-RFIA, Grenoble, 27-29 Novembre 1985, 281-295.

(GUYOMARD,86) M. GUYOMARD, J. SIROUX. Suggestive and Corrective Answers: a Single Mechanism. Structure of Multimodal Dialogues Including Voice. NATO Workshop, Sept 1-5 1986, Vénaco France.

(GUYOMARD,87) M. GUYOMARD, J. SIROUX. An Experimental Oral Dialogue Specification. NATO ASI on Recent Advances in Speech Understanding and Dialogue Systems, July 5-18 1987, Bad Windsheim, F.R. Germany.

(KAPLAN,82) S.J. KAPLAN. Cooperative Responses from a Portable Natural Language Query System. Artificial Intelligence 19 (1982), 165-187.

(PALABRE,87) ESPRIT Project P 1015, Palabre. Deliverable WP 4.3.

(ROULET,81) E. ROULET. Echanges, Interventions et Actes de Langage dans la Structure de la Conversation. Etudes de Linguistique Appliquée, No 44, Oct-Déc 1981, 7-39.

(SIROUX,85) J. SIROUX, D. GILLET. A System for Man-Machine Communication Using Speech. Speech Communication. Vol 4 No 4, 1985, 289-315.

(SIROUX,86) J. SIROUX. Pragmatics in a Realisation of a Dialog Module. Structure of Multimodal Dialogues Including Voice. NATO Workshop, Sept 1-5 1986, Vénaco France.

(TAYLOR,86) M.M. TAYLOR. Natural Dialogue is not Natural Language. Structure of Multimodal Dialogues Including Voice. NATO Workshop, Sept 1-5 1986, Vénaco France.

(VAN KATWIJK,79) A.F.V. VAN KATWIJK, F.L. VAN NES, H.C. BUNT, H.F. MULLER, F.F LEOPOLD. Naive Subjects interacting with a Conversing Information System. IPO Annual Progress Report 14, 1979, 105-112.

(WATERWORTH,86) J. WATERWORTH. Interactive Strategies for Conversational Computer Systems. Structure of Multimodal Dialogues Including Voice. NATO Workshop, Sept 1-5 1986, Vénaco France.

DIALORS : un système de dialogue oral simulé pour une tâche restreinte

D. Luzzati

LIMSI/CNRS - BP 30 - 91400 ORSAY CEDEX - FRANCE

ABSTRACT

DIALORS is a train timetable automatic dialogue system including a skimming parser (ALORS) and a dialogue module (DIALOG). Our purpose is to simulate, presently with a keyboard input, communication between users of a public telephone information service and what they believe to be a machine (an operator using a vocoder and adapted behaviour).

1. PRESENTATION

DIALORS est un système automatique de dialogue, qui intègre un analyseur sélectif (ALORS) et un dialogueur (DIALOG), dans le cadre d'une tâche de renseignements horaires. Son objectif est de reproduire, actuellement à partir du clavier, des conversations effectives, réalisées avec machine simulée, et de se substituer autant que possible à l'opérateur qui simule la machine.

1.1. Le corpus

Le corpus utilisé a été réalisé à la SNCF dans le cadre du GRECO CP, avec le soutien financier du CNET, à partir du protocole d'expérimentation défini et testé dans le projet ORSO (1) . Il s'agit de mettre des correspondants réels en situation de communication humain-machine afin de pouvoir se fonder sur la façon dont ils s'expriment et se comportent effectivement dans une telle situation.

Pour imposer l'image de la machine, on utilise d'une part un vocoder pour modifier la voix du compère (méthode parfois appelée "Wizzard of Oz" (2)) ; ce dernier manifeste d'autre part un comportement qui correspond à ce que les correspondants attendent et acceptent d'une machine qui dialogue. Définir un tel comportement est certes subjectif et aléatoire. Il n'en demeure pas moins qu'il est difficilement concevable d'implanter un système de dialogue sans tenir compte des contraintes techniques et des réalités psychologiques que cela suppose. Notre propos est d'ailleurs d'exploiter, pour l'analyse comme pour le dialogue, les limitations et les contraintes que s'imposent et acceptent les locuteurs pour parvenir à un type de dialogue qui, tout en étant limité à certaines tâches simples, soit susceptible d'une implantation.

La réalisation des enregistrememnts a été prise en charge par le laboratoire d'ergonomie de Paris V (3). Le compère comprenait pratiquement tout, ce qui ne choquait nullement les correspondants, et il utilisait pour l'essentiel un ensemble d'énoncés préconstruits, ce qui nous a permis de faire l'économie d'un véritable système de génération. Ce corpus, composé de 146 communications, a été séparé en deux parties. L'étude de la première (corpus-étalon) a servi à définir les fonctionnalités du système. La seconde (corpus-test) nous sert quant à elle à son évaluation. Pour la mise au point de l'analyseur comme pour celle du dialogueur, on n'a considéré comme pertinents que les phénomènes rencontrés dans le corpus-étalon, et on a évalué stabilité du système à partir du corpus-test. On s'est par ailleurs permis quelques simplifications : le système ne traite que les demandes d'horaires, et il se limite à reconnaître d'autres tâches (tarifs, réservations) ; la base de donnée ne comporte que les trajets Paris-Dreux et Dreux-Paris.

1.2. Entrée clavier

Si, à partir d'un corpus oral, on a ainsi développé un système avec entrée au clavier, c'est pour faire porter notre travail sur les niveaux cognitifs : analyse d'énoncés spontanés et gestion du dialogue en particulier. Il serait en effet difficilement envisageable de faire avancer l'étude de ces niveaux en se fondant sur les performances actuelles des systèmes de reconnaissance. Il est en revanche intéressant pour le développement de tels systèmes d'avoir une idée des performances qui permettraient la réalisation de dialogues pseudo-naturels de ce type.

Dans le développement de l'analyseur, on s'est par ailleurs efforcé d'être compatible avec une détection de mots dans la parole continue, détection qui accepte les mots étranger au vocabulaire et qui fonctionne sans syntaxe, en fournissant par exemple un treillis de mots. Cette compatibilité est tout d'abord inhérente à la stratégie même de l'analyse sélective qui, en ce qui nous concerne, se suffit de 33% des productions des correspondants. L'ensemble des opérations s'effectue ainsi avec un lexique inférieur à 100 items (noms propres non compris). Il s'agit en somme d'obtenir des résultats acceptables à partir d'éléments aussi rudimentaires que possibles. On s'est par ailleurs efforcé de ne pas tenir compte des monosyllabes, et on a confondu les homophones sous une même étiquettes (*de* et *deux* par exemple).

Actuellement, l'ensemble du système est implanté, même si on envisage d'affiner certaines règles de gestion du dialogue. L'évaluation de l'analyseur est terminée, alors que celle du dialogueur doit attendre que la version définitive de l'ensemble des règles soit arrêtée. L'évaluation de la résistance de l'analyseur aux faux rejets et aux fausses alarmes est par ailleurs en cours.

2. L'ANALYSE (ALORS)

2.1. Principes

ALORS traite tous les énoncés des correspondants, que ce soient des requêtes complexes ou de simples assertions. Il en fournit une représentation interne que DIALOG interprète pour répondre ou pour relancer le dialogue. En pratique, l'ensemble des problèmes d'analyse ne se pose que dans les requêtes complètes. Pour bien les distinguer des questions de dialogue, on s'est donc fondé sur les 125 requêtes initiales d'horaire (65 dans le corpus-étalon, 60 dans le corpus-test).

ALORS est un analyseur sélectif (4) (5), c'est-à-dire que son objectif n'est pas de comprendre la totalité d'une requête mais simplement de reconnaître certains de ses éléments afin d'en extraire une représentation sous forme de schéma : gare de départ (GD), gare d'arrivée (GA), jour (J) et plage horaire (PH) notamment (cf. figure 1). C'est donc un analyseur déterministe, fondé sur une grammaire sémantique, qui exploite autant que possible la restriction du lexique. Il utilise par ailleurs un grand nombre de règles pragmatiques,

c'est-à-dire dépendantes du contexte. La syntaxe n'intervient enfin que de façon locale.

Comme il s'agit d'énoncés véritablement oraux, c'est-à-dire d'énoncés fondamentalement non normés, il faut tenir compte de tous les phénomènes de bruit ou de dislocation que cela suppose. Le problèmes est moins de déterminer des règles de succession que de pallier leur absence, et il est exclu de recourir à des stratégies guidées par la syntaxe. Même les approches flexibles (6) ne sont pas satisfaisantes, car il s'agit moins de tolérer des erreurs par rapport à une norme que de parvenir à une analyse alors que cette norme n'existe pratiquement pas.

2.2. Fonctionnement

En fait, sémantique, syntaxe et pragmatique sont pris en compte conjointement, tout au long des six niveaux essentiels, qui interviennent simultanément à propos de chaque item (cf. figure 1) :

une e paris tours e le dix huit janvier départ e le matin pour arriver à tours vers neuf heures et demie neuf heures quarante cinq e

1 *une paris tours dix huit janvier départ matin arriver tours neuf heures demie neuf heures quarante cinq*

2 *- paris tours dix huit janvier - matin arriver tours neuf heures demie neuf heures quarante cinq*

3 *paris tours* **dix huit** *janvier matin arriver tours* **neuf heures** *demie* **neuf heures quarante cinq**

4 *paris tours (dix huit janvier) - matin arriver tours (neuf heures demie neuf heures quarante cinq)*

5 *paris tours (18 jv) matin ⟨ ⟩ tours (9h30-9h45)*

6 ($horaire$ départ/arrivée)(T) (*paris*)(GD) (*tours*)(GA) (*18 jv*)(J) (*matin*)(PH1) - ($vers$ *9h30-9h45*)(PH2)

TACHE (T) : $HORAIRE$ départ / arrivée

GARE DE DEPART (GD)	:	paris
GARE D'ARRIVEE (GA)	:	tours
JOUR (J)	:	18 jv
PLAGE HORAIRE (PH)	:	matin / $vers$ 9h30-9h45

Figure 1 : Fonctionnement de ALORS ($$ = décision par défaut).

1 - Sélection des items appartenant au lexique de l'analyseur (77% du vocabulaire est ainsi ignoré).

2 - Sélection des items pertinents pour l'analyseur. Dans *après tout je voudrais un train après 10 h* par exemple, seul le second *après* demeure.

3 - Filtrage des étiquettes à partir de l'environnement. Dans *je voudrais l'heure d'un train le 10 janvier à partir de 10 heures 10* par exemple, les différentes occurrences de *heures* et de *10* doivent être discriminées.

4 - Traitement des noyaux syntaxiques locaux. Ces noyaux syntaxiques, de longueur variables (de 2 à 8 items), doivent pouvoir être disloqués, par l'insertion de bruit (*le 10 e voyons janvier*), ou parfois même d'éléments signifiants (*demain après-midi 10 janvier*).

5 - Transformation en une représentation interne. Dans l'exemple cité, *arriver* (représenté ⟨ ⟩) est par exemple interprété comme un marqueur qui signifie que la PH qui précède doit être considérée comme PH de départ, et celle qui suit comme PH d'arrivée.

6 - Affectation, éventuellement par défaut, des variables du schéma. Le monosyllabe *vers* est ainsi systématiquement réintroduit, alors qu'il ne fait pas partie du lexique de l'analyseur, de même qu'une règle de contigüité permet d'écarter le second *Tours* de la sélection de GD et de GA.

2.3. Evaluation

Pour le lexique comme pour la syntaxe, le système apparaît rapidement stable, puisque aucune modification n'est nécessaire pour traiter le corpus-test. Certains phénomènes curieux sont même confirmés, tels que la quasi-absence de *quart* et de *demie*, ou l'absence totale d'indication négative de l'heure (* *10 h moins 10*). On peut ainsi considérer que s'il se rencontre des formulations qui viennent à l'esprit et que ALORS n'est pas capable de traiter, la pertinence de ces formulations est plus en cause que les performances de l'analyseur.

L'évaluation de ALORS en termes de taux d'erreur est difficile, voire impossible. Les différents types d'erreur n'ont en effet pas le même poids selon qu'il s'agit d'erreurs sur la nature de la tâche, sur l'identification de GD ou de GA, sur la précision de J ou de PH. L'appréciation de ces erreurs est ensuite difficile à définir, car on peut se fonder aussi bien sur la différence entre l'interprétation fournie par ALORS et celle d'un ensemble de témoins naïfs que sur la propriété ou la non propriété de la réponse fournie par l'ensemble du système, ce qui ne revient pas nécessairement au même.

En tout état de cause, on peut considérer qu'il reste 4 erreurs avec le corpus-test, et que 1 problème demeure non traité dans le corpus-étalon (le dialogue peut rectifier l'ensemble de ces erreurs par l'adjonction d'un échange supplémentaire) :

- Erreur quant à la nature de la tâche (1). *je vous prie* est interprété comme une demande de tarif.
- Erreurs quant à l'affectation de GD/GA (2). Dans un cas, le système interprète deux gares limitrophes comme le couple GD/GA, alors qu'il y avait ellipse de GD. Dans l'autre (corpus-étalon), il effectue une mauvaise affectation, du fait de son ignorance de monosyllabes comme *à* ou *de* (*aller à Verdun de Paris*)
- Erreur liée à une ignorance lexicale (1) : *turbo*).
- Erreur d'interprétation (1). Dans *vendredi 5 14 décembre*, le système comprend qu'il s'agit du vendredi 5 du mois en cours ou du mois suivant.

3. LE DIALOGUE (DIALOG)

3.1. Principes

DIALOG est fondé sur un modèle de dialogue (7) (8) issu de l'analyse du corpus (9). Ce modèle distingue trois niveaux d'échange : question-réponse principale (QP RP), question-réponse secondaire (QS RS), question-réponse incidente (QI RI), associé à un système de liaisons (cf.figure 2).

M1: SNCF bonjour énoncez votre demande s'il vous plaît
C1: horaire des trains pour Poitiers e 28 décembre 84 à partir de 18 heures 30
M2: pouvez-vous reformuler votre demande s'il vous plaît
C2: e pour horaire des trains pour le vendredi 28 décembre 84 à partir de 14 heures
M3: ne quittez pas .....
M4: un express de première et deuxième classe part de la gare d'Austerlitz à 15 heures 54 minutes arrive à la gare de Poitiers à 18 heures 53 minutes ce train vous convient-il
C3: non
M5: que désirez-vous
C4: à partir de 18 heures
M6: ne quittez pas .....
M7: un express première et deuxième classe pare de la gare de Paris-Austerlitz à 19 heures 20 minutes arrive à la gare de Poitiers à 21 heures 45 minutes ce train vous convient-il
C6: il y a première et deuxième classe
M10: oui il y a première et deuxième classe
C7:et entre 18 heures 05 et 19 heures 20 il n'y en a pas y a pas un supplément un train supplémentaire
M11: pouvez-vous reformuler
C8: un train supplémentaire entre 18 heures 05 et 19 heures 20
M12: ne quittez pas .....
M13: un corail première et deuxième classe part de la gare de Paris-Austerlitz à 18 heures 45 minutes arrive à la gare de Poitiers à 21 heures 29 minutes ce train vous convient-il
C9: e première et deuxième classe il est première et deuxième
M14: oui il est première et deuxième classe
C10: oui ce train me convient
M15: désire-vous d'autres renseignements
C11: il a bien lieu demain
M16: oui ce circ ce train circule demain
C12: bon d'accord
M17: désirez-vous d'autres renseignements
C13: non non non
M18: ne quittez pas

_⟨⟨QP(*QI* RI)/*RP*⟩_⟨QS/*RS*(QI *RI*)⟩⟨QS(*QI* RI)/*RS*(QI *RI*)_(QI *RI*)⟩⟩_

Figure 2 : Modèle de dialogue appliqué à une communication du corpus SNCF :
Q/R principale ; Q/R secondaire ; Q/R incidente. *machine* ; correspondant ; question ; réponse ; liaisons / = temps d'attente

Le couple QP RP constitue le noyau de base du modèle, car il reflète l'organisation thématique des communications. Chaque QP définit un espace de travail, dans lequel évoluent les QI, espace qui est remis en cause lors de l'apparition d'une nouvelle QP (changement de tâche ou de trajet). La liaison QP-1/QP, qui peut véhiculer, par le biais des liaisons précédentes, des informations fournies dans QP-(1+n), s'effectue selon trois modes essentiels :

- Destruction de QP-1 dont aucun élément ne sera repris dans QP (changement de trajet).
- Fusion entre QP-1 et QP, ce qui revient à effectuer une gestion sélective de l'ellipse (trajet retour).

- Inhibition de QP-1, afin de pouvoir y revenir après traitement de QP (demande de tarif ou de réservation sur un même trajet).

Ces problèmes de liaison QP-1/QP sont d'eux mêmes caducs dès lors que la requête comble à elle seule l'espace de travail ce qui, même lorsque cela conduit à une redondance, est extrêmement fréquent. Compte tenu de l'usage, il est par ailleurs inutile de s'embarrasser des problèmes inhérents au traitement des questions multiples, dans la mesure où, face à un opérateur comme face à une machine présumée, les correspondants reformulent systématiquement la partie non traitée, souvent d'ailleurs avec des modifications.

Les QS sont des questions elliptiques qui se limitent à modifier l'espace de travail défini par la QP précédente, et elles appellent toujours une réponse de même nature que la RP. Parfois, elles modifient J ou PH (*dans la soirée, demain à la même heure*), mais la plupart des QS supposent simplement une action sur le type de recherche dans la base de données (*train suivant-précédent*). La multiplication des QS conduit alors soit à une procédure de balayage (train n, n+1, n+2 ...), soit à une procédure d'encerclement (train n, n+1, n-1, n+2, n-2 ...).

Les QI se situent sur un autre axe que QP et QS . Ce sont soit des demandes de reformulation, soit des demandes de précision. Lorsqu'elles sont le fait de la machine, elles suivent en général une QP, et il s'agit de compléter un espace de travail afin de pouvoir fournir une réponse. Lorsqu'elles sont le fait du correspondant, elles suivent presque toujours une RP ou une RS, et elles posent les problèmes les plus délicats. L'interprétation des demandes de reformulation est souvent difficile : à la suite d'un renseignement concernant un train qui part à 6 h 57, on ne peut pas par exemple interpréter indifféremment *6 h 57 merci* et *merci 6 h 57*. Ces interventions sont en effet soit des demandes de reformulations, c'est-à-dire des QI, soit de simples phénomènes d'écho, c'est-à-dire des interventions de liaison qui équivalent à un *oui* assertif (cf. figure 6, C7). Pour les demandes de précision, qui reviennent sur une partie de la réponse préalable, on n'a pas encore fixé toutes les règles, qui doivent aboutir à un juste milieu entre deux options :

- Fournir délibérément un ensemble d'informations suffisamment vaste pour que la réponse escomptée ait de fortes chances de s'y trouver ce qui, dans un certain nombre de cas, n'est pas satisfaisant.
- Chercher à fournir exclusivement la précision demandée, ce qui est souvent difficile et parfois dangereux (maniement de *oui/non*). Cela peut induire par ailleurs un dialogue beaucoup plus "naturel", qui fasse perdre le bénéfice d'une communication bien hiérarchisée.

Ce modèle n'a en effet pas à se soucier des problèmes de tour de parole qui, dans une telle situation de communication, sont systématiquement respectés. Il permet cependant un degré de complexité qui dépasse amplement la complexité effective des communications. Il autorise ainsi de longues suites de questions : une réponse pour le modèle peut fort bien être formulée à la modalité interrogative. Il autorise également les différentes formes d'enchâssement (Q/R <QI1 RI1> <QI2 RI2> ou Q/R <QI1<QI2 RI2>RI1>), car il opère sur deux axes : l'axe horizontal des QP/RP-QS/RS qui comporte de simples décrochements, et l'axe vertical des QI/RI qui peut être activé une ou plusieurs fois à chaque noeud de l'axe horizontal, et qui autorise autant d'enchâssements qu'on le souhaite (cf. figure 3).

3.2. Fonctionnement

DIALOG fonctionne à partir de la manipulation de schémas (10). Chaque intervention du correspondant donne lieu à l'instanciation d'un schéma. La mise en relation de ce schéma instancié avec le schéma passif, issu des instanciations précédentes, donne lieu à un nouveau schéma actif qui deviendra le shéma passif lors de l'intervention suivante. L'ellipse est ainsi gérée à partir du schéma actif qui constitue un espace de référence, que cet espace soit plein (ellipse autorisée) ou qu'il soit vide (ellipse interdite) (cf. figure 4).

Chaque communication sélectionne un chemin dans un arbre théorique, qui reflète l'ensemble des dialogues possibles. Ce qui est surtout intéressant, c'est que ce modèle est à la fois dynamique et formel, c'est-à-dire que le système peut s'en servir dans le courant du dialogue, pour prévoir à chaque instant les types d'interventions qui sont susceptibles de survenir, et qu'il dispose de critères formels pour différencier ces interventions.

Suivant le type d'intervention qu'il vient de faire (question introductrice, RP, RS, QI, RI), le système sait quelles sont les trois types d'interventions qui peuvent à chaque fois effectivement survenir (cf. figure 5)). Deux de ces interventions correspondent toujours à un type d'instanciation du schéma. Reste à chaque fois les

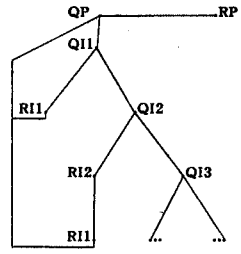


Figure 3 : Schéma du modèle de dialogue. L'axe des QI peut être activé à n'importe quel noeud de l'axe des QP/QS, à l'initiative du correspondant (+RP/RS) comme à celle de la machine (+QP/QS).

1	C : je voudrais un train dreux paris	T = [HORAIRE] GD = dreux GA = paris J = [??] PH = [??]
2	M : quel jour désirez-vous partir ? C : samedi dans la matinée	T = [HORAIRE] GD = dreux GA = paris J = samedi PH = matinée
3	M : train express No 3054 1ère et 2ème classe départ Dreux à 9 h 25 arrivée Paris-Montparnasse à 10 h 12 sans supplément : ce train vous convient-il ? C : pouvez-vous me donner l'heure d'un retour dimanche dans l'après-midi	T = [HORAIRE] GD = paris GA = dreux J = dimanche PH = après-midi
4	M : train express No 3067 1ère et 2ème classe départ Paris-Montparnasse à 16 h 35 arrivée Dreux à 17 h 32 sans supplément : ce train vous convient-il ? C : et dans la soirée	T = [HORAIRE] GD = paris GA = dreux J = dimanche PH = soirée
5	M : train omnibus No 6427 1ère et 2ème classe départ Paris-Montparnasse à 18 h 44 arrivée Dreux à 19 h 48 sans supplément : ce train vous convient-il ? C : plus tard s'il vous plaît	T = [HORAIRE]+ GD = paris GA = dreux J = dimanche PH = 18 H 44
6	M : train corail No 3065 1ère et 2ème classe départ Paris-Montparnasse à 19 h 19 arrivée Dreux à 20 h 16 sans supplément : ce train vous convient-il ? C : et pour les tarifs	T = TARIFS GD = GA = J = PH =
7	M : les questions relatives aux TARIFS ne font pas partie de mon domaine : désirez-vous un autre renseignement ? C : bon vous m'avez dit dix neuf heures vingt c'est ça M : train corail No 3065 1ère et 2ème classe départ Paris-Montparnasse à 19 h 19 arrivée Dreux à 20 h 16 sans supplément : ce train vous convient-il ?	T = [HORAIRE]+ GD = paris GA = dreux J = dimanche PH = 19 H 20

Figure 4 : Fonctionnement de DIALORS par manipulation de schémas.

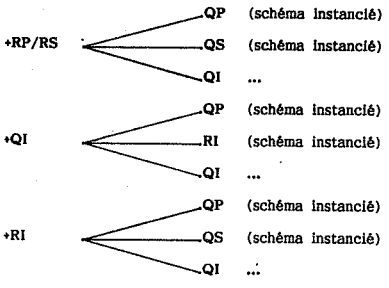


Figure 5 : Situations d'interprétation des interventions des correspondants suivant les types d'intervention du système

CORPUS

M1 : SNCF bonjour quel renseignement désirez-vous obtenir?
C1 : j'aimerais avoir les renseignements le les horaires de train e de gare montparnasse pour aller vers à calais s'il vous plait
M2 : s'agit-il de calais-maritime ou de calais-ville
C2 : calais-ville
M3 : ne quittez pas .....
M4 : vers quelle heure désirez-vous partir
C3 : vers à peu près midi s'il y en a
M5 : ne quittez pas .....
M6 : le train express 2013 part de la gare de paris-nord à 9 h 25 arrive à calais-ville à 13 h 01
C4 : et y en a pas avant avant neuf heures vingt-cinq y en a pas d'autres
M7 : ne quittez pas .....
M8 : oui le train express 2003 corail 1ère et 2ème classe part de paris gare du nord à 6 h 57 mn arrive à calais-ville à 10 h 23 mn ce train vous convient-il?
C5 : je n'ai pas compris l'heure de départ
M9 : oui le train express 2003 corail 1ère et 2ème classe part de paris gare du nord à 6 h 57 mn arrive à calais-ville à 10 h 23 mn ce train vous convient-il?
C6 : mais le départ ne peut pas être à dix heures cinquante-sept si l'arrivée est à dix heures aussi ou alors je ne comprends pas
M10 : départ paris gare du nord à 6 h 57 mn
C7 : ah six heures cinquante-sept merci
M11 : désirez-vous un autre renseignement?
C8 : non ça va merci
M12 : ne quittez pas .....

DIALORS

MM1 : SNCF bonjour quel renseignement désirez-vous obtenir?
CC1 : j'aimerais avoir les renseignements le les horaires de train e de gare montparnasse pour aller vers à dreux s'il vous plait
MM2 : quel jour désirez-vous partir?
CC2 : demain vers à peu près midi s'il y en a
MM3 : train omnibus No 6413 1ère et 2ème classe départ Paris-Montmarnasse à 12 H 28 arrivée Dreux à 13 H 33 sans supplément : ce train vous convient-il?
CC3 : et y en a pas avant avant douze heures vingt huit y en a pas d'autres
MM4 : train express No 3061 1ère et 2ème classe départ Paris-Montmarnasse à 11 H 22 arrivée Dreux à 12 H 9 sans supplément : ce train vous convient-il?
CC4 : je n'ai pas compris l'heure de départ
MM5 : je répète : train express No 3061 1ère et 2ème classe départ Paris-Montmarnasse à 11 H 22 arrivée Dreux à 12 H 9 sans supplément : ce train vous convient-il?
CC5 : mais le départ ne peut pas être à douze heures vingt deux si l'arrivée est à douze heures neuf ou alors je comprends pas
MM6 : je répète : train express No 3061 1ère et 2ème classe départ Paris-Montmarnasse à 11 H 22 arrivée Dreux à 12 H 9 sans supplément : ce train vous convient-il?
CC6 : ah onze heures vingt deux merci
MM7 : désirez-vous un autre renseignement?
CC7 : non ça va merci
MM8 : au revoir

Figure 6 : Exemple de dialogue simulé à partir d'une communication réelle.

QI du correspondant, que le système peut analyser à partir de procédures spécifiques. Si cette analyse est infructueuse, l'ultime recours consiste à solliciter une reformulation, c'est-à-dire à procéder à une QI enchâssée.

3.3. Evaluation

L'évaluation à partir du corpus-test n'étant pas encore effectuée, on se limitera à quelques observations qui concernent le corpus-étalon :

- La simplification de la base de données évite de se trouver confronté à un certain nombre de problèmes liés à la longueur du trajet ou à des particularités locales (cf. figure 6, M2-C2).
- Un certain nombre de situations réclament un échange supplémentaire (12 occurrences). Il s'agit pour l'essentiel de la non compréhension des déterminants monosyllabiques indiquant J (*ce soir, cet après-midi*), de l'indication relative de l'heure (*6 h*), ou d'une ambiguïté potentielle non levée par l'opérateur qui simule la machine (cf. figure 6, MM2).
- Dans d'autres situations, le système parvient à faire l'économie d'un échange. Il s'agit en général soit d'une formulation de PH ressentie comme trop vague (*dans la journée*), soit d'une indication de J ou de PH noyée dans flot de parole.
- Différents problèmes enfin n'ont pas été pris en compte, soit parce qu'on n'en rencontre qu'une seule occurrence (J = *17 12 85*), soit parce qu'une ambiguïté est difficile à lever (*il n'y en a pas avant midi, je n'en veux pas avant midi*).

CONCLUSION

La figure 6 donne un exemple de communication simulée à partir de DIALORS, qui comporte notamment une simplification induite par la limitation de la base de donnée, deux exemples de traitement de Questions Initiales par reformulation (MM5-MM6), et un de phénomène d'écho (CC6). Dans la mise au point du système, la partie linguistique du travail, c'est-à-dire l'étude du corpus, est beaucoup plus importante que l'implantation proprement dite. Cette importance accordée au corpus dénote le parti-pris de se fonder sur l'usage, c'est-à-dire de privilégier la performance aux dépends de la compétence.

Cela suppose également que le système est surtout applicable à des tâches du même type que les renseignements horaires. On peut en revanche en tirer des enseignements plus généraux pour la modélisation à partir de corpus, l'analyse d'énoncés non normés, ou le développement de modèles de dialogues à la fois dynamiques et formalisables.

BIBLIOGRAPHIE

(1) Luzzati D., *ORSO : projet pour la constitution et l'étude de dialogues homme-machine*, Note Scientifique, LIMSI, septembre 1984.
(2) Falzon P., *Langages opératifs et compréhension opérative*, Th. 3ème cycle, Paris V, 1986.
(3) Spérandio J.C. & Létang-Figeac C., *Simulation expérimentale de la synthèse vocale en dialogues oraux de communication homme-machine : étude ergonomique*, GRECO CP, CNRS, 1986.
(4) DeJong G., *Skimming stories in real time*, Ph.D.Th., C.S.D., Yale University, 1979.
(5) Luzzati D., *ALORS : un analyseur sélectif adapté au traitement d'énoncés oraux spontanés*, Note Scientifique, LIMSI, novembre 1986.
(6) Hayes P.J. & Mouradian G.V., *Flexible parsing*, A.C.L. 1980.
(7) Roulet E. & al., *L'articulation du discours en français contemporain*, Berne, P. Lang, 1985.
(8) Taylor M., Néel F. & Bouwhuis D., *The Structure of Multimodal Dialogue*, North Holland, 1987.
(9) Morel M.A. & al., *Analyse linguistique d'un corpus d'oral finalisé*, GRECO CP, CNRS, 1985.
(10) Bobrow D. & al., *GUS, a frame-driven dialog system*, A.I., vol. 8, p. 155-173, 1977.

SYSTEME DE DIALOGUE ORIENTE PAR LA TACHE: UNE APPLICATION EN AVIONIQUE

[1]

AUTEURS : A. MATROUF, F. NEEL et J.MARIANI.

LIMSI/CNRS - BP30 91406 Orsay - FRANCE

Abstract

The aim of the system presented here is to simulate the human pilot in a spoken dialogue with an air-traffic controller, during training exercises. The study has been developed in cooperation with the CENA (Centre d'Etudes de Navigation Aérienne). We particularly focused on the elaboration of a semantic and pragmatic knowledge representation (using hierarchized frames) and on control strategies liable to minimize the number of exchanges and to correct errors due to bad recognition or to the speaker (mispronunciation, syntactic variations from phraseology...).

I Introduction

Nous présentons un système permettant de traiter un dialogue oral entre des élèves contrôleurs aériens et un simulateur de trafic aérien. Le langage utilisé dans ces exercices de formation est de type opératif [2]. Le système a pour tâche, d'une part, de poursuivre un dialogue avec l'élève et de communiquer les désirs de ce dernier au simulateur sous forme de suites d'instructions du language formel de commandes, et d'autre part, de détecter et corriger les erreurs dûes aux différentes étapes de la compréhension du message.

Nous rappelons en premier lieu les résultats d'études du langage utilisé par les élèves contrôleurs du premier niveau. Cette étude a permis de dégager des connaissances syntaxiques, sémantiques et pragmatiques. Nous justifierons ensuite la représentation choisie pour chaque type de connaissances, ensuite nous présenterons l'analyseur, les mécanismes de contrôle de validité d'un message puis de correction éventuelle et enfin l'interprétation et le dialogue.

II Analyse du langage des élèves contrôleurs

Les études des communications verbales des contrôleurs aériens réalisées par JANET [1] et FALZON [2] ont permis de mettre en évidence la grande variabilité des formes d'expressions possibles d'un même signifié. Cette observation a conduit à proposer, à partir des fréquences d'apparition des messages de chaque catégorie, un corpus réduit de messages-types capable de recouvrir toute la phraséologie permettant de sélectionner un lexique restreint (132 mots + les chiffres + les lettres + les noms propres).

[1]Cette étude est développée en collaboration avec le CENA (Centre d'Etude de Navigation Aérienne) dans le cadre du contrat no 85/C0005

Une autre étude du langage des élèves contrôleurs du premier niveau effectuée par le CENA a permis de définir un corpus et un lexique réduisant les mots-clefs à 57 au lieu de 132. Ce corpus est construit selon un compromis entre la phraséologie officielle et la phraséologie utilisée réellement par les contrôleurs. Ce premier corpus ne prétend pas recouvrir la phraséologie dans son ensemble. Le lexique comporte trois sortes de mots: Noms propres (20), les mots clefs (45) et les paramètres(chiffres et lettres).

Etude détaillée du corpus

Dès qu'un avion entre dans le secteur géré par le contrôleur, il signale sa position à ce dernier, déclenchant ainsi un dialogue qui va durer pendant toute la traversée du secteur. Le contrôleur assure simultanément un dialogue avec chaque avion présent dans son secteur; ce dialogue peut être intense ou se limiter à l'initialisation et à la terminaison; par conséquent pour chaque avion le dialogue est composé de plusieurs communications et chaque communication est elle-même composée de plusieurs échanges, un échange étant un message du contrôleur suivi d'un message du pilote et inversement.

Un message provenant de l'élève contrôleur peut être soit une question, soit une instruction soit une information. Une question est en général une demande de paramètres. L'instruction est une requête de modification ou de maintien de paramètres. Les paramètres concernent les catégories suivantes: cap, vitesse, taux, balise, transpondeur, niveau. Un message peut donc être caractérisé par sa catégorie et son type (question, instruction ou information).

Les critères de catégorisation peuvent être liés au message lui-même, à la communication, à l'échange ponctuel ou au contexte. Un message se compose ainsi de plusieurs constituants: l'action, le sujet, les paramètres, le mode d'exécution, ... Les valeurs que peuvent prendre les différents constituants du message changent selon la catégorie et le type du message.

L'étude de l'aspect sémantico-pragmatique des messages a permis de dégager des règles régissant les différents constituants d'un message ainsi que des règles régissant les constituants du message en fonction du contexte courant.

III Représentation des connaissances

Le modèle proposé est fondé sur la théorie des schémas [4]. La philosophie de ces derniers est de représenter chaque objet ou concept par sa description (ensemble de champs); puis lors de la compréhension, de vérifier dans quelle mesure un texte peut se rapporter à la description ainsi formalisée.

L'avantage de cette représentation est de pouvoir traiter un même contexte sémantique décrit par des syntaxes différentes; elle est particulièrement bien adaptée au langage des contrôleurs qui, malgré la phraséologie officielle, utilisent de nombreuses formulations pour un même signifié. Le deuxième avantage est de faciliter la correction d'un message incomplet ou entaché d'erreurs.

A chaque catégorie de messages on associe un schéma: Toutes les connaissances nécessaires à l'analyse et à la compréhension du message sont représentées dans le schéma. A chaque champs sont associés différents indicateurs: VAL, SUC ... , servant de directives pour l'instanciation.

Exemple de schéma concernant la catégorie *"niveau"*

```
(*NIVEAU
   (indicatif   FONC dindicatif )
   (action      VAL ( maintenez descendez ...))
                SUC ( sujet sens))
   (quest       VAL ( quel-est-votre ...))
                SUC ( sujet ))
   (sens        VAL ( la-descente la-montee)
                SUC ( sujet )
                PRD ( action quest ))
   (sujet       VAL ( niveau niveau-de-vol )
                SUC ( param )
                PRD ( action sens))
   (param       ISA  PNIVEAU )
   (mexec       VAL ( rapidement ))
   (methode        INIVEAU))

  (PNIVEAU
   (digit1    VAL ( 0 1 2 3 4 5 6 7 8 9 ))
   (digit2    VAL ( 0 1 2 3 4 5 6 7 8 9 ))
   (digit3    VAL ( 0 5 ))
   (valeur    FONC VNBNIV)
   (sem       (valeur inf 600)
              (valeur sup  30)
              (valeur mul   5)))
```

Représentation des contraintes sémantico-pragmatiques

Le rôle principal de ces connaissances est la détection des erreurs commises par les différentes étapes de l'analyse qui peuvent provenir de la catégorisation, de l'instanciation, de la reconnaissance ou même du locuteur.

Deux principes sont utilisés: la limitation du domaine de variabilité et la redondance de l'information [10] dans le message: par exemple, dans le message *" descendez au niveau 230 "*, l'information apportée par *"descendez"* est incluse dans l'information *"niveau 230 "* plus la connaissance du niveau actuel.

En dialogue oral, même finalisé, la redondance existe mais est difficile à cerner et donc à automatiser, il s'agit donc de rassembler toutes ces informations et choisir une représentation adéquate.

Nous avons distingué trois sortes de contraintes:

1) Contraintes propres à chaque champ d'un schéma correspondant à une catégorie: par exemple, pour la catégorie "niveau", le paramètre doit être compris entre 30 et 600 et multiple de 5. Ces contraintes sont intégrées dans le schéma.

2) Contraintes régissant deux champs d'une même catégorie (inter-action entre les champs)

Exemple : si l'action est le maintien, le mode d'exécution et la direction ne doivent pas être mentionnés et le paramètre être egal au paramètre actuel.

Puisque ces contraintes dépendent de la valeur du champ (donc du mot) et sont communes à plusieurs schémas, nous les avons intégrées dans un dictionnaire des mots.

Exemple:

```
(MONTEZ (categ          *NIVEAU)
        (conj            monte)
        (sem  (param inf actuel))
        (note              0))
```

3) Contraintes globales pour toutes les catégories.

Chaque contraintes est representée par un triplet (champ condition valeur).

Représentation de l'univers de l'avion

A chaque avion présent dans le secteur aérien du contrôleur on associe toutes les informations le concernant: état vertical (stable / en montée / en descente) état horizontal (stable / virage droite / virage gauche),niveau... Ceci permet au système d'avoir une image de l'état de l'univers de chaque avion, initialisée à l'établissement du premier contact et mise à jour à chaque interaction avec le simulateur. Le système garde aussi dans un historique du dialogue une trace des seuls messages susceptibles de provoquer une modification de certains paramètres.

IV Système de dialogue

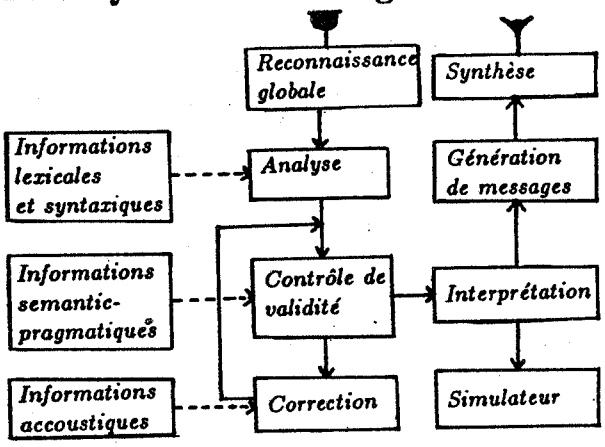


Fig 1: Synoptique du système de dialogue

L'entrée vocale utilisée est la carte MOZART permettant la reconnaissance globale de mots enchaînés pour un vocabulaire restreint d'une centaine de mots [5]. Le langage utilisé est "VLISP".

IV.1 Analyse du message

L'analyse se fait en deux étapes: catégorisation puis instanciation.

IV.1.1 Catégorisation

La catégorisation consiste a déterminer la catégorie du message en utilisant les différentes sources d'information qui sont :

le message lui même: Plusieurs mots-clefs d'un message indexent la catégorie correspondante. Un message correct ne renferme pas des mots indexant plusieurs catégories, mais les erreurs dûes à la reconnaissance globale font que des mots d'un message peuvent parfois indexer des catégories différentes: par exemple, le message prononcé est *"Descendez au niveau 4 9 0"* et le message reconnu est *"Descendez au niveau cap 9 0"*, les mots *"descendez"* et *"niveau"* indexent la catégorie NIVEAU et le mot *"cap"* indexe la catégorie CAP. La catégorie retenue

pour un message est donc celle qui a le plus de représentants dans le message.

l'échange ponctuel : Quand le système pose une question à l'élève contrôleur, l'univers des possibilités est limité et le système peut prédire la catégorie de la réponse.

l'historique du dialogue L'apparition de certains messages ou suites de messages permet parfois de prédire le message suivant: par exemple, avant que le contrôleur ne donne une instruction au pilote concernant une catégorie, en général il demande d'abord la valeur actuelle. Donc si le message précédent est une question portant sur une catégorie, le message suivant a de fortes chances d' être de la même catégorie. De même si le message précédent est une instruction sur une catégorie le message suivant a de fortes chances de ne pas être de la même catégorie.

le contexte courant : Certains états de l'univers de la tâche déterminent les messages autorisés et ceux qui ne le sont pas; certains messages sont impossibles à certains moments: par exemple, si l'avion est stable les messages concernant le taux de montée ou de descente sont inadmissibles.

IV.1.2 Instanciation

L'instanciation consiste à affecter à chaque champ du schéma correspondant à la catégorie déterminée préalablement, le mot du message qui lui correspond. L'exploration du message est orientée par les directives correspondant à chaque champ.

Stratégie d'instanciation

Pour chaque champ du schéma

- si l'indicateur = VAL, le champ prend le premier mot du message appartenant à l'ensemble des valeurs indiquées, et le mot est supprimé du message.
- si l'indicateur = FONC, la procédure correspondante est appelée et doit ramener la valeur du champ.
- si l'indicateur = ISA indique une récursivité.

Le processus d'instanciation se termine si le message est vide ou si tous les champs sont instanciés.

IV.2 Contrôle de validité

Le contrôle de validité consiste à déclencher toutes les règles de contraintes sémantico-pragmatiques associées au schéma. Il commence par les contraintes globales puis les contraintes locales du schéma et enfin les contraintes associées à chaque valeur de chaque champ. Si une règle n'est pas vérifiée, le champ concerné est marqué par '?' et éventuellement par le type de l'erreur.

Exemple:

```
Message reconnu:
  "air-france 93 tournez droite 4 2 3 0"
Schema controle
(g1     (categ      *CAP)
        (indicatif  AF93)
        (action     prenez)
        (sujet      .?)
        (direction  droite)
        (param      ISA g2))
(g2     (valeur ?))
```

IV.3 Correction des erreurs

La correction d'un message peut se faire par l'utilisation d'informations accoustiques (matrice de confusion), par retour vers le système de reconnaissance ou par retour vers le locuteur. La stratégie de correction consiste à considérer le schéma erroné comme principal et produire un schéma secondaire puis faire la fusion des deux schémas.

IV.3.1 Correction par matrice de confusion

Des tests ont été effectués avec quatre locuteurs sur un corpus de 50 phrases prononcées 2 fois, afin de déterminer les différents types d'erreurs (rejet 11,3%, confusion 56,6%, ...).

Ces tests ont permis de définir une matrice de confusion entre les mots avec pour chaque mot une liste, ordonnée selon la fréquence de confusion, de tous les mots susceptibles d'être confondus avec ce mot.

```
Exemple : (9       : noeuds 2)
          (montez : la-montee maintenez)
```

La correction se fait en deux étapes: détermination d'un ilôt de confiance puis Correction à partir de cet ilôt.

Durant la première étape, le système cherche dans le schéma erroné un champ bien instancié (ilôt de confiance) voisin (prédécesseur ou successeur selon une syntaxe locale) d'un champ erroné; à partir du mot correspondant à ce champ dans le message, on essaye de retrouver le mot correct en utilisant d'une part la syntaxe locale et d'autre part la matrice de confusion. Le mot non compatible est remplacé par un mot avec lequel il peut être confondu et qui figure dans la liste autorisée des prédécesseurs ou successeurs du mot considéré comme ilôt de confiance, et l'instanciation est recommencée sur tout le message ce qui permet de prendre en compte la répercution dûe à cette correction.

Exemple:

Message reconnu *"tournez gauche 4 2 3 0"*

Après le contrôle de validité, les champs *"sujet"* et *"param"* sont marqués par "?" Après correction du sujet le message reconnu devient *"tournez gauche cap 2 3 0"*.

Un champ corrigé peut lui même devenir un ilôt de confiance permettant de corriger d'autres erreurs.

Exemple:

Message prononcé *"tournez gauche cap 2 3 0"*

Message reconnu *"tournez gauche 4 2 droite 0"*

Première correction *"tournez gauche cap 2 droite 0"*

Deuxième correction *"tournez gauche cap 2 3 0"*

Cette technique ne corrige que les erreurs de confusion et utilise la syntaxe locale, ce qui veut dire que si une règle syntaxique locale n'a pas été définie pour ce mot erroné, il ne peut pas être corrigé. Elle doit être étendue aux autres types d'erreurs (ajout, élision ...). La matrice de confusion définie à l'aide des tests décrits plus haut ne peut pas être définitive, puisqu' elle dépend en partie des locuteurs. De plus le vocabulaire peut être étendu ou réduit; l'idéal est que le système soit capable d'apprendre lui même les caractéristiques du locuteurs et du système de reconnaissance, il peut ainsi avoir un modèle adaptable des deux. Mais cela exige des corrections manuelles pour expliciter les mots ou les séquences de mots erronés ainsi que le type de l'erreur.

IV.3.2 Correction par retour vers le système de reconnaissance

En cas d'échec de la correction par matrice de confusion d'un message erroné, on peut envisager de recommencer la reconnaissance sur le même signal, après avoir changé les seuils de rejet, de supprimer ou de changer la syntaxe utilisée par le système de reconnaissance et enfin de masquer certains mots jugés indésirables.

IV.3.3 Correction par retour vers le locuteur

Ce mode de correction est sollicité en dernier. Il a l'avantage d'être fiable mais pour des questions d'ergonomie il est souhaitable de l'utiliser le moins possible [10].

Le système demande au locuteur de préciser le ou les champs marqués par "?", en lui rappelant les parties du message déjà reconnues. La réponse du contrôleur peut être le message entier, une partie du message ou seulement les valeurs demandées, mais il ne doit ni changer de catégorie ni poser de questions.

La correction se fait par la création d'un schéma secondaire pour la réponse, et, en supposant que la prononciation de la réponse soit meilleure la deuxième fois, on accorde plus de confiance au schéma secondaire; donc la fusion des deux shémas se fait alors par écrasement du secondaire sur le principal.

Le schéma résultant est de nouveau validé: s'il renferme encore des champs marqués "?", le système recommence; si la correction n'aboutit pas au bout de deux questions, le schéma est abandonné et le système exige la répétition de l'intégralité du message.

IV.4 Interpétation

A chaque catégorie est associée une fonction d'interprétation qui est déclenchée une fois que le schéma est validé. Son rôle est de:

1) produire l'ordre (ou la suite d'ordres) à envoyer au simulateur de trafic aérien

2) produire le message à transmettre au contrôleur.

3) mettre à jour l'image de l'univers de l'avion.

IV.5 Dialogue

Le dialogue proprement dit est relativement simple dans la mesure où la phraséologie prévoit, pour un type de message donné, une réponse appropriée. L'échange, dans le cas normal, n'excède pas deux messages.

Trois types peuvent être distingués:

- le message est une information; un accusé de réception peut lui être répondu mais n'est pas obligatoire.
- le message est une question et exige une réponse.
- le message est une instruction et un collationnement est attendu.

Ces différents échanges sont indépendants et peuvent intervenir dans un ordre quelconque. Cependant, on a vu qu'en cas d'erreurs ou d'incohérence, un dialogue de correction (cf IV.3) devient nécessaire. Et les contrôles de validité mis en jeu font alors intervenir l'ensemble des connaissances.

V Conclusion

Un système de dialogue oral ne doit pas seulement être capable de comprendre les messages mais aussi être capable de détecter les erreurs introduites par le système de reconnaissance de la parole ou produit par le locuteur et éventuellement les corriger de façon interne afin de minimiser le nombre d'échanges avec lui. Contrairement au langage naturel qui pose beaucoup de problèmes pour la compréhension,le langage utilisé dans un dialogue entre spécialistes dans le cadre d'une tâche ne nécessite pas des structures complexes pour la compréhension des messages, car les concepts sémantico-pragmatiques utilisés sont clairement définis, précis et dégagés de toutes ambiguités; Cependant l'observation de dialogues réels montrent la propension des locuteurs à adopter de nombreuses variantes. Le système présenté ici tente de permettre une plus grande flexibilité tout en assurant une grande robustesse. Des tests d'évaluation sont prévus et seront comparés aux résultats obtenus avec le système de reconnaissance seul.

Références

[1] E. JANET Analyse du langage employé dans les communications d'un système de contrôle. Rapport technique NO 12 INRIA 1982.

[2] P.FALZON *"Understanding a technical langage, A schema-based apprach"*. Rapport technique NO 237 INRIA 1983.

[3] P.FALZON *"Les communications verbales en situation de travail: Analyse des restrictions du langage naturel"*. Rapport technique INRIA 1982.

[4] Ganesan KALYAN *"A unified framework for representation and recognition of speech using case frames"*. Speech Technology sept/oct 1986.

[5] J.L.GAUVAIN *"Reconnaissance de mot enchaînés et détection de mots dans la parole continue"*. Thèse de troisième cycle, Juin 1982.

[6] A. AOUATI *"Utilisation des technologies vocales dans une application multicanaux"*. Thèse de docteur ingénieur, Décembre 1985.

[7] K.H.LOKEN-KIM,R.W.RAGAN,Jr.M.G.JOOST *"Speech Application for a flexible manufacturing center"*. AVIOS Septembre 1986.

[8] M. GUYOMARD, J.SIROUX *"Le dialogueur: un intermédiaire entre l'utilisateur et l'application"*. Séminaire "Dialogue homme-machine à composante orale", Nancy, Decembre 1984.

[9] J.PITRAT *"Textes ordinateur et compréhension"*. ed EYROLLES 1985.

[10] J.HOWES *"Dialogue supervision and error correction strategy for a spoken man/machine interface"*. In *"The structure of Multimodal Dialogue"*. ed M.Taylor, F.Néel, D.Bouwhuis. North.Holland 1987 (à paraitre).

SEGMENTATION
ETIQUETAGE

16e Journée d'Étude sur la Parole - Société Française d'Acoustique - Hammamet, 5-9 oct 1987.

ETIQUETAGE, SELECTION, RECONNAISSANCE EN PAROLE CONTINUE

ANDREEWSKY A. DESI M. DEVILLERS L. RINGOT P.

CNRS - LIMSI - ORSAY

We describ a threshold-free system for automatic labelling of continuous speech, a self-validating system of references selection and results for centisecond recognition correlated whith different types of segmentation.

I. Conditions d'analyse.

Après avoir échantilloné le signal à 10 Khz, on procède à un filtrage d'ordre 1 (afin d'accentuer les hautes fréquences) puis on effectue une FFT sur 256 points que l'on convolue avec une fenêtre de Hamming. On ne conserve que 128 points du spectre d'amplitude, que l'on répartit sur une échelle de Bark afin d'obtenir un spectre sur 16 canaux. On décale l'observation d'une demi-fenêtre (128 points) et on recommence l'analyse jusqu'à la fin du signal. On obtient donc une suite de spectres représentant chacun 12,8 ms de parole.

II. Le corpus.

Il est constitué:

a. De 800 phrases phonétiquement équilibrées, prononcées avec un débit normal et sans contrainte par un locuteur masculin. Ce corpus contient un vocabulaire d'environ 2000 mots différents.

b. De 10 phrases prononcées par 10 locuteurs (5 hommes, 5 femmes) issues d'un corpus du GRECO (PEQ02A) utilisées pour tester l'étiquetage automatique en multi-locuteur.

c. De 20 phrases prononcées par un locuteur italien féminin utilisées pour tester l'étiquetage automatique en multi-lingue.

III. Etiquetage automatique.

L'objectif de l'étiquetage automatique est le suivant:
Sachant ce qui a été réellement dit, on se propose de placer des phonèmes sur le signal de parole correspondant.
On notera qu'il n'est fait appel à aucun seuil dans tout ce qui suit.

III.1. Les classes phonétiques.

O (occlusives) : /p, t, k, b, d,g/
F (fricatives) : /f, v/
S (sifflantes) : /s, z/
X (chuintantes) : /ʃ, ʒ/
N (nasales) : /m, n, n/
L (liquides) : /l, r/
I (semi-voyelles): /y, w, ɥ/
V (voyelles) : /a, e, ɛ ,i, ɔ ,o , u, y , ø , œ , ɛ̃ , ɑ̃ , ə̃ , ɔ̃ , ɑ /
DEB, FIN

En vue du traitement qui suit, on associe à toute chaîne phonétique une chaîne de classe. Par exemple à /abrypt / (abrupte), on associe la chaîne: DEB.V.O.L.V.O.O.V.FIN

III.2. Courbe d'énergie théorique.

Deux méthodes ont été développées afin d'obtenir l'allure théorique d'une courbe d'énergie (CE) associée à une chaîne phonétique. Elles sont décrites dans [12] et nous présentons ici, succintement, la méthode actuellement utilisée. Le principe de base consiste à utiliser l'évolution de l'énergie d'un phonème à l'autre, et ce au cours de l'émission d'une chaîne phonétique continue. Dans cette approche, les voyelles se situent toujours aux maxima d'énergie. Deux voyelles ou plus qui se suivent sans hiatus, sont regroupées en un seul maximum. Tant qu'à partir d'une voyelle ou d'un groupe de voyelles l'énergie théorique des phonèmes successifs décroit, on est toujours dans le maximum. Dès que l'énergie se met à croître, c'est que l'on est passé par un minimum. A partir du minimum d'énergie et jusqu'au maximum vocalique suivant, il peut s'insérer des alternances secondaires entre une occlusive suivie d'une liquide, entre une occlusive suivie d'une occlusive ... Ces principes ont été matérialisés par une matrice carrée de dimmension 10. Ci-dessous, on trouve la partie de la matrice indispensable au traitement de la chaine

/abrypta/ (abrupte).

	O	L	V	DEB	FIN
O	@	½	+	ø	ø
L	-	-	+	ø	ø
V	-	-	=	ø	-
DEB	ø	ø	+	ø	ø
FIN	ø	ø	ø	ø	ø

Pour un couple de phonèmes:
colonne: phonème de gauche (PHG).
ligne: phonème de droite (PHD).

"+" signifie que l'énergie de PHD est supérieure à celle de PHG.
"-" signifie que l'énergie de PHD est inférieure à celle de PHG.
"½" signifie que l'on peut avoir une alternance secondaire entre PHG et PHD, même chose pour "@".
"=" signifie que les énergies de PHG et de PHD sont similaires au sens de l'algorithme.
"ø" est une séquence impossible.

A l'aide de cette matrice, la chaîne "DEB.V.O.L.V.O.O.V.FIN" produit la chaîne d'alternance:
"0 1 0 ½ 1 0 @ 1 0".
des variations théoriques de la CE associée à l'énoncé /abrypt ə /, qui permet de déduire le nombre d'extréma pertinents à retenir sur la CE réelle.
Une modification de la classification phonétique et éventuellement de la matrice permet l'adaptation à d'autres langues.

III.3. Le lissage de la courbe d'énergie réelle.

La CE réelle comprend beaucoup plus d'extrema qu'il n'y en a sur la courbe théorique (CT). Il faut donc transformer CE jusqu'à obtention d'une courbe de même allure que CT. Pour cela, on effectue un premier lissage sur l'axe des énergies en ne conservant que le nombre d'extrema relatif aux alternances principales (ex: 0 1 0). On lisse ensuite selon le temps en éliminant un nombre d'extrema relatif aux alternances secondaires (½ ou @). On passe alors aux procédures d'étiquetage.

III.4 Procédures d'étiquetage.

Dans un premier temps, on associe aux extrema sélectionnés par le lissage, des étiquettes phonétiques consonnantiques aux minima, et vocaliques aux maxima. La valeur et l'ordre des différentes étiquettes sont donnés par la procédure de transformation des chaînes phonétiques. Dans un deuxième temps, on essaie de dissocier à l'intérieur des pseudo-syllabes (0½1 ou 0@1), les étiquettes multiples consonnatiques, chaque fois que la présence d'une fluctuation énergétique intra-syllabique rend cela possible. Les cas de figure alors possibles sont décrits dans SHERPA [4].

III.5 Procédures de contrôle.

Les procédures de contrôle détectent les phrases qui présentent un "risque" d'étiquetage défectueux. Elles peuvent être modifiées selon la langue ou le locuteur. Actuellement, deux critères sont opérationnels. L'un vérifie que les occlusives d'un énoncé sont placées sur les points les plus bas de la CE, l'autre vérifie que deux étiquettes consécutives sont séparées par au moins deux spectres. Lorsque l'étiquetage est jugé défectueux par les procédures de contrôle, on modifie l'importance relative des lissages énergétique et temporel. Un nouvel étiquetage est alors proposé et cela jusqu'à 6 tentatives possibles. Une phrase est rejetée après la sixième tentative.
Une fois la courbe étiquetée, on génére des références phonétiques comprenant une étiquette phonétique et un triplet de spectres.

III.6 Résultats.

Les phrases du corpus comprennent au plus 10 mots. La qualité de l'étiquetage est évaluée par rapport aux performances de l'étiquetage manuel réalisé par un phonéticien, et par rapport au fait que les occurrences d'une même chaîne phonétique sont étiquetées de la même manière, et enfin en rétroaction avec les résultats du système VARAP dont la description suit.
Pour le corpus mono-locuteur français, on a les résultats suivants :

Phrases rejetées : 15%
Etiquettes bien placées: 95%
Etiquettes complexes : 15%
complexes dissociées : 50%

Les premiers tests effectués en multi-locuteur français montrent:
- Le premier critère de contrôle est trop sélectif et rejette ainsi des phrases pourtant bien étiquetées.
- La qualité de l'étiquetage reste la même pour les phrases acceptées par les procédures de contrôle.

Il en va de même pour le corpus italien. Il est prévu d'étendre les deux corpus précédents afin de poursuivre les tests.

IV. VARAP.

VARAP est un système qui séléctionne les références d'un corpus préétiqueté automatiquement, sans utilisation de seuils, en faisant appel uniquement à la capacité de reconnaissance des dites références.

IV.1. La distance.

La comparaison des références (spectres étiquetés) se fait au sens de la distance suivante:
Soient deux spectres Oi et Xi, i variant de 1 à 16 (Oi, Xi sont les valeurs de chacun des spectres sur 16 canaux). Formons di = Oi - Xi [i = 1...16] soit Di = di+1 - di, alors la distance entre deux spectres est donnée par la formule:
D = |D1|+....+|D16|

La distance entre deux références tient compte du fait que chaque référence est constituée de trois spectres consécutifs prélevés au voisinage de l'extremum étiqueté et choisis en minimisant leurs distances respectives, l'extremum faisant toujours parti de la référence ternaire.

IV.2. Le mode de sélection.

La sélection se fait en deux étapes en comparant le fichier CORPUS à un fichier dit VALID construit dynamiquement à partir de CORPUS comme suit:

- La première référence de CORPUS est mise dans VALID.
- Les références suivantes de CORPUS sont alors comparées une à une à toutes les références de VALID. Dans l'expérience effectuée, on affiche les 15 premières références qui constituent ainsi un treillis. Cela n'est possible qu'à partir du moment où il y a déjà au moins 15 références dans VALID.
- Sur chaque treillis, un scrutin majoritaire est effectué, qui tient compte du rang des références et de leur nombre dans le treillis. Les quatre premiers phonèmes majoritaires sont ainsi retenus.
- Si le premier des quatre phonèmes retenus est de même valur phonétique que la référence corpus analysée, alors une opération dite d'incrémentation est effectuée sur les références qui ont contribuées à la reconnaissance.

Remarque 1: Une réference R est validée par les références antérieures mais elle peut être invalidée par les references suivantes non encore traitées. On a donc été contraint de faire un deuxième passage de VARAP afin de pallier à cet inconvénient.

Remarque 2: Dans VARAP, chaque treillis comprend, outre la suite des références phonétiques, le contexte gauche et droit de chaque référence du treillis, la distance entre le candidat à reconnaître et les références du treillis, les compteurs d'incrémentation des références du treillis indiquant le nombre de fois où lesdites références ont participé à une bonne reconnaissance en scrutin majoritaire.

- Si le premier des quatre phonèmes retenus est de valeur phonétique différente de la référence CORPUS analysée alors:
  - ou bien la référence CORPUS à reconnaître ne se trouve pas dans le fichier des références validées, et alors elle est introduite dans le fichier VALID.
  - ou bien la référence CORPUS qu'il faut reconnaître se trouve déjà dans le fichier VALID, alors il y a confusion. Ce cas ne se produira que dans le second passage de VARAP. On effectue alors un nouveau scrutin majoritaire par souci de cohérence avec les incrémentations. Si après ce dernier scrutin, la référence est reconnue, la validation est dite incrémentale, sinon il y a validation simple.

Les algorithmes de dépouillement du système VARAP comprennent:

A. Un tableau donnant la compression du corpus des références phonétiques, phonème par phonème, après les deux passages de VARAP, ainsi que le nombre total d'incrémentations, de validations, de confusions.
B. Un tableau donnant la relation entre la confusion et l'incrémentation par ordre décroissant de cette dernière. Il montre en particulier qu'à partir d'une incrémentation égale à dix, les confusions deviennent très faibles: 12 confusions en tout pour l'ensemble du corpus, et qu'à partir de 20 incrémentations, il y a rarement plus de une confusion.
C. Une matrice de confusion construite automatiquement.
D. Un fichier des treillis issus de VARAP.

Les expériences de validation par autocohérence ont été réalisées sur 500 phrases. Après le premier passage de VARAP, sur 9952 références du corpus, il y a eut 4810 incrémentations et 4002 validations. Après le second passage de VARAP, le nombre de références conservées est de 4375, il y a 5973 incrémentations 644 nouvelles validations et 2195 confusions.

Pour terminer, disons que le but du système VARAP est multiple:

- Vérifier la qualité de l'étiquetage automatique utilisé.
- Construire un système de sélection qui ne fasse pas appel à des seuils et qui teste l'aptitude à reconnaître des références étiquetées du corpus.
- Avoir un outil modulaire complet qui, entre la procédure et la constitution du dictionnaire de références, permet de paramétrer toutes les décisions intermédiaires. Entre autre l'adaptation au locuteur, l'utilisation de plusieurs types de distances différentes, enfin la conservation ou la mise à l'écart des références de l'apprentissage en fonction de leur aptitude à reconnaître.

Enfin, notons qu'un VARAP II est en cours de construction dans lequel il n'y a pas sélection d'emblée des références mais vérification d'abord de la cohérence du corpus avec lui même et sélection en fonction des résultats d'un dépouillement automatique.

V. Reconnaissance.

Les expériences suivantes ont été réalisées:

1) Une reconnaissance sur des références phonétiques obtenues à partir de l'étiquetage automatique

et n'ayant pas participé à l'apprentissage. On obtient des taux de 90 à 93% de reconnaisance sur un treillis de 4 candidats.

2) Une reconnaissance centiseconde où tout triplet de spectres de la phrase à reconnaître est comparé aux références du corpus. On obtient alors une suite de treillis pour chaque spectre de l'énoncé et on voit apparaître des plages phonétiques. Des heuristiques sont en cours de développement afin de proposer une suite de phonèmes à partir de la suite des plages.
Toutes ces expériences ont été menées sur un corpus de 200 phrases en utilisant 3 types de spectres :
- spectres d'amplitude
- spectres de logarithme d'amplitude
- spectres en remplacant dans chaque canal l'amplitude par le 30 millième de l'amplitude totale du spectre.
Il n'est pas possible actuellement de donner des résultats décisifs permettant d'opter pour une méthode faisant appel à un type de spectre donné, et un mixage des trois approches ouvre certaines perspectives.
Nous avons constaté que la reconnaissance est de meilleure qualité lorsque l'on utilise comme dictionnaire celui issu de l'étiquetage (et non celui produit par VARAP). Les temps de calcul sont évidemment supérieurs mais nous développons actuellement une méthode d'accès rapide au dictionnaire qui permettra d'une part d'accroître la taille du corpus d'apprentissage et d'autre part de faire de la reconnaissance locale pour améliorer les résultats de l'étiquetage automatique.

3) Avec le dictionnaire issu de VARAP, nous avons effectué des reconnaissances sur le corpus multi-locuteur. Les premiers résultats font apparaître que la qualité de la reconnaissance n'est pas aussi stable qu'en mono-locuteur. Nous envisageons à long terme un traitement d'adaptation au locuteur.

VI. Conclusion et développements.

En résumé, l'adaptation de l'étiquetage automatique à d'autre locuteurs ou d'autres langues semble ne poser que des problèmes d'ajustement au niveau des procédures de contrôle. Le système VARAP doit être amélioré et orienté essentiellement vers la détection des confusions intra-corpus sans chercher à sélectionner des représentants de classe. Les recherches sur les techniques d'accès à un grand dictionnaire de références phonétiques doivent être poursuivies afin de réduire les temps de calcul nécessaires à chaque expérience. L'élaboration d'une station de travail pour l'étiquetage est en projet.

BIBLIOGRAPHIE.

[1] ANDREEWSKY A. DESI M. FLUHR C. POIRIER F. "Une méthode de mise en correspondance d'une chaîne phonétique et de sa forme accoustique", 11ème ICA, Revue d'Accoustique, 1983, p245.

[2] ANDREEWSKY A. DESI M. POIRIER F. "Le Système SHERPA - de l'étiquetage automatique à la reconnaissance par analyse ternaire", 5ème Congrès RFIA, 1985.

[3] ANDREEWSKY A. DESI M. RINGOT P. "Le système de sélection et validation auto-adaptatif pour la reconnaissance de la parole continue". 11ème congrès internationnal des sciences phonétiques. Tallin, URSS, 1987.

[4] DESI M. POIRIER F. "Le système SHERPA: étiquetage et classification automatique par apprentissage pour le décodage automatique de la parole continue", Thèse en Sciences, Paris-Sud, Orsay, 1985.

[5] DESI M. RINGOT P. ANDREEWSKY A. "étiquetage automatique du signal de parole continue à l'aide de la variation relative d'énergie des séquences de phonèmes". 11ème congrès internationnal des sciences phonétiques, Tallin, URSS, 1987.

[6] MARIANI J. "ESOPE: un système de compréhension de la parole continue". Thèse d'Etat, Université PARIS VI, 9 juillet 1982.

[7] MARIANI J. "Méthodes en reconnaissance phonétique", 11 ème ICA, Toulouse, p. 125-137, 1983.

[8] MERCIER G. "Accoustic-phonetic decoding and adaptation in continuous speech recognition", Automatic Speech Analysis and Recognition, Reidel Publishing Co,1982.

[9] MERCIER G. GERARD M. GILLET D. NOUHEN-BELLEC A. QUINTON P. SIROUX J. "Le système de reconnaissance de la parole continue KEAL", 12ème JEP, GALF, Montréal, 1981.

[10] MICLET L. VICARD D. "Reconnaissance des parties stables de parole continue pour le décodage accousti-co-phonétique", 15ème JEP, Aix en Provence, 27-30 mai 1986.

[11] LAZREK M. HATON J.P. "Segmentation et identification des phonèmes dans un système de reconnaissance automatique de la parole continue". Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle, Paris, janvier 1984, p. 5.

[12] RINGOT P. ANDREEWSKY M. DEVILLERS L. DESI M. PARISSE C. "Segmentation et reconnaissance en parole continue à l'aide des références issues du système VARAP". 11 ème congrès internationnal de sciences phonétiques, Tallin, URSS, 1987.

LA SEGMENTATION ET L'ETIQUETAGE DES GROUPES CONSONANTIQUES DE LA BDSONS

Denis AUTESSERRE et Mario ROSSI

Institut de Phonétique d'Aix-en-Provence
U A 261 CNRS "Parole et Langage"

ABSTRACT

This study presents the first applications of our segmentation and labelling method for consonants groups comporting a R and the glide w ,using the BD SON data base. Three problems are successively addressed: the contextual modifications of polymorphic consonant R ,the combinatory nature of the different phases of R ,and the characteristic dynamic features of the glide w . The over and/or under segmentation corresponding to inserts and omissions,resulting from these problems,lead us to question the very notion of consonantic groups.

INTRODUCTION

L'existence des phénomènes de coarticulation ou de coproduction, maintenant bien connus et abondamment décrits [1] , rend souvent difficile la mise en correspondance des unités phoniques discrètes avec le signal acoustique de parole. Or les experts, qui s'engagent à fournir des segments [2] aux consultants de la base de données des sons du français, sont contraints de prendre une décision adaptée à chacune des situations particulières rencontrées. Certains indices de segmentation induisent, sans contestation possible, une frontière interne à une unité phonétique : ainsi, par exemple, entre la tenue et l'explosion d'une consonne occlusive non voisée Mais même dans les cas qui paraissent assez faciles à segmenter, passages d'une voyelle à une consonne occlusive ou constrictive non voisée (et vice versa), se pose le problème assez délicat de la délimitation des phases transitoires (établissement ou coda de la voyelle). Toutes ces difficultés s'aggravent nettement lorsqu'il s'agit de séparer les éléments constitutifs d'un groupe consonantique, surtout s'il comporte une ou plusieurs consonnes vocaliques. La segmentation peut s'avérer impossible à réaliser et il vaut mieux recourir alors à l'utilisation d'un symbole d'indétermination segmentale, plutôt que de placer, sur le signal acoustique, des limites qui résultent d'une décision hasardeuse et, par là-même, non reproductible. Mais cet effacement des frontières est-il fortuit ? Ne peut-il conduire à une réinterprétation, au niveau phonologique, de la notion même de groupe consonantique [3] ? Nous nous efforcerons d'apporter ici des éléments de réponse à ces multiples questions.

PREAMBULE

Les principes généraux de segmentation ont été exposés lors d'une précédente communication aux J.E.P. [2] , nous n'y reviendrons pas. En revanche, la présentation hiérarchisée et le nombre des étiquettes ont été sensiblement modifiés; en voici le dernier état :

Premier niveau : **MACRO-CLASSES**

VO	:	voyelle orale
VN	:	voyelle nasale
VC	:	voyelle cinétique (diphtonguée)
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 (y compris ʁ)
CO	:	consonne occlusive
X...	:	indétermination segmentale
FIN	:	fin du signal de parole

Deuxième niveau : **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 : **ATTRIBUTS**

v	:	vocalicité
c	:	consonanticité
o	:	oralité
n	:	nasalité
z	:	voisement
s	:	non-voisement
k	:	closion
h	:	souffle
f	:	bruit de friction
b	:	bruit d'explosion

L'analyse et la réflexion portent sur les groupes consonantiques les plus difficiles à segmenter des corpus ACC01 A à ACC05 A, de type "acoustique" de la BDSONS. Ce sont ceux qui incluent la consonne "polymorphe" [4] de type "R" et la consonne vocalique transitoire [w] , en contexte a Les distributions suivantes ont été retenues :

1. $\neq pRV_1$ 3. $V_1 Rp(V_2) \neq$ 5. $\neq pRwV_1$
2. $\neq bRV_1$ 4. $CV_1 RpV_2$ 6. $\neq bRwV_1$

où V représente la voyelle [a] et V la voyelle [ɐ] . Les six mots correspondants "pra", "bras", "harp(e)", "Tarbes", "proie" et "broit" ont été prononcés par les douze locuteurs représentatifs d'usages du français non marqués régionalement : six hommes (SL, BP, JM, FC, LC, JB) et six femmes (NC, RS, LT, PO, XT, MD).

Trois grands types de problèmes seront abordés successivement : ils sont relatifs aux modifications contextuelles de "R", à la nature composite de cette consonne et au comportement particulier de la consonne vocalique transitoire grave [w] .

MODIFICATIONS CONTEXTUELLES DE "R"

Les variantes positionnelles de "R" peuvent être réparties en deux grandes classes :
- la première regroupe les variantes de [ʁ] constrictif, auxquelles on attribue l'étiquette CSR;
- la seconde rassemble les réalisations vocaliques [R] et nécessite l'utilisation de deux étiquettes différentes :
  -CVR, pour les variantes battues qui s'analysent en (segment voyelle + battement + segment voyelle)[n],
  -CAR, qui renvoie à la consonne approximante de type R : le battement n'est plus individualisable et la structure périodique présente une amplitude intermédiaire entre celle d'une voyelle et d'une consonne constrictive voisée.

1. Ce sont ces allophones vocaliques qui apparaissent, au contact de la consonne occlusive bilabiale voisée, dans les mots "Tarbes", "broit" et "bras". Pour ces deux derniers contextes (/occlusive voisée-), c'est l'élément voyelle de CVR qui apparaît en premier, directement suivi du battement (alors que c'est l'ordre inverse qui se manifeste dans le mot "Tarbes"). La nature de ce battement varie grandement, en fonction des locuteurs et des conditions de production, selon les étapes suivantes (Fig.1) :
- la closion est totalement dévoisée LT (ou partiellement dans sa partie finale pour XT)
- la structure acoustique prend l'allure d'une consonne constrictive voisée, avec une amplitude de voisement croissante selon l'ordre suivant des locuteurs : NC, JB, FC, RS, PO, MD.
- les caractéristiques spectrales se rapprochent encore davantage de celles d'une voyelle pour JO, LC et SL.
- le stade ultime est atteint avec BP, qui réalise une consonne approximante CAR, dénuée de tout battement : le signal acoustique, plus riche que celui de l'occlusive voisée, s'accroît progressivement en amplitude.

2. Les assimilations progressive et régressive de dévoisement, dans les cas de "pra", "proie" et "harpe", entrainent de nouvelles difficultés de segmentation. Ainsi dans "pra", au contact de l'occlusive non voisée, la partie dévoisée de l'allophone CSR se réalise comme un bruit dont les caractéristiques sont voisines de l'explosion sourde et il devient, dès lors, assez difficile de séparer le bruit d'explosion de la partie constrictive dévoisée de [ʁ] . Il faut recourir à l'analyse spectrale, encore que ce critère apparaisse peu sûr dans ce contexte où les deux consonnes [p] et [ʁ] possèdent en commun le trait grave. Cependant l'énergie du bruit de friction de[ʁ] est souvent plus forte que celle du bruit d'explosion de l'occlusive (Fig.2). Lorsqu'elle se réalise, dans cette même position, la consonne approximante CAR, pourtant variante de CVR, a un comportement qui la rapproche plutôt de CSR. L'allophone vocalique à battements CVR présente des formes variables de transition après la consonne occlusive :
- dans le cas le plus simple l'élément voyelle est préservé et il est suivi d'un battement (JB);

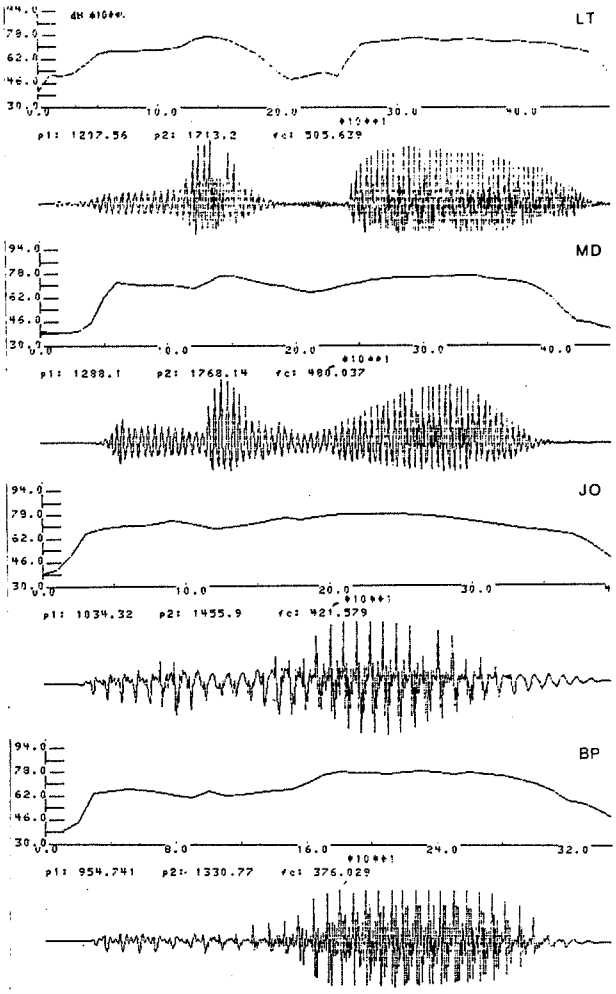


Fig.1 : Quatre réalisations différentes du mot "bras" représentatives des deux allophones de la consonne vocalique [R] : variantes battues 1, 2, 3 et approximante 4

- chez d'autres locuteurs (tel NC) l'élément voyelle est totalement assourdi et se réduit à un bruit qui se confond avec celui de l'explosion : le seul élément individualisé est le battement qui apparaît encore comme un minimum d'énergie significatif.
- enfin, [R] peut se présenter comme une suite (ɔʀ + ɔʀ) où l'élément voyelle, bien que présent, a une énergie fortement atténuée (RS).

Ainsi lorsque l'élément voyelle apparaît, la séparation entre l'occlusive et [R] est relativement aisée; le seul problème persistant concerne le caractère composite de cette consonne qui sera envisagé plus loin. Pour les autres exemples, si les critères retenus pour la segmentation de CSR sont validés, on peut procéder à la séparation de l'occlusive par rapport à la partie dévoisée de [ʁ] . Dans le cas contraire, l'élément battement est suffisamment individualisé pour permettre l'identification de [R] . Ainsi, pour l'allophone vocalique, le problème se réduit à l'identification des critères qui indiquent la présence d'un battement. Dans tous les cas, même ceux où la segmentation s'avère difficile, l'explosion de la consonne occlusive est significativement allongé, ce qui tend à prouver qu'elle contient une phase de[R]. Si, dans le contexte /-occlusive non voisée ("harpe") (Fig.3), les variantes CSR et CAR sont encore partiellement dévoisées au con-

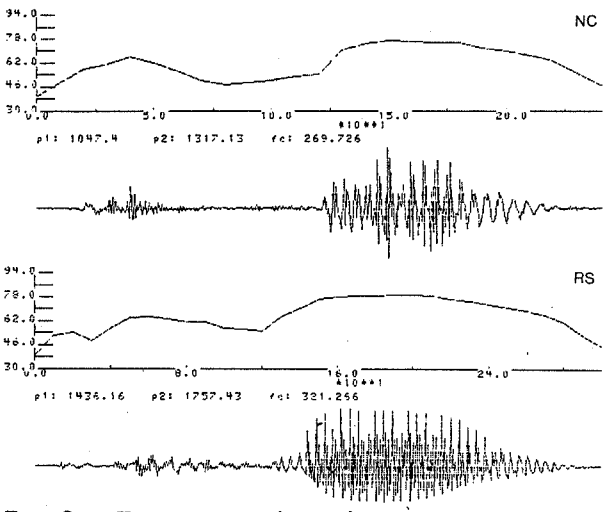


Fig.2 : Variantes dévoisées de la consonne vocalique [R] dans le mot "pra"

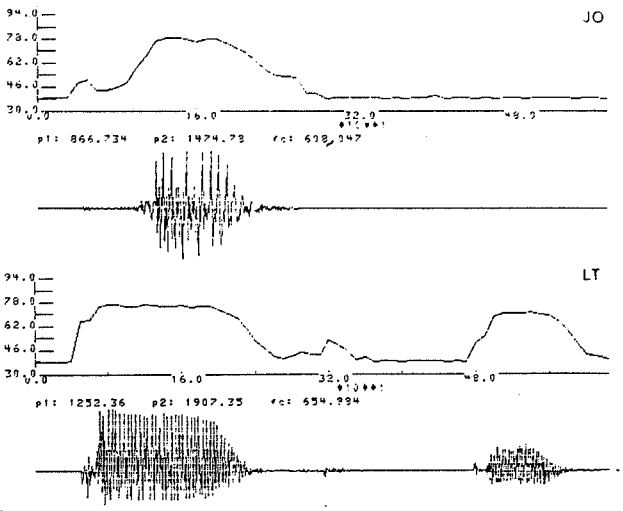


Fig.3 : Allophones constrictif [ʁ] et vocalique [R] représentés dans le mot "harpe"

tact de l'occlusive sourde, l'individualisation du bruit de friction ne présente pas de difficulté particulière puisqu'il est suivi de la tenue silencieuse de l'occlusion. Les locuteurs qui possèdent l'allophone CVR à battements peuvent utiliser deux stratégies :
- ils réalisent le plus souvent, dans cette position, l'allophone CSR (JO) : [R] est alors réduit à un bruit dévoisé entre la voyelle et la tenue de l'occlusive.
- Lorsque l'allophone CVR est prononcé, l'élément voyelle, qui sépare le battement de l'occlusion, se réduit soit à un bruit (LT), soit à une constriction partiellement voisée (LC). Par conséquent, l'identification de [R] devant occlusive non voisée revient à individualiser un bruit précédé ou non d'un battement.

NATURE COMPOSITE DE LA CONSONNE "R"

A son aspect polymorphe, la consonne de type "R" associe souvent un caractère composite qui ne facilite pas son individualisation:

1. L'allophone CSR, mises à part les difficultés de séparation de l'occlusive, évoquées plus haut, ne devrait pas poser de problèmes particuliers de sur- ou de sous-segmentation. La séparation d'avec la voyelle adjacente se pose dans les mêmes termes que ceux qui caractérisent la segmentation des consonnes constrictives.

2. L'allophone CVR, et sa variante CAR, ont un caractère vocalique marqué, bien que différemment réalisé, qui les rend difficiles à séparer de la voyelle adjacente : ainsi l'énergie de CAR augmente progressivement jusqu'à la voyelle. Dans un contexte de voyelles compactes graves, les critères spectraux de séparation sont pratiquement inexistants : on sera donc conduit, dans les cas les plus difficiles, à refuser de segmenter et à utiliser le symbole X- d'indétermination segmentale. En ce qui concerne CVR, il possède en principe la structure suivante : $[v_1 \Gamma v_2]$, ce qui signifie que, placé devant ou après une voyelle, l'élément $[v]$ pourra fort bien fusionner avec cette dernière. Lorsque les timbres de $[v]$ et de la voyelle adjacente sont éloignés, il est possible d'isoler le centre de $[v]$ des voyelles environnantes. Dans bien des cas il sera difficile de séparer, à coup sûr, l'élément voyelle de [R] à battements de son contexte vocalique. Le seul indice qui demeure est l'allongement de la voyelle (JB, RS). Toutefois, chez certains locuteurs (tel JO) l'élément $[v]$ est séparé de la voyelle par une discontinuité de l'énergie, qui est la trace ou l'annonce d'un second battement (Fig.1, réalisation de "bras" par JO).

3. La consonne CVR, lorsqu'elle se réalise comme un battement précédé et/ou suivi de l'élément voyelle $[v]$, se présente comme une unité phonétique composite. C'est le cas dans les contextes suivants :

occlusive $[v \Gamma v]$ V

V $[v \Gamma v]$ occlusive

V $[v \Gamma v]$ V

L'expert sait identifier ces différentes phases et les réunir dans une seule consonne de type "R". Chacun de ces segments est étiqueté :

v : CVR.T.v

Γ : CVR.T.k

Le caractère composite de CVR entraîne, dans une procédure automatique, une sur-segmentation. La reconstitution de l'unité phonétique, à partir des segments identifiés, ne peut être assurée qu'à l'aide d'une grammaire qui prend en compte, à la fois, les critères phonotactiques, la nature du battement (forme et durée), les caractéristiques spectrales de l'élément voyelle, le comportement du bruit d'explosion de la consonne.

LA CONSONNE VOCALIQUE TRANSITOIRE [w] DANS LES GROUPES CONSONANTIQUES

Dans les groupes consonantiques [pRw] et [bRw], les allophones de [R] se comportent différemment au regard de [w] . D'autre part, pour un même allophone, les locuteurs utilisent des stratégies souvent différentes. Dans tous les cas, la protrusion, trait distinctif de [w] , est anticipée dès le début de la syllabe sur tous les éléments du groupe [4] [5] [6] .

1. L'allophone CAR.

Si l'on compare [pRɑ] et [pRwɑ] , la seule différence provient de l'effet de la protrusion sur le deuxième formant, qui s'abaisse progressivement depuis le centre de

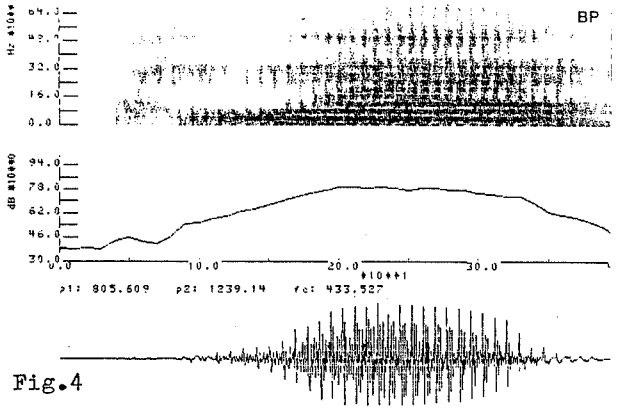


Fig.4

[ɷ] jusqu'au début de [R] . Cette dernière étant grave par nature, a pour effet d'abaisser le F2 de la voyelle adjacente; la protrusion amplifie cette action. Il est pratiquement impossible d'identifier [w] en dehors de la labialisation du groupe [pR] . Tout se passe comme si [w] n'était réalisé que par son trait distinctif, la protrusion, qui affecte le groupe [pR] et le début de [ɑ] . Il est donc vain de vouloir chercher dans le signal un segment qui n'existe pas (Fig.4 "proie" prononcé par BP). Il s'agit, en dernière analyse, dans les groupes occlusive + R + (w) , de savoir distinguer R labialisé de R non labialisé : on est alors en présence, dans le premier cas, du groupe occlusive + R + w , dans le second du groupe occlusive + R . Incidemment on peut voir, semble-t-il, dans ce comportement de [w] , la raison de la présence d'une seule syllabe dans une suite comme "proie", alors qu'il y en a obligatoirement deux lorsque la consonne vocalique transitoire est [j] , par exemple dans "trio", qui doit se réaliser comme un segment non anticipé sur les consonnes qui précèdent.

2. Les allophones CVR et CSR.

Les locuteurs qui utilisent l'allophone CVR, notamment au contact des occlusives voisées, se partagent en trois classes :

a) ceux qui utilisent l'allophone CSR au contact des consonnes sourdes (locuteurs JO, SL); le groupe pRw se réalise de deux façons :

- soit [w] est manifesté par son trait distinctif sur les consonnes précédentes, comme pour l'allophone CAR (locuteur SL),
- soit [w] apparaît comme un segment indépendant (JO); mais alors, à l'écoute on a l'impression d'une réalisation avec diérèse. Le segment [w] > [u] représente le tiers de la durée de la syllabe w .

b) ceux qui utilisent l'allophone CVR, partout sauf en position implosive devant consonne occlusive sourde (ex : "harpe", locuteur RS, où apparaît CSR): les seuls cas relevés s'apparentent à la réalisation de CAR devant w .

c) ceux qui utilisent l'allophone CVR dans tous les cas (JB, LT). Les seuls exemples relevés montrent la réalisation d'une diérèse sur w qui apparaît comme un segment indépendant de durée égale à la moitié de celle de la séquence [wɑ] .

En résumé, tous les locuteurs, qui à un titre ou à un autre utilisent l'allophone CVR, réalisent la variante CSR au contact de [w] , qui ne se présente comme un segment indépendant que lorsqu'il est accompagné de diérèse. Il se comporte alors comme [u] dans "troua", où la limite morphologique interdit la réalisation de [w] .

CONCLUSION

Le polymorphisme et la nature composite de "R" conduisent, soit à une sur-segmentation, soit à une sous-segmentation phonologiques. Dans cette deuxième hypothèse, le problème essentiel n'est pas tant celui de la segmentation que celui de la distinction des spectres labialisés ou non de "R". L'identification phonologique des unités du groupe consonantique a lieu à un niveau supérieur. La sur-segmentation ne peut être réduite que par référence à des critères phonotactiques du niveau phonologique en concordance avec un faisceau d'indices acoustiques. Le recours à des critères du niveau supérieur est indispensable si l'on veut reconnaître une seule et même unité dans toutes les variantes des allophones de cette consonne polymorphe. L'apparition d'un X-, symbole d'indétermination segmentale, sous la touche de l'expert, tranquillisera la conscience pointilleuse de l'automaticien.

En ce qui concerne l'étiquetage, le signal de BDSONS est accompagné d'une transcription, de type normatif, où toutes les unités phonologiques sont représentées, même si [w] n'est pas réalisé comme un segment indépendant. Dans cette dernière hypothèse, l'étiquetage en macro-classes, phases et attributs, ne fait pas apparaître la consonne [w] . L'utilisation de la BDSONS doit donc obligatoirement avoir recours à cette double information de l'étiquetage pour une évaluation correcte de la suite phonologique et de sa réalisation.

Nous aurions pu proposer une autre interprétation des groupes consonantiques 3 qui présenterait non seulement un grand intéret au niveau phonologique mais qui aurait également une incidence sur la stratégie d'analyse des groupes consonantiques : nous aurions pu en effet interpréter [pR] , [pl] pj , [pw] , [pRw] , etc. comme des consonnes occlusives à détente spécifique; le comportement de [w] dans ces groupes nous conduisait en effet dans cette direction.

BIBLIOGRAPHIE

1 Marchal A., "Coarticulation or coproduction", Revue d'Acoustique, 4; 255-258, 1983.

2 Autesserre D. et Rossi M., "Propositions pour une segmentation et un étiquetage hiérarchisé : application à la base de données acoustiques du GRECO communication parlée", Actes des 14èmes JEP; 147-151, Paris, 1985.

3 Hirst D., "Linearisation and the single segment hypothesis", Grammatical representation, Obenauer and Pollok eds, 87-100, 1985.

4 Chafcouloff M., "Quelques variantes de R en français méridional : étude acoustique", TIPA 10; 105-120, 1986.

5 Kozevnikov V.A. and Chistovich L.A., Speech articulation and perception, JPRS 30543, dep. of commerce, 1965.

6 Mac Allister R., Lubker J.F. and Carlson J., "An EMG study of some characteristics of the Swedish rounded vowels", Journal of Phonetics, 267-278, 1974.

7 Bonnot J.F., "Etude phonétique et phonologique de l'activité EMG labiale et vélaire", Université Lille 3, 1987.

PREMIERE APPROCHE DE SEGMENTATION PAR FILTRAGE MORPHOLOGIQUE

A. BEN SLIMANE et B. ZOUABI

ENIT - IRSIT (TUNISIE)

SUMMARY

Speech waveform plots suggest that simple time domain processing should extract useful signal characteristics. The method presented in this paper is a first approach for automatic segmentation of continuous speech signals, using some transformations of mathematical morphology. These techniques are applied to the speech signal in order to explore the various shapes in the speech waveform. A particular waveform is extracted from peak regions which are also delimited by a morphological window, resulting in morphological filtering. The peak regions selected by this filtering are then added in an appropriate window, giving the average magnitude of the speech signal in the channel defined by the morphological window.

Comparision of these average magnitudes with the original waveforms indicates a strong correspondence with spoken phonetic segments.

INTRODUCTION

L'examen visuel du tracé du signal temporel de parole permet de remarquer que certaines propriétés caractéristiques du signal varient en fonction du temps : Le mode d'excitation varie pour un son voisé ou non voisé, les variations d'amplitude des maxima sont conséquentes de même que les variations de la fréquence fondamentale dans les régions voisées. Le traitement temporel du signal peut donner une représentation des paramètres caractéristiques du signal de parole tels que l'intensité, le mode d'excitation, la valeur de pitch et éventuellement les paramètres du conduit vocal tels que les fréquences des formants [1].

L'étude que nous présentons dans cette communication constitue une première approche de segmentation de la parole. Elle utilise certaines transformations de la morphologie mathématique qui sont appliquées au signal temporel de parole afin d'extraire les formes du signal dans les zones voisines des extrèma. Les amplitudes du signal recueillies au voisinage des maxima, sont moyennées dans une fenêtre de pondération, la courbe obtenue après lissage morphologique permet de donner une meilleure représentation des transitions ce qui améliore la qualité de la segmentation.

Nous présenterons tout d'abord le principe de base du traitement morphologique en rappelant quelques définitions relatives aux transformations utilisées. On exposera ensuite les différentes étapes de la méthode de segmentation qui a été élaborée. Enfin, Ils seront présentés quelques résultats relatifs à la segmentation des signaux de parole arabe.

PRINCIPE

L'idée de base de la morphologie mathématique est de comparer les objets à analyser à un autre objet de forme connue, appelé élément structurant. Pour se faire un élément structurant est choisit, il peut être selon l'image traitée un segment de droite, un cercle ou une sphère qu'on fait déplacer de façon que son origine passe par toutes les positions de l'espace image. Pour chaque position, on fait intervenir un test relatif à l'union, l'intersection ou à l'inclusion de l'élèment structurant avec ou dans les objets qui constituent l'image. Tous les points de l'espace correspondant à des réponses positives sont affectés à l'origine de l'élèment structurant et forment ce que l'on appelle l'image transformée. Cette démarche, qui est la base de toute transformation morphologique, permet de faire ressortir de l'image analysée les caractéristiques pertinentes et d'éliminer par conséquent tous les détails inutiles [2][3][4]. La figure (1) donne les résultats d'une érosion d'un ensemble X par un élément structurant circulaire B_X.

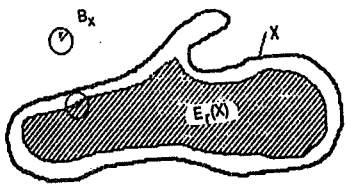


Fig. 1 EROSION

Le traitement morphologique couvre actuellement toutes les applications d'analyse d'image : segmentation, reconnaissance de forme, squeletisation [5][6].

LES TRANSFORMATIONS MORPHOLOGIQUES

Une étude a été déjà élaborée concernant l'adaptation des transformations morphologiques aux signaux temporels [7][8]. Nous rappelons ici les définitions des transformations utilisées :

- érosion et dilatation

Soit un signal échantillonné représenté par x_k ; k = 1,...,N . On définit une fenêtre temporelle de largeur P = L + M ayant son origine placé au point j = L + 1. Cette fenêtre correspond à l'élément structurant et la largeur P = L + M définit la taille de la transformation. Il existe deux transformations de base à savoir l'érosion ou transformation par le minimum et la dilatation ou transformation par le maximum.

- Le signal érodé issu de la transformation par le minimum est défini par l'expression :

$$y_k = \min [x_{k-L},...,x_{k+M}] ; \quad k=L,...,N-M$$

- Le signal dilaté issu de la transformation par le maximum est défini par l'expression :

$$y_k = \max [x_{k-L},..., x_{k+M}] ; \quad k = l, ... , N-M$$

Nous adopterons dans la suite les notations suivantes:
pour l'érosion :

$$y_k = E_p[x_k] \text{ ou } y_k = E_{L,M}[y_k]$$

et pour la dilatation :

$$y_k = D_p[x_k] \text{ ou } Y_k = D_{L,M}[x_k]$$

- OUVERTUR ET FERMETURE

Il est possible d'opérer sur un signal x_k la transformation qui consiste à effectuer une érosion du signal suivie d'une dilatation, dans ce cas on parlera de transformation par ouverture. Le signal transformé y_k est défini par :

$$y_k = D_P[E_P[x_k]] \qquad \text{noté :} \qquad O_p[x_k]$$

La transformation par fermeture consiste à effectuer une dilatation suivie d'une érosion. Le signal transformé y_k est défini par :

$$y_k = E_p[D_p[x_k]] \qquad \text{noté :} \qquad F_p[x_k]$$

Remarque :

Si [L,M] est la fenêtre utilisée pour l'érosion, il fauf alors opérer la dilatation à l'aide de la fenêtre transposée [M,L] :

$$y_k = D_{M,L}[E_{L,M}[x_k]] \quad \text{pour l'ouverture}$$

$$y_k = E_{M,L}[D_{L,M}[x_k]] \quad \text{pour la fermeture}$$

Les transformations par ouverture et par fermeture vérifient les propriétées des filtres morphologiques à savoir : la continuité, la croissance et l'idempotence [10][3].

- Autres transformations

La transformation T telle que :

$$T[x_k] = x_k - O_p[x_k]$$

permet d'extraire les régions des maxima dont la largeur de base est inférieure ou égale à la taille P.

La transformation T telle que :

$$T[x_k] = F_p[x_k] - x_k$$

permet d'extraire les régions des minima dont la largeur de base est inférieure ou égale à la taille P; [11].

SEGMENTATION PAR FILTRAGE MORPHOLOGIQUE

La méthode de segmentation a été élaborée à partir de paramètres extraits d'un détecteur morphologique multicanal que nous avons réalisé. Ce détecteur morphologique a été conçu dans le même esprit que le banc de filtres classique [12][13] à une différence fondamentale, C'est qu'il s'agit d'une sélection de formes et non de filtrage en fréquence. En effet, chaque canal joue le rôle d'un filtre morphologique. Seules sont sélectionnées par le canal les formes d'onde du signal dont la largeur de base correspond à la largeur du canal. Il est à signaler que seul le premier canal est utilisé pour la segmentation, les autres canaux serviront surtout pour la caractérisation des segments phonétiques délimités par le premier canal.

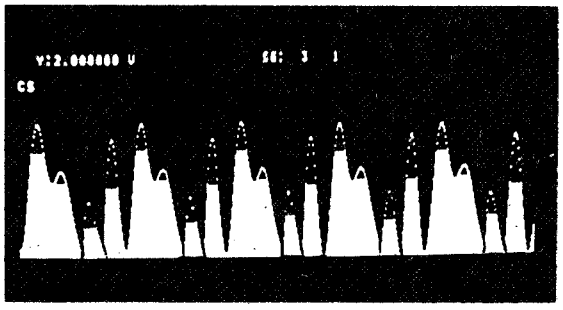


Fig. 2 OUVERTURE

Considérons un canal de faible largeur comprise entre les tailles L1 et L2, ce canal permet de détecter toutes les régions des extrèma dont la largeur de base correspond à [L1,L2]. Pour cela, un filtrage morphologique est opéré sur la valeur absolue du signal en appliquant une ouverture de taille L2. Le signal filtré présente des plateaux de largeur inférieure ou égale à L2 et qui correspondent à toutes les régions des maxima du signal (figure 2). L'extraction des régions des maxima de largeur comprise entre L1 et L2 devient une opération simple après ce filtrage. Il suffit, pour cela, de soustraire le signal ouvert du signal initial et d'éliminer ensuite toutes les régions de largeur inférieure à L1. Pour tenir compte de l'amplitude du signal, on ramène ces régions détectées à leurs amplitudes relatives (figure 3).

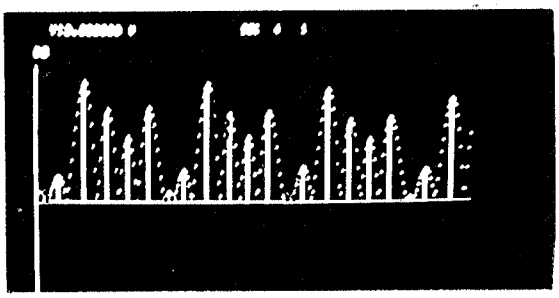


Fig. 3 régions d'extrèma détectées

On effectue ensuite une sommation toutes les 10 ms des échantillons sélectionnés dans le canal. L'ensemble des valeurs obtenues représente l'évolution de l'amplitude moyenne du signal autour des extrèma.

La courbe d'amplitude moyenne obtenue, présente des fluctuations qui peuvent être lissées grâce à un filtre medium réalisé par la composition des deux opérateurs ouverture puis fermeture (OF) [3] de taille égale à 1.

Remarque : La largeur du canal doit être choisie de manière à ne pas détecter les impultions relatives au bruit. Ainsi pour une fréquence d'échantillonnage égale à 10 khz, les tailles adoptées et qui correspondent à un meilleur résultat de segmentation sont comprises entre 2 et 4 périodes d'échantillonnage.

L'observation de la courbe d'amplitude moyenne et du signal temporel, montre une concordance entre les deux signaux. Il est possible d'opérer la segmentation par délimitation des éléments phonétiques à l'aide d'un gradient morphologique.

Le gradient morphologique appliqué à une image à niveau de gris, permet l'augmentation du contraste de l'image, il permet surtout la mise en évidence de contours entre les plages de niveau de gris différents [3][10] ; d'où l'idée d'appliquer le gradient morphologique à la courbe d'amplitude afin d'accentuer les différences de niveau sur cette courbe. Le gradient morphologique est défini par :

$$|\text{grad } A| = \frac{E_1[A] - D_1[A]}{2}$$

A étant le signal d'amplitude. Si le gradient est appliqué à l'aide de l'érosion asymétrique $E_{0,1}$ et de la dilatation asymétrique $D_{0,1}$, il permet la détection des limites des zones montantes. Dans le cas de la détermination du gradient à partir de l'érosion $E_{1,0}$ et de la dilatation $D_{1,0}$ il donne les limites des zones descendantes. Ces deux types de gradient sont utilisés sur la courbe d'amplitude, on obtient alors les limites des zones montantes, des zones stables et des zones descendantes.

La figure (4) donne le résultat de la segmentation du signal de parole pour l'élocution "ASSALOU".

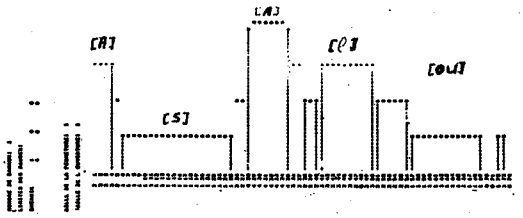


Fig. 4 Segmentation morphologique pour l'élocution "ASSALOU"

CONCLUSION

L'élément structurant associé aux filtres morphologiques utilisés dans la segmentation correspond à une fenêtre temporelle uniforme. Bien que les premiers résultats semblent donner satisfaction, la segmentation peut être améliorée en utilisant des éléments structurants de type non linéaire telsque les élèments semi élliptiques, semi circulaires..., qui ont des propriétées de lissage très intèressantes. D'autre part le filtrage morphologique peut être mieu controlé en utilisant les transformations morphologiques de type conditionnelles ou adaptatives [9].

REFERENCES

[1] B. Gold, L. R. Rabiner, "Parallel Processing Techniques for estimating Pitch Periods of speech in the time domain," J.A.S.A, vol. 46, 1969.

[2] J. Serra, "Introduction à la morphologie mathématique" Cahiers du Centre de Morphologie Mathématique, Ecole des Mines, Fontainebleau, N° 3, 1969.

[3] J. Serra, "Image analysis and mathematical morphology," Academic Press, 1982.

[4] S. Beucher, "Extrema of gray tone functions and mathematical morphology," Ecole des Mines, Fontainbleau, 1983.

[5] F. Meyer, Thèse de Docteur Ingénieur, Ecole des Mines de Paris, 1979.

[6] G. Matheron ,"Les applications idempotentes," Rapport du CGMM, Fontainebleau, 1982.

[7] A. Ben Slimane, "Rapport" ENIT, Tunisie, 1985.

[8] A. Ben Slimane, "Rapport de D E A". ENIT, Tunisie, 1986.

[9] B. Zouabi, N. Ellouze, A. Ben slimane, " Traitement morphologique de signaux unidimensionnels", GRETSI, France, 1987.

[10] M. Coster, J. Chermant, "Précis d'analyse d'images," CNRS, Paris , 1985.

[11] J. Serra, S. R. Sternberg, "Summer School on Mathematical Morphology" Fontainebleau, 1981.

[12] B. Zouabi, A. Guesmi,"Introduction à la Reconnaissance de Parole Arabe", 4e JTEA, Tunisie, 1983

[13] B. Zouabi, N. Rehif, "Système de Reconnaissance Globale de Mots Isolés", Arab School on Sciences and Technology, Maroc, 1983.

DECOMPOSITION TEMPORELLE ET DECODAGE ACOUSTICO-PHONETIQUE

Gérard CHOLLET, Gunnar AHLBOM, Frédéric BIMBOT, Alain VIGIER

ENST Dept. Signal, CNRS UA-820
46 rue Barrault, 75634 PARIS cédex 13, FRANCE.

Abstract

Speech is considered to be an encoded sequence of elementary units (the phonemes of the language). Althought our production and perception mechanisms may at some level use these units, it is difficult to retrieve them automatically from the speech signal (that is to perform an acousic-phonetic interpretation). Each speech unit interacts with its neighbours, creating a phenomenon called "coarticulation". Acoustic-phonetic decoding could be viewed as the process of taking away this coarticulation effect to retrieve the underlying ideal targets associated with phonemes. Such targets may not exist in real speech but they have proved to be useful in acoustic synthesis by rules.

Atal [1] proposed a "temporal decomposition technique" for efficient speech coding. This technique is first applied here to acoustic-phonetic decoding of speech. An interactive technique is developped to infer a limited set of spectral targets and associated interpolation functions from the training samples. Few research axes using this decoding for speech recognition tasks are discussed.

INTRODUCTION

La parole est interprétée comme une séquence de mots, eux-mêmes composés de sons élémentaires (les phonèmes de la langue). Une approche naturelle en reconnaissance de la parole continue est de procéder à une segmentation du signal en unités que l'on compare à des formes de référence. La segmentation en "phonèmes" est difficile car la structure de ces éléments est complexe et dépend du contexte. L'utilisation de segments plus longs (diphones, demi-syllabes, polysons,...) facilite le découpage mais augmente le nombre de comparaisons à effectuer et ne résoud pas tous les problèmes de coarticulation.

ATAL [1] propose une technique de décomposition temporelle du signal de parole afin de coder efficacement l'évolution des paramètres spectraux. Suivant cette formulation, le spectre à l'instant t, $y_i(t)$, est exprimé comme la combinaison linéaire de n spectres g_{ik} (cibles) modulés par des poids $\emptyset_k(t)$ (fonctions d'interpolation compactes).

Les fonctions Ø ont une durée limitée et représentent l'influence réciproque des cibles sur la trajectoire spectrale [2]. Cette décomposition peut être interprètée comme le chevauchement temporel de zones d'influence de phénomènes acoustiques que l'on tente d'associer à des cibles articulatoires [3]. Les fonctions Ø peuvent être décrites par une attaque, une portion stable (éventuellement) et un relâchement.

Nous proposons une technique de localisation et d'étiquetage automatique de la parole continue utilisant une version robuste de la décomposition temporelle. Les cibles spectrales sont évaluées par itérations successives de l'estimation des fonctions Ø et des spectres associés. On montre que les effets de la coarticulation (cibles non-atteintes) peuvent être corrigés par l'inférence de cibles et que la précision de l'étiquetage est de ce fait améliorée par rapport aux techniques classiques (centiseconde et segmentale).

La décomposition temporelle fournit une représentation de l'information qui se prête particulièrement bien à la reconnaissance. Différentes possibilités d'utilisation sont suggérées.

DECOMPOSITION TEMPORELLE

La technique proposée par Atal a été, à l'origine, appliquée au codage de la parole. Pour un segment de parole de durée T, la suite de paramètres spectraux $\{y_i(t)\}$ $(1 \le i \le m, 1 \le t \le T)$ est exprimée en tant que combinaison linéaire de n cibles $\{g_{ik}\}$ $(1 \le i \le m, 1 \le k \le n)$, pondérées par n fonctions d'interpolation $\{\emptyset_k(t)\}$ $(1 \le k \le n, 1 \le t \le T)$:

$$y_i(t) = \sum_{k=1}^{n} g_{ik}\, \emptyset_k(t)$$

Première passe

Une première estimation d'une fonction Ø, sur un intervalle $[t_1, t_2]$ est obtenue grâce à la Décomposition en Valeurs Singulières (SVD) des paramètres spectraux:

$$Y = U D V^t$$

De la matrice U ne sont conservées que les p premières composantes (en général p = 4 ou 5), car elles expriment la majeure partie de la variation des paramètres $y_i(t)$ sur l'intervalle $[t_1, t_2]$.

On maximise ensuite l'énergie des fonctions Ø dans une fenêtre w(t) définie sur l'intervalle $[t_1, t_2]$. Cette maximisation conduit à la recherche de la valeur propre maximale de l'équation

$$R\,b = \mu\,b$$

$$\text{où} \quad R_{ij} = \sum_{t=t_1}^{t_2} w(t)\, u_i(t)\, u_j(t) \qquad 1 \le i \le p,\ 1 \le j \le p$$

Le vecteur propre associé à la valeur propre maximale de l'équation ci-dessus permet de calculer la fonction Ø recherchée sous la forme:

$$\emptyset(t) = \sum_{i=1}^{p} b_i\, u_i(t) \qquad t_1 \le t \le t_2$$

Une fois une fonction Ø déterminée sur l'intervalle $[t_1, t_2]$, avec un éventuel ajustement de l'intervalle, on se décale pour calculer la fonction Ø suivante jusqu'à ce que l'on ait couvert l'intervalle total [1, T]. Au terme de cette première passe on obtient donc une suite de fonctions Ø sur tout le segment à coder. La figure 1 illustre les résultats de cette passe sur le chiffre "five" (base de données TEXAS). On a également représenté les spectres associés à chaque

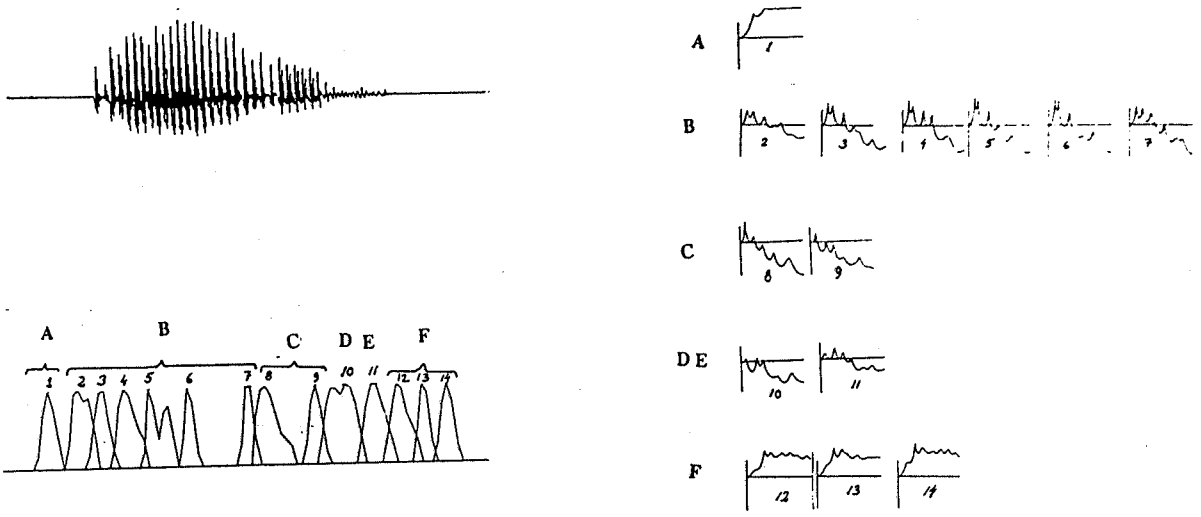


Fig 1: Décomposition temporelle: première passe.

fonction Ø. Les fonctions Ø obtenues au terme de la première passe ne sont pas satisfaisantes: elles sont en nombre trop important et la réunion de leurs supports ne couvre pas tout le segment.

Affinement itératif et regroupement de fonctions

Les cibles spectrales associées à une décomposition temporelle sont alors évaluées par un calcul de pseudo-inverse minimisant l'erreur de reconstruction:

$$G = Y \varnothing^t (\varnothing \varnothing^t)^{-1}$$

de même, les fonctions Ø peuvent être réévaluées par

$$\varnothing = (G^t G)^{-1} G^t Y$$

On procède ensuite au regroupement des fonctions Ø consécutives associées à des cibles proches (somme des fonctions et moyenne des cibles).

Ce calcul est poursuivi de maniére itérative jusqu'à ce qu'il n'apporte plus d'amélioration significative. On peut voir sur la figure 1 quelles fonctions seront regroupées au cours des différentes itérations.

Inférence de cibles

Certaines fonctions Ø ont une portion stable de durée limitée, voire nulle. Il peut s'agir de cibles intermédiaires nécessaires pour modéliser des transitions entre phonèmes spectralement éloignés (par exemple [se] ou [ui]). Dans d'autres cas, il s'agit d'un phénomène de cible non-atteinte.

On peut alors tenter d'inférer la cible visée par le locuteur, en étudiant localement la trajectoire spectrale: la cible intermédiaire est réévaluée en procédant à des régressions linéaires sur le début et la fin de la trajectoire. Une fois cette nouvelle cible inférée, les nouvelles fonctions Ø sont réestimées en minimisant l'erreur de reconstruction ($\varnothing = (G^t G)^{-1} G^t Y$), ou par l'intermédiaire de considérations géométriques.

Les nouvelles cibles obtenues se prêtent mieux à l'étiquetage.

Approximation trapézoïdale

Les fonctions Ø sont approximées par trois segments de droite (critère des moindres carrés). Les cibles sont recalculées une dernière fois. Les fonctions Ø étant systématiquement normalisées à l'unité, leur approximation trapézoïdale permet de limiter l'influence qu'aurait une valeur maximale très élevée et peu représentative du comportement général de la fonction. La figure 2 représente les fonctions Ø et les cibles associées avant et après l'approximation trapézoïdale. Il est à noter que seule une cible a changé de manière significative.

La décomposition temporelle obtenue décrit bien la structure phonétique du signal. Cela se fait au détriment d'une erreur de reconstruction plus importante.

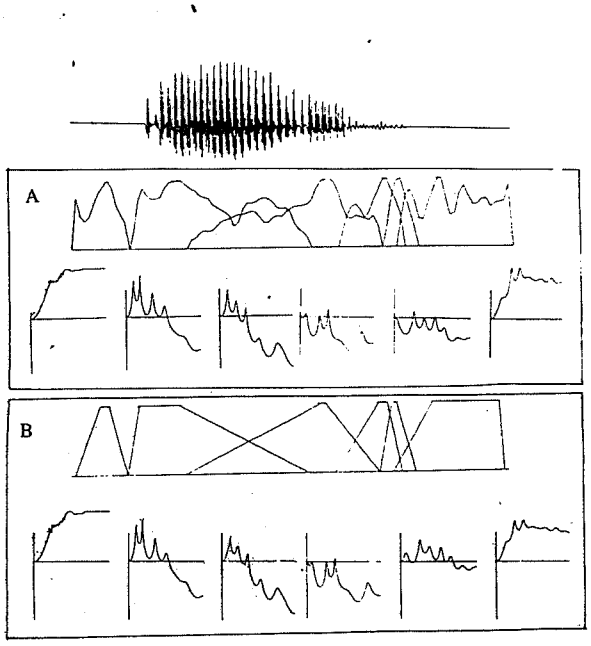


Fig 2: Décomposition temporelle après affinement itératif: A: résultats bruts. B: approximation trapézoïdale.

Jeu de paramètres

Les coefficients LAR ont été retenus pour plusieurs raisons: leur stabilité par combinaison linéaire, leur homogénéité et leur isotropie. Par ailleurs, l'interpolation entre deux jeux de LAR fournit une bonne approximation de transition linéaire entre formants [3].

L'application de cette méthode paraît optimale lorsque les paramètres spectraux sont calculés avec une fréquence suffisante pour rendre compte de rapides transitions du signal (typiquement toutes les 50 ms). Par ailleurs, il faut choisir des paramètres pour lesquels l'interpolation linéaire correspond à une réalité à la fois physique et acoustique: les LAR semblent vérifier ces propriétés.

ETIQUETAGE

A chaque couple $(g_i, \emptyset_i)$ peut être associée une étiquette, en rapport avec la nature phonétique de la cible: phonémique ([a], [f]), allophonique ([R vocalique], [R consonantique]), ou intermédiaire ([s -> e], [u -> i])...

La figure 3 montre la matrice de distance pour deux élocutions du mot "six" de la base de données TEXAS issues d'un même locuteur. Les distances euclidiennes sont calculées pour toutes les combinaisons de cibles.

Cette matrice met en évidence des couples de cibles proches. A des événements phonétiques semblables sont associées des cibles spectrales proches. Cela permet de construire un dictionnaire à partir de cibles de références connues et d'étiqueter une forme test en cherchant dans le dictionnaire l'élément qui lui est spectralement le plus proche. C'est une étape dans le décodage acoustico-phonétique. L'information temporelle décrite par les fonctions d'interpolation est essentielle: la fricative [s] et l'explosion du [t] ont des spectres proches; seules leurs fonctions Ø permettent de les différencier.

Un tel type d'étiquetage permet non seulement de localiser des événements phonétiques, mais également d'en connaître la zone d'influence. Il se démarque en cela à la fois des étiquetages centiseconde et segmental.

RECONNAISSANCE ET DECOMPOSITION TEMPORELLLE

A chaque entité obtenue par décomposition temporelle on peut associer un doublet de vecteurs: la fonction d'interpolation $\emptyset_k$ qui contient l'information temporelle et le vecteur spectral g_k.

Par rapport aux méthodes classiques de reconnaissance, on rajoute un étage après l'analyse LPC puisque les observations ne sont plus les paramètres directement issus de cette analyse mais les résultats de la décomposition temporelle. On bénéficie donc des avantages de cette décomposition: des paramètres en nombre plus faible, calculés non uniformément dans le temps et des unités phonétiques décorrélées.

Programmation dynamique

La matrice de distances de la figure 3 peut être utilisée en programmation dynamique. Les faibles distances au voisinage de la diagonale laissent espérer de bons scores en reconnaissance de mots isolés.

Reconnaissance par étiquetage

La phase d'apprentissage d'un tel système consiste à calculer, pour chaque unité phonétique de référence ($i^{ème}$ étiquette), un doublet $(g, \emptyset)_i$. Chaque mot de référence est représenté par une suite d'étiquettes caractéristiques.

Les mots tests sont décomposés en doublets. Pour chacun d'entre eux sont calculées toutes les distances aux références $(g, \emptyset)_i$. L'étiquette associée au doublet test est celle de la référence qui donne une distance minimale. Reconnaître un mot consiste alors à comparer la suite d'étiquettes obtenues avec celles des mots de référence, et à choisir celle qui lui ressemble le plus.

Lien avec les modèles de Markov

Niranjan [5] a fait la comparaison entre la décomposition temporelle et les chaînes de Markov. Il établit une correspondance directe entre une fonction compacte et un état. Dans le cadre d'un tel modèle les observations sont les doublets $(g, \emptyset)$.

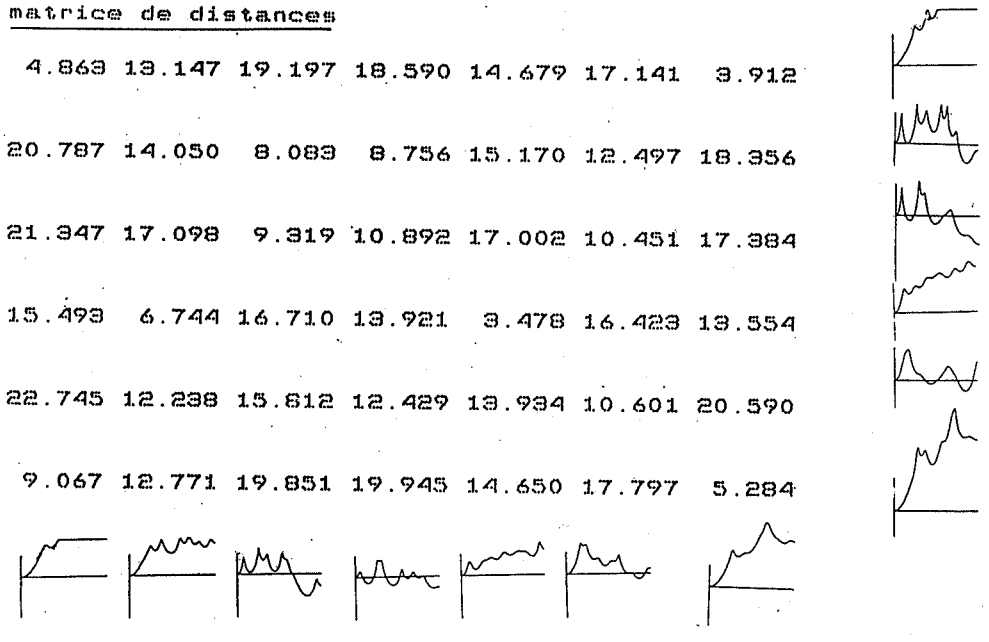


Fig 3: Distances entre cibles de 2 élocutions du même mot [siks], prononcées par le même locuteur.

En reconnaissance de mots isolés [6], un modèle serait construit pour chaque mot, lors de la phase d'apprentissage. Chaque mot test serait par la suite classifié comme étant celui auquel est associé le modèle d'apprentissage de probabilité maximale

La phase d'apprentissage demandant toujours un nombre minimal d'observations, il est clair qu'elle se fera avec beaucoup plus d'unités phonétiques. Les résultats ne devraient qu'en être améliorés.

Filtrage inverse des tests par les références

Le polyson, défini en tant qu'unité segmentale comprise entre deux zones de stabilité spectrale, peut être utilisé comme unité de reconnaissance. Il correspond à une suite de cibles spectrales.

Partant de points d'ancrage reconnus sur le signal (zones stables), on utilise comme références $\{G_i^{ref}\}$ l'ensemble des polysons ayant mêmes extrémités que le segment à reconnaître. Pour chaque G_i^{ref}, on calcule les fonctions $\emptyset_i^{test}$ par filtrage inverse:

$$\emptyset_i^{test} = (G_i^{ref\,t} . G_i^{ref})^{-1} G_i^{ref\,t} Y^{test}$$

où Y^{test} est la matrice des paramètres spectraux entre les deux points d'ancrage.

On choisit la matrice $\emptyset^{test}$ qui minimise l'erreur de reconstruction $Y^{test} - G^{ref} \emptyset^{test}$, en s'étant au préalable assuré de la compacité des fonctions $\emptyset$ de $\emptyset^{test}$.

Le nombre de segments de références $\{G_i^{ref}\}$ est de l'ordre de 7000. On peut cependant pré-calculer et stocker leurs pseudo-inverses. Cette technique pourrait être mise à profit pour du codage à très bas débit (vocodeur à environ 100 bits par secondes).

CONCLUSION

Nous avons décrit une version robuste de la décomposition temporelle. Cette technique s'insère particulièrement bien entre une analyse LPC et des algorithmes de reconnaissance. Elle fournit de plus une représentation pertinente du signal qui peut aussi être utilisée en synthèse.

Références

[1] ATAL (1983) **Efficient coding of LPC parameters by temporal decomposition.** Proc. ICASSP-83, 2.6, pp 81-84.

[2] CHOLLET, GRENIER, MARCUS (1986) **Temporal decomposition and non-stationary modeling of speech.** EUSIPCO, La Hague.

[3] BIMBOT, AHLBOM, CHOLLET (1987) **From segmental synthesis to acoustic rules using temporal decomposition.** XIth Int. Cong. Phon. Sciences, Tallinn.

[4] AHLBOM, BIMBOT, CHOLLET (1987) **Modeling spectral transitions using temporal decomposition techniques,** Proc ICASSP-87, Dallas.

[5] NIRANJAN, FALLSIDE (1987) **On modelling the dynamics of speech patterns** ECSP, 1987, Edimbourg.

[6] RABINER, LEVINSON, SONDHI (1983) **On the application of vector quantization and hidden Markov models to speaker-independant, isolated word recognition,** Bell Technical Journal, Vol 62, n° 4.

RECONNAISSANCE DES FORMES ET SEGMENTATION

Méloni Henri, Bulot Rémi

Groupe d'Intelligence Artificielle
Faculté de Luminy
70 route Léon Lachamp 13288 Marseille Cedex 9

Abstract

We present an approach in Prolog II of the segmentation of the speech signal into pseudo-phonetic events and their labelling. The environment makes use of several techniques, such as signal processing, pattern recognition, automatic learning, knowledge representation and problem solving. Some pattern schemes of parameters and their relationships define units wich correspond to significant portions of phonemes in specific contexts.

I - Introduction

La segmentation du signal de parole en unités pseudo-phonétiques diverses est l'un des premiers problèmes abordés par les chercheurs en Reconnaissance Automatique de la Parole. De nombreuses solutions ont été apportées, aussi bien pour délimiter et étiqueter des intervalles spécifiques associés à des portions de sons connus [1] que pour localiser et identifier les phonèmes en situation de reconnaissance [2]. Les techniques les plus efficaces consistent généralement à déterminer les portions stables et instables du signal à partir de fonctions mesurant les variations temporelles de l'énergie dans différentes zones du spectre. Toutefois les méthodes algorithmiques sont fréquemment inefficaces pour séparer des sons acoustiquement proches tels que des voyelles, des semi-voyelles ou des consonnes vocaliques situées dans une même syllabe.

Nous disposons d'un environnement sous PROLOG II [3] dans lequel on peut définir et traiter des connaissances acoustiques et phonétiques complexes (paramètres, formes, relations, etc.) utilisables pour désigner les limites de nombreux phénomènes caractéristiques des sons du français. La souplesse et la précision des outils disponibles permettent de décrire des événements acoustiques et phonétiques ainsi que les contextes spécifiques dans lesquels ils sont pertinents. Le problème de la segmentation se ramène donc, dans ce cadre, à l'utilisation optimale des informations disponibles au moyen d'une stratégie définie par un ensemble de règles.

L'ensemble des règles qui décrivent les paramètres, les formes, les relations et les événements s'est révélé efficace aussi bien pour la segmentation et l'étiquetage de phrases connues [4] que pour limiter et caractériser les phonèmes et les syllabes en situation de reconnaissance. Les schémas de formes permettent d'éviter de recourir à l'utilisation de seuils absolus sur les paramètres et rendent les règles beaucoup plus indépendantes des locuteurs.

II - Paramètres acoustiques et phonétiques

Pour distinguer les subtiles variations apportées par un phonème dans de nombreux environnements, nous avons été conduits à décrire une grande quantité de paramètres acoustiques et phonétiques. Certains d'entre eux sont calculés systématiquement tandis que d'autres sont évalués lors de l'activation d'une règle qui les utilise. De plus, les conditions de calcul des paramètres peuvent être dynamiquement modifiées en cours de traitement en fonction de connaissances diverses identifiées dans le contexte.

II.1 - Paramètres standard

Les paramètres standard sont évalués sous PROLOG chaque 10 ms à partir du signal et de spectres lissés obtenus au moyen des coefficients de prédiction linéaire [5]. Le vecteur centiseconde est constitué de 22 valeurs qui permettent de décrire efficacement le signal pour les différents types de sons du français. Les conditions de calcul de ces paramètres sont dynamiquement modifiables en fonction de connaissances contextuelles diverses qui déterminent, au moyen de règles, certains arguments du prédicat évaluable qui effectue l'analyse du signal sur un intervalle donné. Les attributs et leurs conditions d'évaluation sont mémorisés pour un ensemble de trames et les éventuels calculs différents obéissent au processus général de backtracking de PROLOG. Les prédicats évaluables qui accèdent aux paramètres standard pour effectuer divers types de traitements utilisent donc des valeurs correspondant toujours aux conditions dans lesquelles les règles les ont contraint d'opérer.

II.2 - Paramètres temporaires

Les paramètres standard ne suffisent pas à rendre compte de tous les phénomènes acoustiques et phonétiques et nous avons défini un grand nombre de paramètres temporaires qui sont conservés sur un intervalle jusqu'à ce qu'une nouvelle règle utilise un autre paramètre provisoire sur une portion de cette même zone [5]. Ces nouveaux attributs, décrits au moyen de règles, sont calculés à partir des paramètres standard et donc soumis à un second niveau de backtracking.

Les paramètres temporaires les plus simples sont évalués au moyen de combinaisons de paramètres standard, de scalaires ou d'autres attributs provisoires. Des prédicats évaluables permettent non seulement d'effectuer les opérations classiques sur les paramètres (addition, différence, produit, division, homothétie, etc.) mais également de réaliser des décalages temporels ou de calculer des fonctions d'instabilité. Ces paramètres, définis de manière déterministe, rendent compte de la plupart des phénomènes acoustico-phonétiques concernant la segmentation du signal et nous permettent de caractériser contextuellement des événements peu marqués. Cependant, certains attributs discontinus tels que les pics spectraux conduisent à des évaluations indéterministes de paramètres importants (trajectoires d'un pic, formants, affaissement d'un pic ou d'un formant, etc.).

III - Reconnaissance de formes

L'évolution temporelle des paramètres est schématisée par des formes. Certaines indiquent de manière régulière la présence d'un phénomène acoustique particulier dans un environnement plus ou moins précis. Nous privilégions toujours, lorsque cela est possible, les formes qui caractérisent la totalité d'un événement depuis son apparition jusqu'à sa disparition ; elles sont généralement modélisées par des collines et des vallées. Les positions relatives de plusieurs formes ainsi que certaines de leurs associations constituent des informations déterminantes pour la description des événements acoustico-phonétiques. En plus des limites d'un phonème, ces connaissances permettent d'affecter aux segments des étiquettes pseudo-phonétiques. Les phonèmes et les syllabes sont considérés comme des événements phonétiques particuliers regroupant des unités plus atomiques.

III.1 - Définition des formes pertinentes

L'ensemble des outils proposés dans notre environnement [4] permet de symboliser des schémas de formes qui caractérisent l'allure globale d'une infinité de courbes. A partir des prédicats évaluables de base, nous décrivons, pour divers paramètres, tous les schémas qui correspondent à des phénomènes pertinents dans un contexte donné. Pour la segmentation, seuls les modèles de collines et de vallées sont utilisés ; chaque classe de formes correspond à une instantiation particulière des arguments d'un schéma (seuil de bruit, émergences à gauche et à droite, micro-variations tolérées, largeur minimale). Ainsi, les événements acoustiques significatifs déterminés par le paramètre mesurant la densité des passages par zéro du signal correspondent aux 3 types de collines présentés dans la figure 1.

III.2 - Relations temporelles entre les formes

Les formes et les événements acoustico-phonétiques s'organisent temporellement suivant des relations significatives de la réalisation des sons. Les fonctions les plus utilisées permettent d'évaluer, pour deux intervalles, leur adjacence, leur coïncidence ou l'inclusion totale ou partielle de l'un dans l'autre. Ainsi, le phonème /t/ sera généralement caractérisé, dans un grand nombre de contextes, par une importante vallée d'énergie suivie d'une

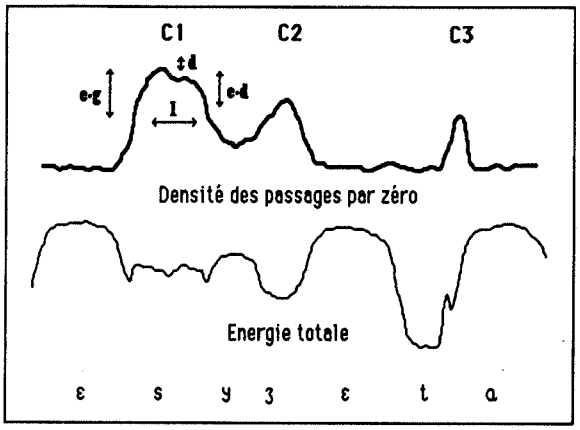


Figure1 : schémas de collines caractérisant des événements fricatifs

petite colline d'énergie coïncidant avec une colline étroite de densité des passages par zéro (figure 1). Pour d'autres phonèmes, placés dans des contextes peu favorables, leur trace est beaucoup plus difficile à saisir et n'est décelable qu'au moyen de relations complexes sur des paramètres particuliers.

IV - Evénements acoustiques et phonétiques

Représentés au moyen des formes de base, les événements acoustiques caractérisent l'évolution temporelle d'un paramètre ou la conjonction de plusieurs propriétés acoustiques du signal. Ils déterminent généralement des segments infra-phonémiques et ne reçoivent pas d'interprétation phonétique. De plus ils font rarement l'objet d'une définition spécifique au moyen de règles mais apparaissent implicitement dans les clauses décrivant les événements phonétiques. Leur fonction essentielle est la délimitation et l'étiquetage de zones du signal sur lesquelles un phénomène apparaît, se déroule, s'amplifie ou disparaît. On peut également envisager des événements qui regroupent plusieurs formes et événements coïncidents ou successifs.

Les événements phonétiques, identifiés à partir des événements acoustiques, des formes et des relations, constituent des unités que l'on peut associer directement à des phases spécifiques de phonèmes et de transitions (constriction, occlusion, explosion, etc.) ou à des regroupements de segments acoustiquement proches. Des règles contextuelles réunissent ensuite ces éléments pour désigner les limites des phonèmes ou décomposent certains d'entre eux à partir de critères plus fins pour séparer les voyelles des consonnes vocaliques qui les entourent (figure 2). Les clauses qui définissent ces connaissances opèrent dans des contextes souvent très différents suivant qu'il s'agit d'événements "évidents" ou de segments tributaires de l'identification préalable de l'environnement. Ces règles sont indépendantes du locuteur et opèrent une partition peu ambiguë d'un énoncé en macro-classes pseudo-phonétiques.

Les événements phonétiques sont parfois hiérarchisés lorsque certains d'entre eux sont décomposés en unités plus courtes. C'est le cas notamment lorsqu'un noyau vocalique long est segmenté plus finement au moyen d'événements acoustiques divers (succession de portions monotones de formants, coïncidence de zone stables de certains pics, formes particulières d'un paramètre, etc.). Nous avons défini plus d'une centaine

d'événements phonétiques qui caractérisent contextuellement les phonèmes et portions de phonèmes du français (plusieurs types de noyaux vocaliques et consonantiques, des explosions, des constrictions, des occlusions, des transitions, etc.). Par exemple, un type d'événement vocalique peut être caractérisé dans certains environnements par une importante colline d'énergie totale sur laquelle coïncide une montée d'énergie dans les basses fréquences. Ces propriétés sont traduites au moyen de la règle suivante :

```
evenement-voc(<voc(5), z> ) ->
   forme(<colline1-er0, z> )
   inferieur(5, longueur(z))
   voise(z)
   ou (coincidence-sur(z, colline1-ebf) ,
       coincidence-sur(z, colline1-ap1) ) ) ;
```

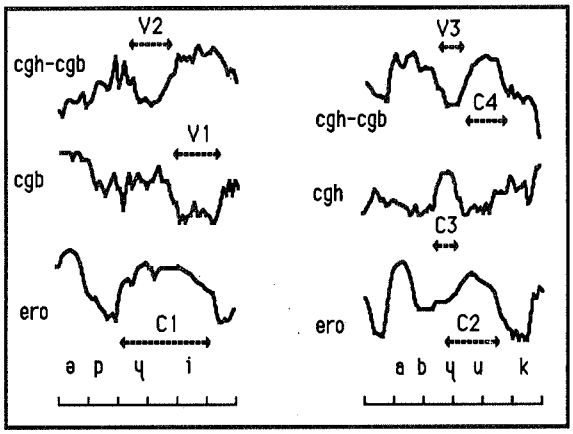


Figure 2 : Segmentation à l'intérieur de groupes vocaliques

Les phonèmes sont projetés sur les différentes macro-classes pseudo-phonétiques de façon quelquefois indéterministe. Les rares ambiguités interviennent, dans des contextes particuliers, entre les consonnes nasales et les occlusives sonores ou entre les voyelles et les semi-consonnes.

Des techniques d'apprentissage automatique à partir d'exemples ont permis d'établir des règles contextuelles de segmentation du signal en unités phonétiques [6]. Le modèle synthétisant les régularités sur l'ensemble d'exemples est obtenu à partir de descriptions exhaustives de ces derniers en termes d'attributs symboliques (traits, phonèmes, syllabes), de formes et de leurs relations évaluées sur un intervalle donné.

V - Segmentation

Les formes et les événements acoustico-phonétiques constituent des unités qui délimitent les traces de phonèmes ou de portions de phonèmes dans un énoncé. On peut utiliser ces connaissances aussi bien pour étiqueter une base de sons connus que pour effectuer la reconnaissance d'une phrase. dans chacune de ces situations, des règles de stratégie définissent une utilisation optimale des informations mémorisées ou déduites.

V . 1 - Segmentation d'un énoncé connu

L'étiquetage d'un corpus enregistré peut conduire à localiser dans le signal des unités très variées (événements acoustico-phonétiques, phonèmes, phones, syllabes, etc.) [7]. Notre travail, dans ce cadre, s'est borné à délimiter les portions stables des phonèmes et les explosions des consonnes occlusives. Ces indications sont utilisées pour effectuer des statistiques et l'apprentissage de concepts sur une base de sons du français.

V . 1 . 1 - Connaissances utilisées

Si nous choisissons convenablement les paramètres standard et provisoires, chaque phonème, dans un environnement particulier, est marqué par une séquence d'associations de formes. Nous avons défini statistiquement la valeur moyenne d'un paramètre pour chacun des phonèmes de manière à déduire immédiatement les schémas attendus pour un phonème dans un contexte précis. Il suffit ensuite de rechercher, dans une zone précisée, les formes correspondantes au moyen des prédicats prédéfinis. Les informations supplémentaires prises en compte concernent la position du dernier son localisé et la durée moyenne du phonème traité.

V . 1 . 2 - Stratégie pour l'étiquetage d'énoncés connus

De manière générale, la recherche des événements acoustico-phonétiques associés aux phonèmes s'effectue de la gauche vers la droite à travers une fenêtre temporelle dont la longueur est calculée en fonction des connaissances disponibles sur l'unité phonétique traitée et sur le contexte dans lequel elle apparait (phonème environnant, position dans la phrase et dans la syllabe, vitesse d'élocution, etc.).

Lorsque les formes recherchées ne se manifestent pas de façon évidente, nous sommes amenés à positionner la fenêtre de travail sur un événement interprétable sans ambiguité pour ensuite effectuer un découpage de la droite vers la gauche jusqu'au dernier phonème localisé.

Cette utilisation descendante des connaissances peut convenir également en situation de reconnaissance pour associer temporellement aux événements acoustico-phonétiques les phonèmes des mots d'une cohorte proposée par la stratégie.

Dans certaines circonstances (rapport signal/bruit insuffisant, phonèmes adjacents acoustiquement proches, ...), des ambiguités demeurent. Le système sollicite alors, de manière interactive, un positionnement d'un curseur sur les trames jugées convenables par l'utilisateur. Il poursuit ensuite la segmentation au delà de l'intervalle pris en compte.

V . 2 - Segmentation en reconnaissance automatique

Dans le cadre de la RAP, les connaissances concernant les limites et la nature des événements acoustico-phonétiques peuvent être utilisées dans des circonstances diverses (de manière remontante ou descendante, sans informations connues ou déduites sur le contexte, après identification des segments adjacents, etc.). Cependant, il existe une hiérarchie des connaissances qui induit une stratégie naturelle de recherche des segments et que l'on peut résumer au moyen de la règle :

segmentation ->
tous-les(événement-vocalique-évident)
tous-les(événement-consonantique-évident)
tous-les(événement-consonantique-secondaire)
tous-les(événement-vocalique-secondaire)
tous-les(sous-segment-vocalique-intégral) ;

V.2.1 - Localisation des événements vocaliques évidents

Les divers événements vocaliques évidents sont désignés par la coïncidence de plusieurs types de collines apparaissant sur des paramètres standard ou provisoires. Il s'agit généralement de schémas très marqués sur des attributs tels que l'energie totale, l'énergie basse fréquence, l'énergie spectrale utile, l'amplitude ou l'émergence du premier formant, les différences d'énergies entre certaines zones du spectre, ... Ces unités, mémorisées dans un treillis, sont toutes désignées par la même étiquette pseudo-phonétique. Elles sont recherchées dans la phrase de la gauche vers la droite et correspondent à la partie vocalique des syllabes (voyelle + certaines semi-voyelles et consonnes vocaliques). Les quelques règles qui définissent les événements vocaliques évidents n'induisent aucune erreur dans le processus de segmentation de phrases énoncés dans des conditions normales.

V.2.2 - Localisation des événements consonantiques évidents

Comme pour les événements vocaliques évidents, ces unités sont décrites par des règles qui n'utilisent pas des phénomènes contextuels déjà identifiés. Leur définition est cependant plus complexe et conduit à la mémorisation de plusieurs types de segments (constrictifs, occlusifs, cosonantique).

Les événements constrictifs sont repérés au moyen de collines importantes sur des paramètres tels que la densité des passages par zéro ou la différence entre l'énergie dans les hautes fréquences et celle des basses fréquences. Ces formes coïncident dans la plupart des cas avec une vallée de l'énergie utile du spectre. Ces unités acoustico-phonétiques désignent, suivant les caractéristiques des formes identifiées, des portions de consonnes fricatives sourdes et sonores ou des phénomènes transitoitres associés à la réalisation contextuelle de certains sons. Dans le cas des consonnes fricatives, nous étendons l'intervalle aux éventuelles petites vallées d'énergie qui l'entourent comme sur l'exemple présenté dans la figure 3.

Les événements occlusifs sont définis par une importante vallée de l'énergie suivie d'une explosion que l'on détecte à partir de collines particulières pour des paramètres tels que l'énergie totale, l'énergie utile, la densité des passages par zéro, les centres de gravité, la différence entre l'énergie spectrale et l'énergie basse fréquence, etc. L'occlusion est marquée également par le buzz que l'on caractérise au moyen de phénomènes marquant une élévation relative de l'énergie dans les basses fréquences (figure 3). Si la plupart des occlusives sourdes sont correctement localisées, la consonne /b/ et quelquefois les consonnes /d/ et /g/, dont les explosions peuvent être peu marquées dans certains contextes, sont étiquetées au moyen d'une autre macro-classe.

Les événements consonantiques évidents désignent des phénomènes acoustico-phonétiques qui n'ont pu être assignés à l'une des deux classes précédentes. Ils sont désignés par la coïncidence de vallées pour les divers paramètres qui mesurent l'énergie du spectre dans des zones privilégiées.

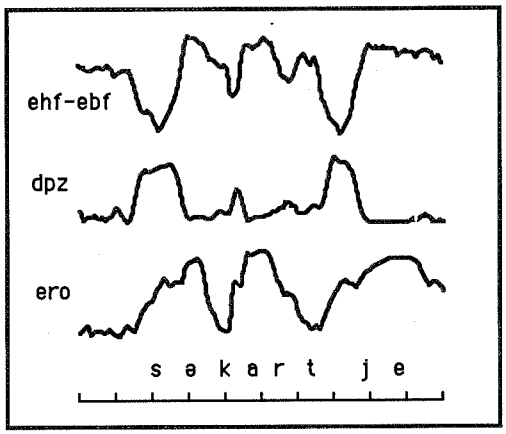


Figure 3 : coïncidences de collines et vallées pour des consonnes

V.2.3 - Localisation des événements secondaires

Après la localisation des événements acoustico-phonétiques évidents et leur mémorisation dans le treillis, il demeure des zones du signal qui n'ont pas été interprétées en termes de macro-classes. Pour ces segments, nous disposons d'un ensemble de règles qui utilisent les connaissances déduites afin de proposer, dans la mesure du possible, une caractérisation acoustico-phonétique. Les clauses sont nombreuses et utilisent parfois des informations complexes. Elles permettent de délimiter des phonèmes ou des transitions apparus dans des contextes spécifiques. Par exemple, les voyelles fermées /i/ et /y/, situées entre deux consonnes fricatives peuvent ne pas apparaître de manière évidente. Cependant, l'identification des segments constrictifs adjacents permet de reconnaître à l'intervalle médian la qualité d'événement vocalique secondaire à partir d'indications distinctes de celles utilisées pour les événements vocaliques évidents et spécifiques à cet environnement. A la suite de cette phase, très peu de "trous" subsistent dans le treillis ; ils ne correspondent généralement pas à des phonèmes "oubliés" mais à des transitions difficilement classifiables.

V.2.4 - Localisation des voyelles dans les noyaux vocaliques

Pour chaque événement vocalique évident ou secondaire nous effectuons une segmentation plus fine pour désigner la portion du noyau correspondant à la voyelle ainsi que les zones qui pourraient indiquer la présence de semi-voyelles ou de consonnes telles que /r/ ou /l/.

La localisation d'une zone caractéristique de la voyelle est relativement aisée à l'exception des voyelles fermées (/i/, /y/, /u/) dans certains contextes (suivies de /r/ dans la même syllabe). Les règles utilisent des connaissances relatives à la stabilité des deux premiers formants ainsi que les positions des maxima des collines pour certains paramètres (énergie basse fréquence et totale, amplitude et émergence du premier formant, différences entre diverses énergies, centres de gravités, etc.).

La recherche d'événements dans le noyau vocalique que l'on peut interpréter comme des phonèmes distincts de la voyelle est délicate et très dépendante du contexte. Nous limitons la segmentation "remontante" aux seuls phénomènes nettements marqués par la présence de formes importantes (collines ou vallées) sur des paramètres standard et provisoires significatifs. Dans de nombreux environnements, les semi-voyelles sont localisées au moyen d'attributs simples (figure 2). Les interprétations trop ambiguës sont reportées à une étape ultérieure où seront disponibles des informations plus précises sous la forme d'hypothèses phonémiques à discriminer.

VI - Conclusion

La technique de segmentation que nous proposons à donné des résultats satisfaisants aussi bien dans le cadre de l'étiquetage d'une base de sons que pour la partition d'un énoncé en macro-classes acoustiques et phonétiques. L'efficacité du système résulte de la souplesse des outils définis sous Prolog II qui permettent de coder des connaissances très complexes sous une forme synthétique et naturelle. Les informations utilisées ne se réfèrent pratiquement pas à des valeurs de paramètres mais à des schémas de formes à leurs relations qualitatives. Celà induit une certaine indépendance à la fois du locuteur et des conditions de saisie de l'énoncé.

En situation de reconnaissance, on peut envisager d'employer les connaissances décrivant les événements acoustiques et phonétiques pour produire de manière remontante des unités donnant accès à des cohortes de mots d'un lexique, ou pour localiser avec précision de manière descendante certains phénomènes devant apparaître dans un contexte connu.

On peut envisager également ce type de segmentation pour délimiter des unités plus importantes comme les mots ou les groupes prosodiques.

Bibliographie

[1] J.E.P. Lannion 1972, pages 317-391

[2] J.E.P. Paris 1985, pages 127-172

[3] Méloni H., Bulot R.
A knowledge based system for acoustic and phonetic decoding of continuous speech ; Congrès International d'Intelligence Artificielle de Marseille, décembre 1986

[4] Gibelli M., Soussi N.
Segmentation du signal de parole ; Mémoire de DEA, Faculté des sciences de Luminy Université d'Aix-Marseille II, septembre 1986

[5] Méloni H., Bulot R.
Paramétrisation du signal et reconnaissance des formes pour le décodage acoustico-phonétique enProlog ; Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle, Antibes 1987

[6] Guizol J.
Apprentissage inductif de règles pour le décodage de la parole Congrès AFCET Systèmes Experts, Avignon 1987

[7] J.E.P. Aix-en-Provence 1986, pages 171-186

PRESENTATION DE LA COMMISSION "ETIQUETAGE LARGE" DU GRECO "COMMUNICATION PARLEE"

COLLECTIF

ABSTRACT

This paper is an introduction to the goals and works of the "Coarse Labelling" commitee of the GRECO "Communication Parlée", bound to work on the data base BDSONS.

INTRODUCTION

Ce papier a pour but de présenter la Commission "Etiquetage Large", mise en place au Printemps 86 par le bureau du GRECO "Communication Parlée" sous la responsabilité de L. MICLET. Etant donné que ce texte est rédigé fin juin, et qu'à cette date trois réunions, seulement (?) ont été tenues, il est évident que des conclusions définitives sont loin de pouvoir être encore tirées. Cependant, un certain consensus sur les méthodes s'étant dégagé, ainsi qu'une définition du travail à effectuer, ce texte aura j'espère une valeur d'information sur les activités du GRECO, qui concernent toute notre Communauté.

1. SITUATION DE L'ETIQUETAGE DIT "LARGE"

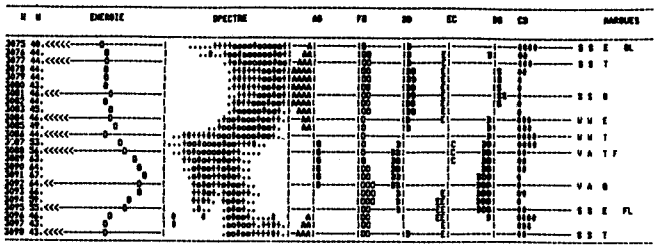
On sait que le GRECO développe une activité d'étiquetage dit "fin", sous deux aspects, temporel et fréquenciel [ABR 85]. Cette activité, de longue haleine, a pour but de fournir sur des bases de données sonores du GRECO (BDSONS) une information aussi exhaustive que possible, telle que des phonéticiens professionnels sont capables de la décrire. Un exemple d'étiquetage temporel fin est donné en figure 1 : on y voit avec quelle précision le signal est décrit, et quel énorme intérêt peuvent tirer les différents utilisateurs d'une base de connaissance résultant d'une telle expertise.

Cet étiquetage fin passe par une première phase "large", dont les principes sont indiqués dans les documents cités en références [MEF 87], [CON 86].

La Commission Etiquetage Large se devait donc de se coordonner avec l'activité précédente, bien que les buts en soient légèrement différents.

La nécessité de l'Etiquetage Large en tant que tel, et non pas comme première étape à une décision plus fine, provient en premier lieu des utilisateurs de BD SONS qui pratiquent le décodage acoustico- phonétique automatique - c'est-à-dire qui font de la reconnaissance de la parole.

- *Exemple n°3 :* /S W A/ dans /swasan(t)/ échantillons 3075-3098



1. 3075-3085 fricative /S/ de caractéristiques analogues à celles qui ont été vues avant. [S S F G L] [S S I] [S S O 2]

2. 3086-3087 un segment de transition apparaît, les formants présentent une inflexion, l'indice CD est instable. [W W I]

3. 3088-3094 tenue et coda de la voyelle orale /A/.
La première phase est fusionnée (enchevêtrement spectral) avec la semi-voyelle. [V A I F] L'ensemble est plutôt grave et ouvert. [V A O]

Fig. 1 (MEF 87)
Exemple d'étiquetage fréquentiel fin

Tous ceux qui ont travaillé dans le domaine, ont fait, nolens volens, de l'étiquetage large, de manière parfois semi automatique, souvent manuelle. Les buts en sont pour eux évidents : il faut réunir une base de données (de type centisenconde, phonèmes ou diphones) pour l'apprentissage et le test du système de décodage automatique - elle doit être bien sûr étiquetée ; c'est-à-dire que chaque élément doit porter un petit badge indiquant comment le système "devrait" le reconnaître. L'ensemble de ces badges (étiquettes) forme le système phonétique (si on ose dire) du programme de reconnaissance. Il va sans dire que chacun a le sien, puisqu'il est défini à des fins propres, sans vrai besoin de références extérieures - dans un premier temps.

Néanmoins, le besoin de comparer les performances des systèmes, voire de les faire collaborer, implique de plus en plus une uniformisation, ou tout au moins une compatibilité de ces ensembles de codes. La mise à disposition des laboratoires de BDSONS induit naturellement son étiquetage large : les mêmes données décrites par les mêmes étiquettes, voilà une solide base d'échange d'informations et de résultats.

Pourquoi l'étiquetage fin de BDSONS ne pourrait-il pas servir à cet effet ? D'abord parce qu'il est prévu sur une longue durée ; ensuite, qu'il est trop riche pour ce que savent faire les programmes de décodage automatique. Enfin, ceux-ci fournissent en général un treillis phonétique non exhaustif, destiné à un décodage lexical et syntaxique, et non pas une description intégrale du contenu acoustique du signal : les évaluer avec les étiquettes "fines" serait les mettre en dehors de leur champ délibéré d'action.

D'autres utilisations sont envisagées pour ces étiquettes : d'abord une fonction de repérage dans BD SONS, indispensable à tous les chercheurs quelqu'ils soient. C'est trop évident pour qu'on y insiste largement - mais c'est capital. Ensuite, éventuellement, l'aide à des tâches du type construction de dictionnaire de diphones, pour la synthèse, par exemple.

2. LES OPTIONS RETENUES PAR LA COMMISSION

Donnons tout de suite les principales caractéristiques de l'étiquetage large tel que la commission l'a défini, avant d'expliciter le pourquoi de ces choix.

- On cherche à disposer rapidement d'une large base d'apprentissage et de test pour la reconnaissance.
- On n'étiquette pas tous les phonèmes de la base, mais seulement ceux qui sont facilement "repérables". Les autres sont simplement indiqués.
- On étiquette ce qu'on entend et ce qu'on voit ; il ne s'agit pas d'une transcription de la cible, mais de la réalisation.
- On se limite aux symboles de l'API.
- On définit un format d'étiquetage normatif ; chacun étiquette ensuite de la manière qu'il veut (à la main, ou par programme de cadrage).
- On peut donner des indications supplémentaires ; mais le principe est de n'étiqueter que ce qui est sûr.
- On ne fait pas de segmentation.

Ces principes diffèrent légèrement de ceux définis pour la première étape de l'étiquetage fin ; en particulier sur le fait de n'étiqueter qu'une partie des phonèmes - ceci nous a semblé un comprois acceptable entre les équipes pour éviter au maximum la part d'interprétation (ou d'expertise) de l'étiqueteur. Mais cela ne gêne pas la fonction d'adressage de BDSONS, puisque un phonème dont ne peut pas indiquer un endroit précis de réalisation est néanmoins en général encadré par deux points précis : les "centres" de ses voisins ; il est alors facile d'examiner le signal à la loupe sur l'endroit litigieux. D'autre part, la volonté d'étiqueter ce qu'on entend ou voit est délibérément tournée vers la fonction "décodage acoustico-phonétique", et comporte une part de subjectivité dans certains cas. L'étiqueteur indique alors, en cas de doute, quelle alternative est proposable.

L'étiquetage fin permettra (éventuellement) de lever l'ambiguïté indiquée, si nécessaire.

Une autre raison de ces différences est la volonté de faire vite, et d'étiqueter de manière utilisable aussi rapidement que possible une partie du corpus distribué aux laboratoires. Ce qui ne veut pas dire faire mal, ni créer un étiquetage contradictoire ou incompatible avec les autres en voie de réalisation.

Le sous-ensemble de phonèmes à étiqueter a été défini comme suit : les voyelles, les fricatives sourdes, les plosives sourdes. Il a été décidé d'en indiquer un endroit précis de réalisation certaine, et non pas de le segmenter (pour des raisons évidentes). Dans le cas d'un doute (cf le dernier phonème sur la figure 2), l'alternative est indiquée. Les phonèmes non repérés par un point exact dans le signal sont indiqués sans adresse précise. Sous forme "informatique", la figure 2, qui indique le travail réalisé à la main et à l'oreille par l'étiqueteur, devient la figure 3, fichier de type caractères ASCII qui reflète la même information.

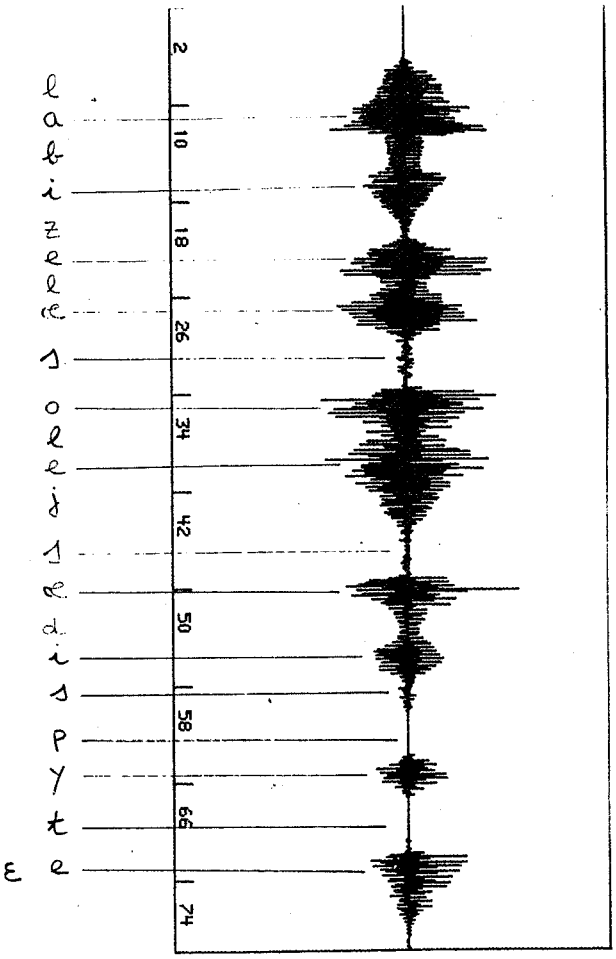


Fig. 2
Exemple d'étiquetage large

3. AVANCEMENT DES TRAVAUX AU 30.6.87

A l'heure où cet article est écrit, les différentes équipes (une dizaine) intéressées par cette action sont en train de comparer leur pratique d'étiquetage (dans les normes ainsi définies) sur la même phrase - cette étude préliminaire a été suggérée par la Commission Etiquetage, de façon à vérifier la cohérence de nos hypothèses. Les résultats seront exploités pour d'éventuelles modifications. Ensuite, la base de données à étiqueter sera répartie selon les désirs et les moyens de chacun. Celle-ci est un sous-ensemble à la fois de BDSONS (bien sûr), mais aussi de la partie de BDSONS distribuée en 1987 sur les 10 cassettes PCM.

Elle comprend les corpus suivants :

LABIS	par 30 locuteurs
PEØ01 à PEØ05	par 11 locuteurs
REC01 et REC02	par 7 locuteurs
RNE01 et RNE02	par 10 locuteurs
RNF01 et RNF02	par 7 locuteurs (dont 5 communs avec les précédents)
ALA01 et ALA02	par 10 locuteurs
ALR01 et ALR02	par 10 locuteurs

soit environ 3 heures de parole.
Les nouvelles les plus récentes seront données par oral aux 16° JEP.

BIBLIOGRAPHIE

[MEF87] Manuel d'étiquetage fréquenciel fin janvier 87. Documentation CNRS GRECO n° 39 (CERFIA et ICP/INPG).
[ABR85] C. ABRY et Al.. Propositions pour la segmentation et l'étiquetage d'une base de données des sons du français. 14° JEP - Paris 1985.

Rapports des réunions de la Commission Etiquetage Large. Disponibles auprès de L. MICLET.

[CON86] Conventions phonétiques. Documentation CNRS. GRECO n° 39. Commission Etiquetage BDSON. Le 28/03/86 (Disponible auprès de L.J. BOE).

en-tête			
NORME	ADRESSE DE LA REALISATION	TRANSCRIPTION	COMMENTAIRES
l	?	l	
a	adr 1	a	?
b	?	b	
i	adr 2	i	
'	'	'	'

Fig. 3

Fichier ASCII d'étiquetage large

PROSODIE

16e Journée d'Étude sur la Parole - Société Française d'Acoustique - Hammamet, 5-9 oct 1987.

TIMING EXTRINSEQUE ET TIMING INTRINSEQUE: LE TEMPS EST-IL UNE VARIABLE CONTROLEE?

Jean-François P.BONNOT

Laboratoire de Phonétique et Centre de Recherches en Informatique
et Combinatoire, UA CNRS 1099
Université du Maine, BP 535, F-72017 Le Mans Cedex

ABSTRACT

In this paper, we try to show that time is a controlled variable. We first deal with some central issues of the action theory. Then, we propound an alternative view based on the concepts of active temporal patterning, (pre)programming, preparatory adjustments and articulatory target. We borrow some examples from our works (cineradiography and electromyography).

1. DYNAMISME, INVARIANCE, ORDRE SERIEL

1.1. Dans un article datant de 1980, Fowler et coll. |1| soulignent que les théories "classiques" de la coarticulation (i.e. les théories linguistiques), obligent à imaginer des mécanismes complexes dont le rôle est de permettre le passage d'un plan abstrait (phonologique) à un plan concret (chaîne parlée). Pour Fowler et coll., les segments sont dynamiques et co-produits, et ils conservent des propriétés invariantes. La sériation et l'individuation ne sont pas affectées, malgré les contraintes liées à l'exécution. Prenant l'exemple du contrôle musculaire de la production vocalique, ils indiquent qu'un modèle basé sur la théorie de l'action doit tenir compte de 4 propriétés: les voyelles constituent une classe naturelle, distincte de celle des consonnes et de toutes "les autres classes d'action". Ensuite, les voyelles ont des propriétés invariantes qui ne sont nullement affectées lors d'une mise en contexte. Enfin, ces propriétés invariantes ne ressortissent pas à des mouvements invariants. Il s'agit ici du fait que les position de départ peuvent être très différentes suivant les environnements. Ceci converge vers une quatrième propriété: les voyelles sont dynamiques. Les cibles articulatoires ne sont pas des formes canoniques. Fowler et coll. considèrent que c'est là un concept statique: dans cette perspective, les mouvements effectués en direction, ou à partir de la cible ne seraient que des "produits annexes", dont la seule utilité serait de passer d'un état statique à un autre état statique. Ils suggèrent encore que l'ensemble des voyelles forme une classe de mouvements équivalents. En ce sens, font partie de cette classe tous les mouvements déterminant la forme et la longueur du conduit vocal, étant entendu que les paramètres varient selon les voyelles. Une voyelle donnée se définit de la même façon, mais les équivalences ne fonctionnent qu'à l'intérieur d'un sous-ensemble. Par ailleurs, les cibles des contoïdes, à la différence de celles des vocoïdes, ne pourraient être exprimées en termes de forme globales du tractus. Une brève parenthèse est ici nécessaire: il est facile de faire surgir de nombreux contre-exemples (cf. ce même volume: Bonnot |2|); dans tous les cas où les consonnes sont caractérisées par un "travail accessoire" (Troubetzkoy |3|), il faut tenir compte de l'ensemble du conduit. C'est en particulier ce que l'on observe pour les consonnes emphatiques de l'arabe.

1.2. Kelso et coll.|4,5|, s'ils relèvent que les variations cinématiques ont leur utilité pour la description du comportement spatio-temporel des articulateurs, n'en insistent pas moins sur le fait que c'est la dynamique elle-même qui est à l'origine des déplacements:"We observe both invariant and systematically varying features of motion when stress and speaking rate are changed. Perhaps most important, our results, analyzed geometrically and interpreted from a dynamic perspective, do not require the assumption that time itself is a controlled variable. Instead, the form of articulatory trajectories over time is seen as a consequence of a control structure whose dynamic parameters are functionnaly equivalent to those of a mechanical mass-spring system (...)". La durée ne serait donc pas une variable contrôlée: les segments (consonnes et voyelles) pourraient bien sûr être définis comme le résultat d'équations de contraintes, mais surtout, l'organisation générale de la chaîne ressortirait au même processus. Le fait que les structures articulatoires évoluent dans le temps permettrait donc de faire l'économie d'un système de régulation du timing.

2. PROGRAMMES ET "SIGNAUX D'ERREUR"

2.1. Il faut reconnaître que les expériences de Fowler (et coll.) et de Kelso (et coll.) sont remarquablement contrôlées. Il est toutefois possible de faire valoir un certain nombre d'arguments autorisant un recentrage de la théorie, et montrant (a) que celle-ci ne rend compte que d'une facette de la production de la parole - alors qu'elle est présentée comme ayant vocation universelle - et (b) que le temps est bien une variable contrôlée. Les 2 points sont intimement liés car, dès lors que l'on fait intervenir les notions de programme (et de préprogrammation), on doit nécessairement considérer que les événements articulatoires sont organisés en fonction d'une succession de contraintes, entièrement ou partiellement prédictibles, et entraînant d'éventuelles corrections au cours de l'exécution. Quelle que soit la définition que l'on donne du "programme", on peut souscrire à cette remarque très générale de Monsell |6|:"use of terms like "program", "retrieval", etc. tends to provoke aversive reactions among some action theorists, especially those advocating a "dynamic" perspective (...) Some mechanism must nevertheless specify to such a system, the sequence of parameterisations required."

2.2. Kelso |7| concède que des informations afférentes, concernant la vitesse et la position peuvent être fournies par les fuseaux neuromusculaires et par les structures articulaires; néanmoins, les unités de temps restent définies uniquement en fonction des variables relatives à l'état du système (les relations spatiales entre articulateurs). Il est pourtant raisonnable de supposer que la mise en évidence de phénomènes de feedback implique l'existence d'une modélisation temporelle active: Granit |8| relève que la contribution des fuseaux neuromusculaires, qu'ils agissent seuls ou en coopération avec les organes peauciers et articulaires, ne se résume pas à produire une simple "servo-assistance", mais consiste plutôt à fournir au système de contrôle des informations dont l'importance est capitale pour l'exécution des mouvements complexes. Granit est très clair:"An error detector can act sensibly only if it is related

to the purpose of the movement in which it operates (...) Clearly, the concept of error is tied to the concept of a goal (souligné par Granit). More often than not, the nervous system incorporating the goal will be cortical (souligné par nous). D'autre part, Gentil |9| souligne (p.25) que l'exécution des mouvements de la parole est sous le contrôle de circuits sensori-moteurs: les afférences vers le cortex moteur peuvent emprunter une voie directe, ou passer par les aires somesthésiques, ou encore transiter par le cortex prémoteur. Gentil ajoute que "tous ces circuits ont de courtes latences et peuvent contribuer aux ajustements permanents." Abbs |10|, qui est cependant l'un des promoteurs de la notion d'équivalence motrice, remarque que la parole est "préméditée", qu'elle est le résultat d'un sur-apprentissage, et qu'elle implique l'existence de cibles bien définies (Abbs veut parler ici de "stimuli auditivement acceptables"). Dans une autre contribution, Abbs |11| note que les données recueillies chez les primates non humains et chez l'Homme, permettent de penser que la région pariétale postérieure (7b) est chargée de la sélection de cibles abstraites (par exemple les traits phonologiques chez l'Homme), tandis que les régions frontales (aires 4, 6-44) sont spécialisées dans l'exécution de ces cibles (p.218).

2.3. Se fondant sur de nombreuses données expérimentales, Hoffer |12| propose un scénario séduisant, qui a l'avantage de réunir les propriétés biomécaniques du muscle et un contrôle moteur actif, au sein d'un même processus "unifié". Tout d'abord, Hoffer remarque qu'un mouvement aussi complexe que celui consistant à frapper une balle de golf, doit être en partie préprogrammé, et donc réalisé en boucle ouverte, étant donné la vitesse d'exécution élevée. Il estime toutefois que l'information afférente est largement utilisée:"in golf, the CNS stipulates the sequential timing and levels of activation of different muscles (...) The path, orientation and velocity of the club-head are then completely determined by the visco-elastic properties of the body, which are reflexively regulated, and by the mechanical impedance of the club in motion. Knowledge of the results, combined with proprioception, is then used to make adjustments either in the parameters of the physical plan,(...) or in the internal template of the motor program" (nous soulignons). Il est particulièrement important d'insister sur le rôle de ces "gabarits" (templates), qui nécessitent un apprentissage et sont indispensables à l'activation séquentielle des groupes musculaires. Nous retrouvons, exprimées d'une autre façon, des notions auxquelles nous avons fait appel ailleurs (Bonnot et coll.|13|, Bonnot |14|): les gabarits correspondent à la macrostructure temporelle (sequencing; sériation des éléments de la chaîne) et les composants "connaissance du résultat + proprioception" peuvent être mis en parallèle avec la microstructure (phasing; "accord fin", pour reprendre une expression de Hoffer). Il y a toutefois une différence: il semble bien que pour Hoffer, les 2 composantes soient dissociées dans le temps. Les ajustements interviendraient pour "affiner" le geste suivant; nous pensons au contraire qu'il y a simultanéité: dans des conditions de production "normales", les principales procédures d'évaluation et de correction sont disponibles de longue date (phénomène de sur-apprentissage; voir aussi ci-dessous le point de vue de Requin). A plusieurs reprises, nous avons pu mettre en évidence des phénomènes de ce type. Ainsi, l'étude de logatomes "équilibrés", constitués de structures syllabiques simples et itératives, permet de conclure à une organisation hiérarchisée de la programmation motrice. L'hypothèse d'un mécanisme de préprogrammation est fermement étayée par nos résultats: par exemple, la durée des temps de latence (TL) de reprise d'activité de l'élévateur (LP) passe de 127 msec (|mVmVmV|) à 117 msec (|mVmVbV|), puis à 105 msec (|mVbVmV|) et enfin à 95 msec (|mVbVbV|) (|mmm-mmb|, NS; |mmm-mbm|, $p<0.05$; |mmm-mbb|, $p<0.01$). La durée variable du TL paraît donc correspondre - au plan de l'exécution - au temps de "lecture du programme": celui-ci est plus complexe pour les occurrences possédant plus de consonnes nasales. D'autre part, de nombreux éléments vont dans le sens d'un rééquilibrage du timing en cours de production. Nous emprunterons une illustration à l'une de nos expériences portant sur l'opposition VCV vs. VC≠≠CV en français. Chez tous les témoins, le TL de LP (|-mV vs -m≠≠mV|) est plus long pour les réalisations de type C≠≠C:

	|m|	|mm|		
JFB1	81	100	NS	
JFB2	91	153	$p<0.05$	TL de LP
VDW	67	103	$p<0.05$	en msec.
R	77	114	NS	

Le sens de la variation est absolument systématique, même lorsque le rapprochement n'est pas significatif statistiquement. Ceci peut être interprété comme une preuve limitée de l'existence d'un mécanisme d'exploration temporelle se développant au minimum dans le cadre syllabique. Nous n'affirmons toutefois nullement qu'il y ait là matière à valider ie très puissant modèle de Kozhevnikov et Chistovich |15|. En fait, il semble surtout s'agir d'une révision de l'organisation temporelle dans le cadre CC. D'autre part, chez JFB (1 et 2), le TL de l'orbiculaire supérieur (OOS) présente des durées extrêmement proches les unes des autres, qu'il s'agisse d'une consonne simple ou "double", et cela pour |p pp b bb m mm|: ces comportements divergents s'expliquent bien, car les lèvres constituent l'articulateur primaire, et l'occlusion bilabiale doit être produite sans délai; par contre, le mouvement ascendant du voile jouit d'une plus grande liberté.

2.4. Certains événements sont donc prévus à l'avance. Dans cette perspective, l'état initial du système, ou un état intermédiaire, si la séquence est longue, joue le rôle de référence. De nombreux physiologistes insistent d'ailleurs sur les ajustements préparatoires. Requin |16| souligne par exemple que les réorganisations du substrat postural qui anticipent les conséquences de l'exécution d'un mouvement, doivent être considérées comme des processus préparatoires à part entière. Requin propose d'intégrer à la préparation programmée les phénomènes d'ajustements, qui interviennent par le biais de systèmes exploratoires. Il observe que ceci justifie le point de vue de Semjen |17|, et permet donc d'étendre la notion aux processus ressortissant au contrôle sensoriel, rendant ainsi caduque la classique opposition entre mouvements programmés - qui seraient effectués en boucle ouverte - et mouvements contrôlés - qui seraient réalisés en boucle fermée (p.385-386). Il est d'ailleurs permis de penser qu'une tâche motrice peut débuter avant que la totalité de la séquence ne soit programmée. C'est ce que suggèrent Stelmach et coll.|18| pour l'écriture manuelle: le programme moteur serait mis en place par tranches. Ce point de vue rejoint celui de Lindsley |19|, qui a travaillé sur l'encodage d'unités de 1ère articulation: un 1er dispositif serait chargé d'assurer la fluence initiale; l'exécution serait retardée jusqu'à ce que l'encodage du second élément soit terminé. Il y aurait ensuite fonctionnement parallèle de deux niveaux de programme. On doit naturellement reconnaître qu'il n'est pas aisé de décider si l'on a affaire à une programmation globale - précédant totalement la mise en oeuvre - suivie d'une recherche des unités, ou si c'est la programmation proprement dite qui se poursuit, alors que l'exécution est engagée. La seconde solution paraît mieux à même d'intégrer harmonieusement les données périphériques, et surtout les "simulations" en provenance du feedback interne.

3. LES CIBLES ARTICULATOIRES

3.1. La vérification expérimentale de ces hypothèses par des travaux portant sur la parole, mais aussi sur le mouvement en général, réhabilite les descriptions faites en termes de cibles articulatoires: en effet, si, (par exemple) le TL de LP est d'autant plus long que la suite de la séquence comporte zéro, un ou deux

|m|, c'est très vraisemblablement parce qu'il faut que le système de programmation évalue les contraintes liées aux passages successifs d'une position contractée du voile à une position (légèrement) abaissée et inversement. C'est un fait bien attesté que certaines structures et donc que certains gestes bénéficient d'une plus grande latitude que d'autres. On peut même penser, avec Perkell |20|, qu'il existe des "trajectoires relativement invariantes". Si l'hypothèse d'une organisation temporelle intrinsèque en sort renforcée, il n'en va pas de même de celle attribuant aux équations de contraintes - et par voie de conséquence à la co-production - un rôle central dans la production de la parole: nous avons vu que de nombreux travaux (cf.notamment Bonnot |14|, Sussman et Westbury |21|) ont bien mis l'accent sur l'importance de la sensibilité au contexte. Simon |22| avait parfaitement raison d'écrire, dès 1967, que "la parole se réalise essentiellement par un ensemble de mouvements et (...) ces mouvements sont prédominants. |Mais| les tenues |articulatoires| si variables et si dégradées qu'elles puissent être parfois, sont une nécessité pour la compréhension"(p.208-209). Cette position a reçu tout récemment le soutien de Fujimura |23| et de Kent |24|. Pour Fujimura, le "iceberg pattern" est constitué par un mouvement élémentaire, et au moins chez un sujet donné, il est relativement invariant. On observe ces modèles lorsque l'articulateur est utilisé pour un trait inhérent spécifiant le lieu d'articulation. Il existe des relations temporelles lâches entre les "icebergs" et il est donc possible de mettre en lumière des stratégies de dissimilation ou de "répulsion", aussi bien que des phénomènes de lissage entre mouvements consécutifs. A nouveau, nos données cinéradiographiques et électromyographiques |13,14,25| confortent ce point de vue: s'agissant du 1er aspect, nous avons montré qu'en arabe saoudien, les géminées (|t-tt, t-tt|) présentent une "accentuation" des propriétés articulatoires des simples correspondantes. La non emphatique est caractérisée par une fermeture plus marquée ainsi que par un relèvement et une antériorisation de de la masse linguale. Quant à l'emphatique, elle offre une constriction postérieure accrue et une "extension de l'emphase" en direction de la zone uvulaire. On observera que ces faits concernent tout particulièrement les parties d'organes impliqués dans la réalisation du lieu d'articulation et, pour les emphatiques, du "travail accessoire". Dans les deux cas, on note une plus grande stabilité de l'articulation. Ceci peut être considéré comme une stratégie d'individuation segmentale.

3.2. Bien entendu, il est possible d'interpréter ces faits en termes de renforcement articulatoire. Cette option suppose que l'on considère la <u>force</u> comme étant la variable contrôlée: la durée plus grande de |tt| et de |t̪t̪|, leur plus grande stabilité, etc., ne seraient que de simples conséquences de l'augmentation globale de l'effort physiologique; cette analyse rejoint, par d'autres voies, celle de Kelso et coll., puisqu'elle met presque uniquement l'accent sur les caractéristiques bio-mécaniques du système. Or, Semjen et coll.|26| ont réalisé une expérience évaluant la coordination des contrôles du temps et de la force au sein d'une séquence motrice composée de 4 "coups" portés par le doigt sur une touche. L'un de ces coups devait être produit plus fortement (stress+) ou plus faiblement (stress-) que tous les autres. On observe que les intervalles précédant et suivant le coup marqué "négativement" subissent un allongement identique à ceux précédant et suivant le coup marqué "positivement". Les facteurs périphériques (bio-mécaniques) ne peuvent donc être tenus pour responsables de la production d'un degré de force plus élevé. On doit plutôt conclure que ces allongements sont à mettre en rapport avec des processus centraux, contrôlant les modifications de force: "this suggest that, in addition to the plan established during the RT, real-time control processes are also involved in the execution of the sequence. These processes could comprise a memory search for the target force level and/or updating the executive program with specification of the level of force"(p.181). En ce qui concerne l'opposition géminée vs. simple, on peut admettre qu'il s'agit, traduite en termes spatiaux, d'une manifestation du contrôle temporel fin (phasing). Kent |24| remarque que "the temporal distortions that arise in neuropathologies such as apraxia of speech or ataxic dysarthria indicate that iceberg pattern are not physiologically fixed but rather are optimized movements for the normal conditions of speech production" (p.341). Dans le cas présent, les stratégies "d'optimisation" opèrent en fonction du contexte immédiat (C vs C≠C). Même si les caractéristiques générales sont prédéterminées, cela n'empêche pas que les contrôles moteurs évaluent les conséquences du geste <u>en fonction de la durée disponible</u>: en arabe, les géminées sont toujours beaucoup plus longues que les simples. Nous ne suggérons pas que la force ne soit pas un élément important. Au plan strictement physiologique (mouvements des membres), Stein |27| est très réservé quant à la possibilité qu'une variable particulière (force, viscosité, rigidité, vélocité, longueur etc...) soit contrôlée à l'exclusion de toutes les autres: il y a vraisemblablement interaction complexe entre les indices. C'est ce que les grammairiens arabes avaient intuitivement perçu : Cantineau |28| souligne que "le tašdid ou "renforcement" des consonnes, qui est le terme arabe correspondant à "gémination" n'est pas compris dans la liste des ṣifât al-ḥurûf |qualités des consonnes|, probablement parce qu'il ne modifie pas leur nature propre, mais prolonge seulement leur tenue"(p.25). Durée et force semblent donc survenir de façon concomitante, <u>même si le contrôle de la 1ère composante jouit d'une autonomie très importante</u>. Ceci rejoint les conclusions de Semjen et coll., qui estiment que le timing est partiellement indépendant de la force, bien qu'il soit modulé par des mécanismes centraux la contrôlant. Traduit en termes phonologiques, et rapporté à la situation de l'arabe, on peut dire que la longueur est le trait contrastif, alors que la force est l'un des traits configuratifs.

3.3. Fujimura souligne encore que ces stéréotypes (les "icebergs") sont essentiellement décelables dans les sections d'approche ou de sortie du noyau syllabique ("into or out of consonantal occlusion, out of or into a specific (perhaps only stressed) syllable nucleus"(p.230)). Ceci vaut pour le mouvement proprement dit. Au plan immédiatement supérieur, qui est celui de l'activité EMG, certains faits viennent conforter cette analyse: d'assez nombreux sujets présentent une activité d'OOS en 3 phases (2 bouffées séparées par un intervalle de quasi silence électrique) lors de la production d'occlusives bilabiales. Chez le sujet JFB, dans des trisyllabes CVCVCV où C=|b,m|, la 1ère bouffée et l'intervalle sont toujours investis d'un pouvoir différenciatif beaucoup plus important que la seconde bouffée (pour cette dernière, aucune opposition n'atteint le seuil $p<0.05$ en C3). Il est possible d'en conclure que le muscle (OOS) est impliqué de deux façons différentes dans l'encodage moteur. Les contrastes organisationnels (|b vs.m|) qui caractérisent la 1ère bouffée indiquent que l'on a affaire à un composant dynamique du système d'encodage, particulièrement sensible au contexte et donc, très malléable. Le timing plus constant de la seconde bouffée montre qu'elle occupe une place hiérarchiquement "inférieure" mais, en même temps, elle pourrait servir de repère permettant la répartition temporelle des autres événements plus fluides. Elle correspond à une section transitionnelle (sortie de la phase d'occlusion et entrée dans le noyau vocalique). Les chevauchements articulatoires que l'on observe fréquemment pour les groupes de consonnes totalement ou partiellement hétérorganiques (cf.Rochette |29| et Marchal |30|) sont dès lors explicables sans qu'il soit besoin de recourir à la notion de co-production. En effet, ce terme est utilisé par la plupart des théoriciens de l'action dans un cadre de mécanique physiologique pure (absence (presque) totale de contrôle). Il convient d'ailleurs de reconnaître que Marchal |30| n'adopte pas une perspective aussi radicale, puisqu'il parle de "boucles tactiles"). Lors de la production d'un groupe, le programme est

confronté à une nécessité absolue: au sein d'un domaine temporel dont les bornes sont placées de façon contraignante, il faut situer deux ou plusieurs composants relativement invariants, les "icebergs". On n'est pas tenu de voir là des phénomènes ne mettant en jeu que des contraintes purement articulatoires. La prise en compte des segments déjà réalisés et la prévision des unités à venir explique ces télescopages. Comme l'écrit encore Kent |24|, "the component movements for a given phonetic target must be correctly timed with respect to the phonetic context. Acceptable timing relations are few and seem to be specified with respect to movement onsets, offsets, or velocity maxima"(p.240). Certains aspects des travaux de Gentil et Gay |31| viennent également à l'appui de notre thèse: les modèles d'activité sont très différents, suivant que l'on a affaire à la production de parole ou à la mastication. Dans le 1er cas, les mouvements du maxillaire sont à la fois plus simples et mieux synchronisés. Ceci nous paraît être compatible avec l'idée qu'il existe des cibles proprement phonologiques et relativement abstraites, et que ces cibles sont contrôlées neurologiquement. Cela montre par la même occasion qu'on ne saurait faire l'économie de ce niveau. Tous ces éléments permettent de mieux mettre en évidence les interrelations entre cible et contrôle programmé de la durée. Les cibles, comme l'ont montré Simon |22| ou Perkell et Nelson |32|, peuvent ne pas être atteintes, ou être dégradées: il devient alors indispensable de corriger la trajectoire pour les unités à venir, en tenant compte de la vitesse de production, mais aussi - et cela est très important - de la nature du système phonologique et des principes phonotactiques. Les travaux très récents de Sereno et coll.|33| font apparaître que l'acquisition par les enfants des stratégies coarticulatoires est progressive. Il y a là une indication supplémentaire que ces conduites ne ressortissent pas à une systématique purement physiologique, mais qu'elles sont la conséquence d'un contrôle temporel "volontaire".

REFERENCES

|1| Fowler C., Rubin P., Remez R. et Turvey M."Implications for Speech Production of a General Theory of Action" in Butterworth B.(ed) Language Production,1, Academic Press, New York, 1980, 373-420.

|2| Bonnot J-F.P. "Conventions linguistiques et naturel acoustico-physiologique: peut-on parler de règles de coarticulation?" 16e JEP, Hammamet, 1987.

|3| Troubetzkoy N.S. Principes de phonologie, réimp. 1970, Klincksieck, Paris.

|4| Kelso J., Vatikiotis-Bateson E., Saltzman E. et Kay B."A Qualitative Dynamic Analysis of Reiterant Speech Production: Phase Portraits, Kinematics, and Dynamic Modelling" JASA, 1985, 77, 266-280.

|5| Kelso J. et Tuller B."Intrinsic Time in Speech Production: Theory, Methodology and Preliminary Observations" Haskins Labs: Status Report on Speech Research, SR-81, 1985, 23-39.

|6| Monsell S. "Programming of Complex Sequences: Evidence from the Timing of Rapid Speech and Other Productions" Experimental Brain Research Series, 15, Springer, Berlin, 1986, 72-86.

|7| Kelso J. "Pattern Formation in Speech and Limb Movements Involving Many Degrees of Freedom" Experimental Brain Research Series, 15, idem, 105-128.

|8| Granit R."Multiple Roles of Muscular Afferents" The Behavioral and Brain Sciences, 1982, 5, 547.

|9| Gentil M. Organisation temporelle du système articulaire: contributions musculaires aux gestes labiaux, linguaux et mandibulaires, Thèse Etat, 1986, Strasbourg.

|10| Abbs J.H."Invariance and Variability in Speech Production: A Distinction Between Linguistic Intent and Its Neuromotor Implementation" in Perkell J.S. et Klatt D.H.Invariance and Variability in Speech Processes, L.Erlbaum, Hillsdale, NJ., 1986, 202-225.

|11| Abbs J.H. "A Speech-Motor-System Perspective on Nervous-System-Control Variables" The Behavioral and Brain Sciences, 1982, 5, 541-542.

|12| Hoffer J.A. "Central Control and Reflex Regulation of Mechanical Impedance: The basis for a Unified Motor Control Scheme" The Behavioral and Brain Sciences, 1982, 5, 548-549.

|13| Bonnot J-F.P., Chevrie-Muller C., Arabia-Guidet C., Maton B. et Greiner G.F."Coarticulation and Motor Encoding of Labiality and Nasality in CVCVCV Nonsense Words" Speech Communication, 1986, 5, 83-95.

|14| Bonnot J-F.P.Contribution à l'étude phonétique et phonologique de l'organisation temporelle de l'activité électromyographique labiale et vélaire. Coarticulation et processus d'encodage moteur, Thèse Etat, Strasbourg, 1986, 704 pages.

|15| Kozhevnikov V.A. et Chistovich L.A.Speech, Articulation and Perception, Joint Publ.Service US Dept. of Commerce, Washington DC, 1965.

|16| Requin J."Toward a Psychobiology of Preparation for Action" in Stelmach G. et Requin J.(eds)Tutorials in Motor Behavior, North-Holland Co, 1980, 373-398.

|17| Semjen A."From Motor Learning to Sensorimotor Skill Acquisition" J.of Human Movement Studies, 1978, 3, 182-191.

|18| Stelmach G., Mullins P. et Teulings H."Motor Programming and Temporal Patterns in Handwriting" in Gibbon J. et Allan L.(eds)Timing and Time Perception, Annals of the New York Academy of Sciences, 1984, 423, 168-182.

|19| Lindsley J.R. "Producing Simple Utterances: Details of the Planning Process" J.of Psycholinguistic Research, 1976, 5, 331-353.

|20| Perkell J.S."Coarticulation Strategies: Preliminary Implications of a Detailed Analysis of Lower Lip Protrusion Movements" Speech Communication, 1986, 5, 47-68.

|21| Sussman H. et Westbury J. "The Effects of Antagonistic Gesture on Temporal and Amplitude Parameters of Anticipatory Labial Coarticulation" JSHR, 1981, 46, 16-24.

|22| Simon P.Les consonnes françaises. Mouvements et positions à la lumière de la radiocinématographie, Thèse Etat, Klincksieck, Paris, 1967.

|23| Fujimura O. "Relative Invariance of Articulatory Movements: An Iceberg Model" Invariance and Variability etc., 1986, 226-234.

|24| Kent R.D."The Iceberg Hypothesis: The Temporal Assembly of Speech Movements" idem, 234-242.

|25| Bonnot J-F.P. "Etude expérimentale de certains aspects de la gémination et de l'emphase en arabe" Trav.Inst.Phonétique Strasbourg, 1979, 11, 109-118.

|26| Semjen A., Garcia-Colera A. et Requin J."On Controlling Force and Time in Rhythmic Movement Sequences: The Effect of Stress Location" in Timing and Time Perception, etc., 168-182.

|27| Stein R. "What Muscle Variable(s) Does the Nervous System Control in Limb Movements?" The Behavioral and Brain Sciences, 1982, 5, 535-577.

|28| Cantineau J."Cours de phonétique arabe" Etudes de linguistique arabe, Klincksieck, Paris, 1960.

|29| Rochette C.Les groupes de consonnes en français Klincksieck, Paris, 1973.

|30| Marchal A. L'électropalatographie: contribution à l'étude de la coarticulation dans les groupes d'occlusives, Thèse Etat, Nancy, 1985.

|31| Gentil M. et Gay T."Neuromuscular Specialization of the Mandibular Motor System: Speech vs.Non Speech Movements" Speech Communication, 1986, 5.

|32| Perkell J.S. et Nelson W.L."Articulatory Targets and Speech Motor Control: A Study of Vowel Production" in Grillner S. et coll.(eds)Speech Motor Control, Pergamon Press, Oxford, 1982, 187-204.

|33| Sereno J., Baum S., Marean G. et Lieberman Ph. "Acoustic Analyses and Perceptual Data on Anticipatory Labial Coarticulation in Adults and Children", JASA, 1987, 81, 512-519.

HIERARCHISATION DES PARAMETRES ACOUSTIQUES ET IDENTIFICATION DES FRONTIERES

DUEZ Danielle

U A CNRS 261 UNIVERSITE DE PROVENCE 13621 AIX EN PROVENCE CEDEX

ABSTRACT
The perception of subjective pauses appears to be strongly correlated to vowel boundary lengthening. These résults suggest
1) Pause is an integrant part of final lengthening
2) Hierarchisation of boundary should be based on duration parameter.

1) INTRODUCTION

L'étroite corrélation qui existe entre la perception de la pause et les variations des paramètres acoustiques de la voyelle finale de constituant a été soulignée dans une étude précédente (Duez, 1985). La manière dont ces paramètres se hiérarchisent est ici examinée à travers le phénomène de la pause subjective (pause qui ne correspond pas une interruption dans le signal).

Il s'agit 1) de vérifier l'hypothèse selon laquelle la pause ferait partie de l'allongement final, 2) de mettre aussi en évidence le rôle effectif des paramètres dans l'identification de la frontière.

2) PROCEDURE EXPERIMENTALE

a) Corpus

Deux phrases sont lues de façon neutre, sans marquer ni pause silencieuse, ni accent externe par un locuteur ne présentant pas d'accent régional.

"Il faut que j'entérine cet acte"

"Il faut que Jean terrasse ce monstre"

A partir de ces deux phrases, deux phrases sont obtenues par concaténation.

"Il faut que j'enterrasse ce monstre"

"Il faut que Jean térine cet acte"

Les phrases concaténées présentent une structure prosodique acceptable, mais sont dépourvues de contenu sémantique.

b) Modifications.

La durée, la fréquence fondamentale, et l'intensité des voyelles cibles /ãɛ̃/ sont obtenues à partir des tracés oscillographiques, de la courbe de fréquence et de la courbe d'intensité (méthode de détection mise au point par Teston, 1972). Les valeurs relevées pour la phrase (a) servent de référence pour les manipulations effectuées sur la phrase (b), et réciproquement.

Phrase (a), "Il faut que Jean térine cet acte (durée: 300ms, écart fo: 2 tons, écart intensité (10dB).

Phrase (b), "Il faut que j'enterrasse ce monstre" (durée: 120ms, écart fo: 2 tons, écart intensité (0dB).

Les manipulations sont faites avec vocoder prédictif à l'ICP (programme M Proso, 1978). Pour les modifications de durée, nous avons procédé par pas de 25% (seuil légèrement supérieur au seuil de durée déterminé par Rossi (1971, page 45). La durée du /ãɛ̃/ de la phrase (a) est ramenée successivement de 300ms à 250ms, 200ms, 150ms, 120ms, la durée du /ãɛ̃/ bref (120ms) est augmentée de 30ms, 80ms, 130ms, et 180ms. Pour les modifications de fo, nous avons procédé par pas de 1/3 de tons (seuil défini par Rossi et Chafcouloff (197) et Di Cristo (1985). Dans la phrase (a), l'écart est successivement ramené à 1/3, 1/6, 1, 1,3, 1,6, 2 tons, la procédure inverse est adoptée pour la phrase (b). Les modifications appliquées pour l'intensité sont de 3dB, ce sont les valeurs obtenues par Rossi (1971). Dans la phrase (a), l'écart est réduit de 3dB, 6dB et 10dB, et réciproquement pour la phrase (b).

Quatorze phrases sont ainsi générées à partir de la phrase (a), quatorze également à partir de la phrase (b).

c) Tests

Le test s'organise en trois sessions auxquelles participent 10 sujets. Au cours d'une session, chacune des séries de phrases est présentée au sujet en quatre ordres aléatoires. Une interruption d'une heure est maintenue entre chaque session. La consigne est d'appuyer sur un signal chaque fois qu'une pause est perçue après la syllabe /ʒãɛ̃/.

d) Analyse

Pour chacun des échantillons (336 en tout) sont calculés le nombre de réponses (nombre de sujets x nombre d'écoutes), et le temps de réaction (intervalle existant entre la première période de la voyelle /ãɛ̃/ et le début du clic).

3) RESULTATS

a) rôle de la durée

L'examen des figures 1 et 2 révèle une influence similaire de la durée pour les deux séries de phrases et l'étroite corrélation qui existe entre la durée de la voyelle et le pourcentage de réponses (les coefficients de corrélation sont respectivement de 0.98 et 0.99, p=0.001).

Dans les phrases (a) où les variations de fo et d'intensité sont élevées, la réduction progressive de la durée s'accompagne d'une réduction du taux d'identification. Pour une durée de 320ms, on obtient un pourcentage élevé (88%), pour une durée comprise entre 200ms et 250ms, le taux d'identification est moyen (62% et 63%), pour des durées de 150ms et 120ms, le taux est faible (28% et 39%).

Dans les phrases (b), on observe la même tendance avec une amplitude plus marquée pour la courbe de réponses: pour la durée non modifiée, on relève 12% de pauses subjectives, pour la durée maximale (300ms), 94% de pauses, les pourcentages relevés pour les durées de 150ms, 200ms et 250ms sont respectivement de 43%, 79% et 82%.

Certaines différences peuvent cependant être relevées. Lorsque la durée est de 120ms, le pourcentage de pauses subjectives est plus élevé dans les phrases (a): on peut y voir l'effet conjugué de l'intensité et de fo. Le seuil d'identification est en revanche moins fin dans cette même série pour les autres durées.

Les conditions de l'expérience, et en particulier le nombre limité d'échantillons peut être à l'origine de cette différence: le sujet peut avoir été amené à se créer une référence plus fine dans les phrases où les variations paramétriques sont faibles. Des raisons linguistiques peuvent aussi avoir joué. Dans les phrases (b), les variations de fo et d'intensité sont élevées, l'attente d'un allongement plus marqué peut en avoir résulté. Les temps de réaction relevés pour les deux séries de phrases confirment en partie cette hypothèse: pour les phrases (a) et (b), une durée de 250ms conduit à un temps de réaction similaire (59ms et 60ms, t=1.11, p=0.9), une durée de 300ms à un temps de réaction significativemet plus long dans les phrases (a): 58ms et 67ms (t=11,2, p=0.001). Pour une durée de 200ms, le temps de réaction est cependant paradoxalement plus long dans cette série de phrases (60ms, 65ms, t=5.5, p=0.001).

Malgré ces divergences, on peut noter dans les phrases (a) et (b) qu'un allongement de l'ordre de 50% conduit à l'identification d'une pause-frontière, ce qui rejoint les résultats obtenus par Rossi (1981) et Di Cristo (1985) pour la réalisation acoustique de la frontière de continuation majeure et de groupe intonatif.

b) Rôle de la fréquence fondamentale

L'influence de fo sur l'identification des frontières est peu marquée. Dans les phrases (b), un écart de 2 tons conduit à un pourcentage d'identification nettement inférieur à la moyenne (31%); dans les phrases (a), le pourcentage d'identification reste faible, quelle que soit la réduction de l'écart (se reporter figures 3 et 4).

c) Rôle de l'intensité

L'effet de l'intensité sur l'identification de la pause rejoint celui de fo. Pour les phrases (a), le pourcentage d'identification n'est que peu modifié par la réduction de l'écart: il reste stable et élevé (77%, 84%, 82% et 88%). Pour les phrases (b), la même absence de corrélation peut être relevée: le pourcentage maximum est de 22%, et l'accroissement progressif de l'écart d'intensité conduit à un accroissement progressif qui n'est que de 4% environ (se reporter figures 5 et 6).

4) CONCLUSIONS

Ces résultats nous conduisent aux remarques suivantes:

-La pause fait partie de l'allongement final, ce qui implique que la pause-silence et l'allongement syllabique relèvent d'un même phénomène sous-jacent.

-Les résultats obtenus pour le français rejoignent ceux relevés pour l'anglais et le néerlandais : l'allongement final paraît avoir un rôle universel et fondamental dans l'identification des frontières.

-Une hiérarchisation des frontières fondée sur la durée dérive de ces résultats: elle diffère de la hiérarchisation proposée par Rossi et Di Cristo (1980). Cette hiérarchisation ne peut cependant être enlevée que dans le cadre du contour.

Cette étude, qui ne constitue qu'une première approche expérimentale du problème permet de tester la validité de la méthode, et donne des résultats qui nous paraissent intéressants. Elle doit être reprise et approfondie. Il s'agit d'obtenir un seuil d'identification plus précis, et d'examiner les interactions existant entre les différents paramètres.

BIBLIOGRAPHIE

BOUWHUIS, M. de ROOIJ, J. (1977), Vowel length and the perception of prosodic boundaries, IPO, Eindhoven, 12, 63-68.

Di CRISTO, A. (1985). De la microprosodie à l'intonosyntaxe, Thèse de Doctorat d'Etat, Université de Provence.

DUEZ, D. (1985), Perception of silent pauses in continuous speech, Language and Speech, 28, part 4, 377-389.

ROSSI, M. (1971), L'intensité spécifique des voyelles, Phonetica, 24, 129-161.

ROSSI, M. (1971), Le seuil différentiel de durée, in Papers in Memory of Delattre, Mouton, La Hague, 435-449

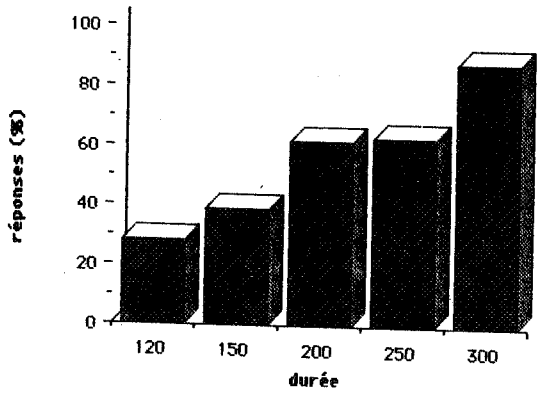
ROSSI, M., CHAFCOULOFF, M. (1972), Recherche sur le seuil différentiel de fo dans la parole, Travaux de l'Institut de Phonétique d'Aix en Provence, 1, 179-185.

ROSSI, M., Di Cristo, A., HIRST, D., MARTIN, P., NISHINUMA, Y. (1981), L'intonation de l'acoustique à la sémantique, PARIS, Klincksiek.

Teston B. (1972), Description d'un système de détection de l'intensité et de la mélodie de la voix, Travaux de l'Institut de Phonétique d'Aix en Provence, Vol 1, 129-145.

phrases (a)

Figure 1



phrases (b)

Figure 2

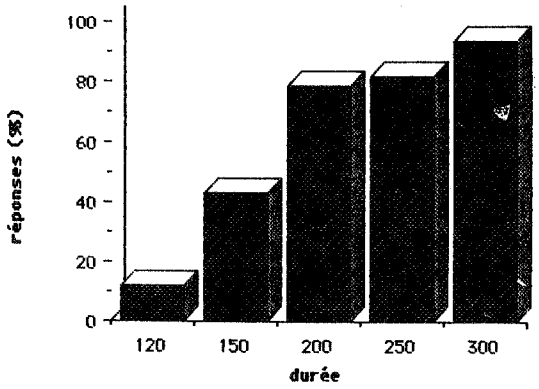


Figure 3

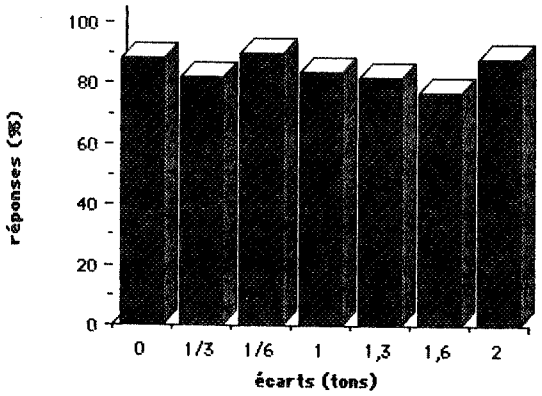


Figure 4

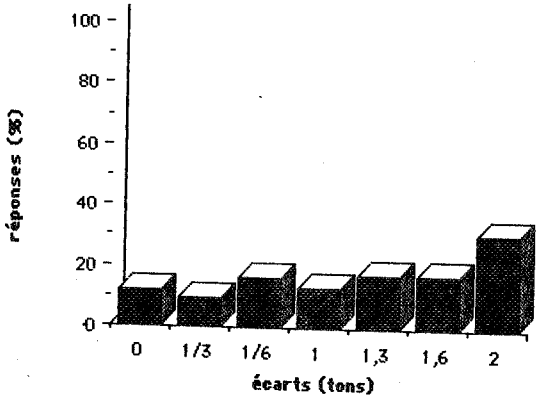


Figure 5

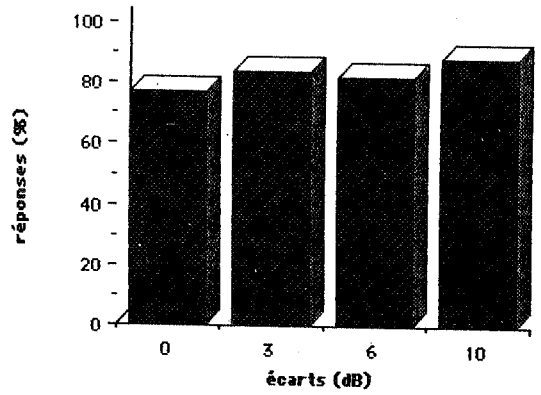
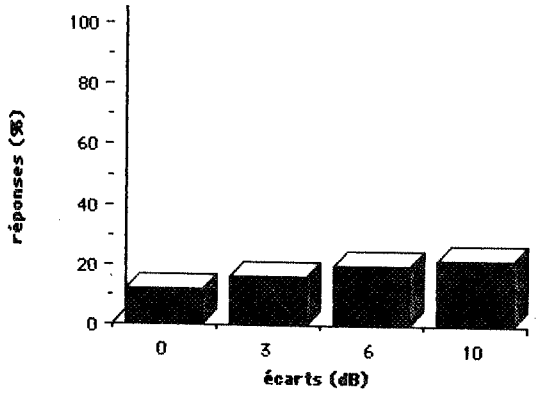


Figure 6



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

F. EMERARD, C. BENOIT.

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

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

I INTRODUCTION

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

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

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

II CONTENU DE LA BASE

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

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

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

1) Informations acquises manuellement :

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

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

2) Informations acquises automatiquement :

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

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

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

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

3) Exemple d'interrogation

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

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

La figure 2 présente la sortie imprimée de tels résultats.

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

```
   89    56. N (Y )EI a ML 2-2 2/3 nv(dv)vo 14    370.[NYEI]
  871    73. L (Y )OE f ML 2-2 2/4 lv(dv)vo 22    443.[LYOER]
 1585    78. D (Y )O  f ML 2-2 2/5 pv(dv)vo 46    444.[DYOT*]
 2393   100. N (Y )EI a ML 3-3 2/3 nv(dv)vo 07    359.[NYEI]
 2500    96. L (Y )EI a ML 2-2 2/3 lv(dv)vo 16    379.[LYEI]
 2816    89. S (Y )ON a ML 4-4 2/3 cs(dv)vn 40    367.[SYON]
 3764    42. S (Y )EU a ML 2-2 2/3 cs(dv)vo 14    402.[SYEU]
 3962    44. S (Y )EU a ML 2-2 2/3 cs(dv)vo 14    348.[SYEU]
 5675    90. V (Y )ON a ML 3-3 2/3 cv(dv)vn 30    354.[VYON]
 5740    61. T (Y )AI f ML 2-2 2/4 ps(dv)vo 16    477.[TYAIR]
```

Figure 2

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

III BASE DE CONNAISSANCES

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

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

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

Figure 1

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

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

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

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

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

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

Figure 3

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

En effet, les interrogations sont hiérarchisées ;
- si, par exemple, la durée d'une consonne est demandée ainsi :
1. la consonne appartient à un mot lexical,
2. ce mot est monosyllabique,
3. la consonne est initiale du mot,
4. ce mot est précédé d'un autre mot,
5. ce mot est suivi d'un autre mot,
6. la consonne est dans un environnement gauche et droit consonantique,

- et si aucune occurence ne répondait à ces six conditions, alors la durée retenue pour la consonne serait celle obtenue au terme de la question 5 (voire de la 4).

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

IV CONCLUSION

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

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

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

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

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

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

V BIBLIOGRAPHIE

[1] EMERARD F. et LIENARD J.S. (1980), "Quelques aspects de la traduction phonétique et prosodique du texte écrit", Revue d'Acoustique, 53, 97-101.

[2] MONNE J. (1983), "Programme de calcul de spectrogramme numérique", in "Traitement du Signal de Parole", ENST, Paris.

[3] BAILLY G. (1983), "Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique. Etablissement d'un modèle de génération", Thèse de Docteur-Ingénieur, INPG, Grenoble.

[4] O'SHAUGHNESSY D. (1984), "Design of a Real-Time French Text-to-Speech System", Speech Com., 3, 233-243.

[5] SORIN C. et BARTKOVA K. (1984), "Synthèse de plusieurs styles d'élocution : invariance et variantes prosodiques", 13èmes J.E.P., Bruxelles.

REMERCIEMENTS

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

La variation des mesures temporelles absolues et relatives de l'articulation de la parole

Eric Keller

Université du Québec à Montreal et
Centre de recherche, CHCN
4565 Queen Mary, MONTREAL, QC H3W 1W5, Canada

Abstract
A central prediction of the proportional hypothesis of articulator timing was examined with respect to two types of temporal variation: (1) Variation due to speech rate and (2) spontaneous variation. It was expected that the calculation of relative (proportional) time measures would induce significant reductions in variation in both conditions. However, only the variation due to speech rate underwent a systematic and significant reduction in variation through the calculation of relative time measures (Nsubj=13, Nobs=2,178). It is concluded that the proportional hypothesis cannot be maintained in its most general form and that the term "temporal invariance" is probably a misnomer in this context. Proportional measures serve only to reduce (by about 50%) variations induced by speech rate and do not eliminate such variation entirely.

Une hypothèse à la fois fort répandue que très simple veut que les événements articulatoires soient distribuées dans le temps comme des marques sur une bande élastique: Lorsque le débit de la parole est rapide, tous les intervalles entre événements articulatoires seraient réduits de façon proportionnelle par rapport à la situation où le débit est lent. Par corrélat logique, les mesures relatives (proportionnelles) seraient moins variables que les mesures absolues. Par exemple, le temps de la partie initiale d'un mouvement articulatoire de /k/ à /a/ dans la syllabe /ka/ varierait moins d'une instance à la prochaine quand il est calculé en tant que pourcentage (ou proportion) du mouvement entier que quand il est mesurée directement.

Ce principe provenant de la motricité générale a été défendu à plusieures reprises dans le contexte de la parole par les auteurs Kelso et Tuller (p.ex. Kelso, Tuller & Harris; Kelso, Saltzman & Tuller, 1986; Kelso & Tuller, 1987). Typiquement, ces auteurs ont rapporté des rapports relativement "invariants" entre la durée du segment /-ap-/ et le cycle vocalique /-api-/ dans la production du logathome /papip/, en dépit de modifications de débit et d'accent. Cependant, trois rapports récents ont remis en question cette hypothèse. Premièrement, Lubker (1986), en tentant de répliquer ces résultats avec des sujets suédois (Kelso et Tuller travaillent avec des sujets américains) n'a pas trouvé le même degré d'invariabilité inter-sujet, et même la variabilité intra-sujet restait relativement importante (aucune analyse statistique a été effectuée). Deuxièmement, Gentner (1987) a remis en question un grand nombre des travaux antérieurs en motricité générale appuyant la notion de proportionnalité. Selon des barêmes statistiques plus strictes, ces résultats serait suspects d'après Gentner. Finalement, Nittrouer, Munhall, Kelso & Tuller (1986) ont brièvement présenté des résultats allant à l'encontre de l'hypothèse proportionnelle antérieurement défendue par deux d'entre eux.

Ces différences pourraient refléter des subtilités empiriques ou elles pourraient être dues à des défauts méthodologiques (ou elles pourraient résulter des deux causes). En fait, deux déficits méthodologiques des recherches antérieures dans le domaine de la motricité de la parole sont qu'ils sont basées sur peu d'informations provenant de quelques sujets seulement et qu'ils négligent généralement d'effectuer une analyse statistique exhaustive. De plus, elles ne vérifient l'hypothèse qu'en examinant les effets de variations temporelles dues au débit et à l'accent (stress) et elles laissent de côté les variations temporelles spontanées. Nous avons décidé d'examiner l'hypothèse proportionnelle dans deux conditions de variation temporelle, (1) en variation due au débit et (2) en variation temporelle spontanée, en supposant que l'hypohtèse devrait s'appliquer de la même façon aux deux conditions. De plus, notre vérification de l'hypothèse est la plus vaste qui n'ait jamais été effectuée (13 sujets et 2178 observations) et elle implique une évaluation statistique rigoureuse.

Méthode

Treize locuteurs natifs du français québecois, entre 20 et 65 ans, 6 F et 7 M, ont contribué des informations pour cette étude (âge moyenne 37,2 ans, é.t. 17,6 ans). Tous ont prononcé la syllabe /ka/ au moins 25 fois (en moyenne 41,9 fois) en quatre conditions:

1. /ka/ en tant que syllabe longue, en répétition normale: /ka:ka:ka:.../
2. /ka/ en tant que syllabe courte, en répétition rapide: /kakaka.../
3. /ka/ en tant que syllabe longue et accentuée, dans le contexte "le macaque assommé".
4. /ka/ en tant que syllabe courte et non accentuée, dans le contexte "le lac à canard".

L'enregistrement articulatoire de ce matériel a été effectué par un système ultrasonique de mesures des mouvements du dos de la langue. (Pour des informations détaillées sur cette méthode, voir Keller & Ostry, 1983 et Keller, 1987). Dans le signal lissé représentant les mouvements verticaux de la langue pour la séquence /-akaka-/ (une courbe ressemblant aux deux bosses d'un chameau) et dans le signal sonore accompagnant, nous avons établi six repères temporels: 1. le début du mouvement descendant pour

le premier /ka/ (vélocité 0, P1), 2. le point de vélocité descendante maximale (P2), 3. le point où le reflet sonore de l'initiation des mouvements laryngaux pour le voisement du /a/ central deviennent visibles (P3), 4. le début du mouvement ascendant pour le /ak/ central (vélocité 0, P4), 5. le point de vélocité ascendante maximale (P5) et 6. le début du mouvement descendant pour le /ka/ final (vélocité 0, P6).

Par rapport à ces points, nous avons mesuré les intervalles suivantes: 1. le cycle entier (Tcycle, P1-P6), 2. le mouvement descendant (Tdesc, P1-P4), 3. le mouvement ascendant (Tasc, P4-P6), 4. le début du mouvement descendant (T1desc, P1-P2), 5. la fin du mouvement descendant (T2desc, P2-P4), 6. le début du mouvement ascendant (T1asc, P4-P5), 7. la fin du mouvement ascendant (T2asc, P5-P6) et le délai linguo-laryngal (LLD, P1-P3).

Des proportions ont été calculées, par rapport à Tcycle, pour toutes les mesures sauf Tcycle, et par rapport aux mouvements respectives, pour toutes les mesures sauf Tcycle, Tdesc et Tasc. Par exemple, pour toutes les 2178 mesures de T1desc, un pourcentage *T1desc/Tdesc*100* a été calculé. A la suite de cette opération, des coefficients de variation (é.t./moy.) ont été calculés pour toutes les mesures de chaque sujet. A cause de non normalité sévère, les coefficients de variation ont été soumis à une transformation logarithmique. Puisque les coefficients de variation logarithmiques (CdeVlogs) provenant de différentes conditions étaient semblables, ils ont été combinés pour fournir un seul indice de variation spontanée par sujet et par mesure. Les comparaisons entre les deux conditions de variation temporelle et entre les mesures relatives et absolues ont été effectuées au moyen de procédures ANOVA pour mesures répétées.

Résultats

1. Dans toutes les comparaisons, la variation temporelle due au débit était significativement plus importante que la variation spontanée (c.-à d. la variation à l'intérieur de chaque condition de débit) (Ncomparaisons=13, p<.001).
2. La variation de toutes les mesures absolues étaient significativement plus importante que la variation des mesures relatives correspondantes (par rapport au cycle, Ncomparaison=7, p<.01; par rapport au mouvement, Ncomparaisons=5, p<.001).
3. Cependant, la réduction de variabilité due au calcul relatif et la réduction de variabilité due à l'élimination des effets de débit étaient en interaction significative (p<.05) pour toutes les mesures temporelles, sauf pour le délai interarticulatoire (LLD, P1-P3). Ceci impliquait que des réductions de variation significatives n'étaient possibles que pour la variation temporelle due au débit et non pour la variation temporelle spontanée (à l'exception toujours du délai interarticulatoire LLD).

Discussion

Cette étude démontre à la fois la puissance que les limites de l'hypothèse proportionnelle. En ce qui concerne les variations massives dues au débit de la parole, un calcul relatif (ou proportionnel) des mesures temporelles peut effectivement servir à réduire la variabilité d'une mesure temporelle. Retenons toutefois que la variabilité n'est pas éliminée par cette opération. (Elle est typiquement réduite à 50% de la variabilité originale). Le terme "invariance" reste donc inapplicable dans le sens stricte du terme.

En ce qui concerne la variation moins importante à l'intérieur d'une condition de débit donnée (que nous avons appelée "variation spontanée"), un calcul relatif ne sert pas à réduire la variation de façon significative, sauf dans le cas du délai inter-articulatoire entre les activités de la langue et des cordes vocales (LLD).

La présence de résultats systématiques et non compatibles avec l'hypothèse de simple proportionnalité permet de rejeter l'hypothèse proportionnelle dans sa forme la plus générale et suggère qu'elle doit être nuancée selon le contexte. Une façon de reconsidérer la variabilité est de la traiter comme étant reliée à certains facteurs, tels que le débit, contexte linguistique, etc. Dans un travail utilisant des données semblables, Munhall (1985) a démontré que la proportion la plus importante de la variation temporelle était due aux modifications de débit induites par l'accentuation ("stress"). Les données présentes vont dans le même sens et appuyent donc une interprétation selon laquelle "l'invariance temporelle" de Kelso et Tuller ne serait qu'un autre nom pour "variation due au débit".

Ceci dit, il reste le problème de la différence entre mesures temporelles simples et la mesure inter-articulatoire LLD qui a été mis en évidence dans ce travail. Actuellement, nous ne disposons pas d'hypothèse particulièrement convaincante qui expliquerait ce phénomène. Son explication exigerait probablement des analyses additionnelles et possiblement le recueil de mesures inter-articulatoires plus variées.

Références

Gentner, D.R. (1987). Timing of skilled motor performance: Tests of the proportional duration model. *Psychological Review, 94,* 255-276.

Keller, E. (1987). Factors underlying tongue articulation in speech. *Journal of Speech and Hearing Research. Juin.*

Keller, E., & Ostry, D. (1983). Computerized pulsed echo ultrasound measurements of tongue dorsum movements. *Journal of the Acoustical Society of America, 73,* 1309-1315.

Kelso, J.A.S., Saltzman, E.E., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics, 14,* 29-59.

Kelso, J.A.S., & Tuller, B. (1986). Intrinsic time in speech production: Theory, methodology and preliminary observations. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelso, J.A.S., Tuller, B., & Harris, K.S. (1983). A 'dynamic pattern' perspective on the control and coordination of movement. In P. MacNeilage (Ed.), *The production of speech* (pp. 137-173). New York: Springer Verlag.

Lubker, J. (1986). Articulatory timing and the concept of phase. *Journal of Phonetics, 14,* 133-137.

Munhall, K.G. (1985). An examination of inter-articulator relative timing. *Journal of the Acoustical Society of America, 78,* 1548-1553.

Nittrouer, S., Munhall, K.G., Kelso, J.A.S., Tuller, B., & Harris, K.S. (1986). *Patterns of interarticulator phasing relations.* Présentation lors de la 112e réunion de l'Acoustical Society of America, 8-12 déc., Anaheim, CA.

Ostry, D.J., & Munhall, K.G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America, 77,* 640-648.

DUREES SYSTEMATIQUES DANS LES RIMES (C)VC EN FONCTION DES SEGMENTS ET DE L'ACCENT

Laurent Santerre

Département de linguistique

Université de Montréal

C.P. 6128, Montréal H3C 3J7 CANADA

Abstract. In Montreal French, on account of the phonological system of short and long vowels checked by the phonetic categories of fortis or lenis consonants, many schemes of relative lenghts are systematically produced in the rhymes. The two distinctions features of timbre and of lenght of the vowels are always relevant in the morphemes; if the vowel is short, the consonant is longer; if the vowel is long, the consonant is shorter. Syllable, morph, word and phrase boudaries have to be taken into account to improve the text-to-speech system.

Introduction

Cette étude fait suite à Santerre 1987 [1] et s'inscrit dans une recherche beaucoup plus vaste sur les paramètres prosodiques du français québécois en vue de la synthèse par règles. On y a établi avant tout la systématique des durées relatives dans les rimes à voyelles longues ou brèves entravées par des consonnes allongeantes ou abrégeantes ou neutres, en s'en tenant à l'accent de fin de syntagme terminal et de fin de syntagme intérieur sujet. La présente étude examine le jeu des durées relatives dans les mêmes rimes, mais cette fois dans diverses positions accentuelles ou différents cas de désaccentuation. Une question importante en effet se pose, celle de savoir si la distinction si nette entre patte et pâte sous l'accent est sauvegardée dans les morphèmes désaccentués, comme dans empattement, empâtement, empatté, empâté, etc. La durée phonologique et morphologique du français québécois est à coup sûr déterminante dans l'accentuation et l'intonation, mais jusqu'à quel point faut-il tenir compte des diverses frontières syntaxiques et morphologiques pour programmer une parole nature? Le français québécois comporte donc obligatoirement 17 voyelles, dont 8 sont des longues par nature en syllabe entravée. Le trait de durée s'ajoute à celui du timbre. Ces voyelles longues sont le /3/ de fête opposé au /E/ bref de faite, le /A/ de pâte opposé au /a/ bref de patte, le /O/ de côte oppos. au /o/ bref de cote, le /eu/ de jeûne opposé au /oe/ bref de jeune, plus les 4 nasales: /en/ de feinte, /an/ de fente, /on/ de fonte et /un/ de défunte.

Les consonnes obstruantes, les seules retenues pour l'instant comme coda dans les rimes, se regroupent par rapport à la durée et à l'allongement; ces distinctions ne sont pas phonologiques, mais phonétiques; les 3 occlusives sourdes sont longues et abrégeantes, les 3 constrictives sourdes sont longues et non abrégeantes, les 3 occlusives sonores sont brèves et non abrégeantes, enfin les 3 constrictives sonores sont brèves et allongeantes; c'est ce qui ressort de Santerre 1987.Il y a donc lieu d'examiner les durées systématiques dans les rimes à voyelles brèves entravées par les quatre types de consonnes, les quatre voyelles longues orales et les quatre nasales sous les mêmes entraves. Les voyelles brèves allongées se rapprochent considérablement des autres longues.

Le nouveau corpus reprend plusieurs mots du premier et présente les morphèmes à l'étude en 7 positions de frontières syntaxiques ou syllabiques susceptibles d'influer sur l'accent ou les degrés de désaccentuation. Le même locuteur québécois a lu ce corpus de la façon la plus neutre possible, sans effet stylistique aucun. Certaines phrases ont été lues plusieurs fois pour vérification. Les mesures ont été faites au mingo, très souvent sur les sonagrammes analogiques et au besoin sur les sonagrammes numériques beaucoup plus précis.

Ce nouveau corpus permet de contrôler des résultats comparables du premier corpus. Un troisième corpus plus élaboré avec le groupe des consonnes /b d g/ permet d'examiner un plus grand nombre de voyelles sans prendre des dimensions trop encombrantes; ces deux corpus font ensemble quelque 280 phrases. Les tableaux qui suivent montrent des résultats sur certains mots en particulier, ou présentent des résultats d'ensemble.

Analyse des frontières. (Exemples avec patte et pâte)

1. Le mot est prononcé isolément; la rime /a t/ ou /A t/ est dans une syllabe précédée et suivie de la frontière de phrase $. Accent primaire.

2. "Aimes-tu les pâtes": frontière du mot et frontière de phrase Pat$. Accent 1.; énoncé de 4 syllabes.

3."Le mot pâte me plaît. La rime est sous l'accent secondaire (2) et suivie d'une frontière intersyntagmatique.

4."Des pâtes maison". La différence avec le cas précédent consiste dans le fait que la frontière qui sépare les deux mots est intrasyntagmatique,

Frontières syll.	1. $(C)VC $	2. #(C)VC $	3. #(C)VC#CV		4.#(c)VC #CV		5.#(c) V!CV		6. #(c) V/CV$		7.+(c) VC+CV	
Accent:	1	1	2	0	2	0	2	0	0	1	0	1
patte /at/	13+20	9+17	8+14		8+12		8+8		9-13		7+13	
pâte /At/	25+15	18+13	13+11		11+10		12+7		12-12		10+8	
chante /ant/	24+11	20+14	14+10		11+11		12+6		12-12		13+12	
attache /aC/	14+25	12+20	11+12		8+11		9+12		9-15		10+12	
vache -	15+23	12+25	13+13		11+14		10+10		10-15		10+14	
relâche /AC/	22+18	20+19	18+14		12+10		13+12		13-15		13+12	
étanche /anC/	26+17	27+18	18+13		15+11		15+11		17-16		15+13	
nage /aJ/	24+14	25+14	14+8		13+9		11+7		14-10		16+9	
âge /AJ/	27+14	28+10	15+7		14+10		14+7		16-7			
mélange /anJ/	28+11	26+11	15+9		15+9		14+6		17-7		17+10	
faite /Et/	12+20	7+17	7+14		7+15		7+7		8-12			
tète -	11+17	11+18	8+13		7+9		8+8		7-12			
fête /3t/	23+13	17+12	13+12		11+11		13+8		11-13			
tête -	26+15	18+13	16+11		15+10		13+6		12-13		11+10	
tinte /ent/	26+11	24+9	16+13		15+10		16+6		15+11		17+11	
Grèce /Es/	12+18	9+23	10+12		9+13		9+10		8-15			
délaisse -		11+21							9-15		9+14	
engraisse /3s/	24+17	20+15	14+11		12+11		12+11		11-15		14+13	
pince /ens/	30+17	25+15	20+14		20+13		17+8		16-13		18+13	
relève /Ev/	18+10	19+10	10+7		9+6		9+6		12-8		11+8	
rêve /3v/	21+7	22+9	15+8		15+8		12+6		14-9			
évite /it/	11+20	9+18	7+10		7+9		7+7		8-13		6+14	
riche /iC/	13+19	9+20	8+13		9+16		10+14				8+14	
arrive /iv/	19+11	20+9	12+7		16+8		11+5		9-8			
avive -		18+8	11+6								14+9	
	V	C										
a E o + b d g: M:	13.5	12.5	11.7	11.1	10.2	7.4	9.3	7.6	10.	6.4	8.4 9.	9.5 7.2
S:	1.9	2.0	2.25	1.55	1.3	1.7	2.1	1.8	2.	0.78	1.9 2.7	1.3 0.9
0 3 + b d g:	21.7	10.2	19.	9.7	16.	8	14.	6.7	14.	6.2	12.5 8.1	11 9
	1.5	2.6	1.6	0.5	1.4	0.8	1.8	0.9	1.4	1.6	2. 0.58	0 0
an on + b d g:	22.5	10.5	20.7	9.5	14.	5.5	13.5	6.5	13.5	6.5	13.2 7.2	14 7.7
	4.5	2.4	3.1	3.8	1.6	1.3	1.9	1.0	2.4	1.29	1.89 2.2	0 1.15

Tableau 1. Durées de la voyelle et de la consonne dans les rimes, selon les positions accentuelles [m= moyenne; s= écart-type].

comme entre un mot et son qualificatif; accent 2.

5."D(e) la pâte à tarte" [dla pAtatart]: la rime à l'étude est divisée par une "frontière" de dérivation syllabique qui peut faire prononcer [pA ta tart]; en français québécois, j'ai souvent observé que la cohésion du morphème empêche cette dérivation syllabique. Dans ce cas, tout le morphème, noyau et coda, est sous l'accent 2.

6. Un empâté: ici la division du morphème /pAt/ est entière; le /A/ est en dehors de l'accent, mais se trouve en position pénultième, tandis que le /t/ tombe sous l'accent 1.

7. Empât(e)ment: l'élision du schwa place le morphème en position pénultième, normalement non accentuée en français, puisque, selon Dell [2], on ne peut avoir 2 syllabes accentuées de suite. La position du morphème dans le mot, sa distance par rapport à l'accent final, la nature de la consonne qui suit, etc, voilà autant de facteurs qui peuvent influer sur la systématique des durées relatives dans la rime, mais que nous ne pourrons pas examiner ici.

Le but de cette étude est double: l'aquisition de connaissances linguistiques fondamentales, et leur application à la synthèse par règles; mais je ne ferai état ici que de ce qui peut être utile pour la synthèse naturelle de la parole. Des tests ultérieurs nous diront s'il faut raffiner ou si l'on peut se permettre de généraliser sur les syllabes non accentuées. Nishinuma [3] montre que ces dernières sont plus sensibles que les syllabes accentuées.

Remarques sur le tableau 1.

1. Dans les 4 premières colonnes, les consonnes fortes (sourdes) qui entravent des voyelles brèves prédominent en durée dans la rime, ex: patte et attache. Au contraire, quand la voyelle est longue par nature, c'est elle qui prédomine en durée, et la consonne se réduit; dans ce cas, l'écart des durées relatives est positif, comme dans pâte et relâche, et il est négatif dans les rimes à voyelles brèves. Les consonnes faibles (allongeantes) entraînent des écarts positifs dans la rime. Quand l'entrave est faite avec une consonne neutre (faible et non allongeante), comme /b d g/, la voyelle brève et la consonne n'entraînent pas des écarts de durée systématiquement positifs ou négatifs. Par contre, si la voyelle est longue par nature, l'écart est nettement positif.

2. Cette systématique des écarts positifs ou négatifs en fonction de la nature des constituants des rimes sous l'accent se retrouvent aussi dans les positions 5 et 6, mais elle peut être influencée par la dérivation syllabique; tous les cas qui dérogent s'expliquent facilement. En position pénultième (7), normalement non accentuable en français selon Dell [2], il semble que les morphèmes longs prennent un accent secondaire plutôt que nul.

Durées vocaliques

Syl.:	1	2	3	4	5	6	7
a +C	14	12	11	8	9	9	10
A -	22	20	18	12	13	13	13
an -	26	27	18	15	15	17	15
a +t	13	9	8	8	8	9	7
A -	25	18	13	11	12	12	10
an -	24	20	14	11	12	12	13
E +t	11	11	8	7	8	7	
3 -	23	17	13	11	13	11	
en -	26	24	16	15	16	15	17
a +J	24	25	14	13	11	14	16
A -	27	28	15	14	14	16	
an -	28	26	15	15	14	17	
i +t	11	9	7	7	7	8	6
i +v	19	20	12	16	11	9	14
$\breve{V}$:m	12.7	9.8	9.0	8.25	8.5	8.9	7.9
S	1.24	1.23	2.12	1.41	1.41	0.17	1.94
$\bar{V}$	24	21.7	15	13.9	12.9	13.7	14.7
	3.57	3.55	2.75	2.7	2.11	2.51	2.75

Tableau 2. Durée des voyelles. Exemples et moyennes tirés du corpus. /C/= ch comme dans attache; /j/ = ge comme dans nage; $\breve{V}$ = voyelle brève ; $\bar{V}$ = voyelle longue.

Remarques sur les voyelles:

1. Dans les 7 colonnes, l'opposition de durée vocalique est pertinente, même en 6 où en raison de la coupure syllabique la voyelle est séparée de sa coda dans le morphème (voir aC et AC); dans le système phonologique, le trait du timbre entraîne nécessairement celui de la durée, c'est pourquoi le /i/ suivi de /v/ garde sa durée de brève (9 cs) dans la colonne 6: la coupe syllabique empêche l'allongement coarticulatoire. Le cas de /aJ/ dans nage ne constitue pas une dérogation à la scission du morphème: il s'agit plutôt d'une trace d'une ancienne durée des /A/ en pénultième qui sont passés à /a/. Cette tendance s'entend encore à Paris (le paquet [PAke]).

2. Dans la synthèse, il n'y aura pas lieu de distinguer la durée des brèves allongées de celle des longues par nature, (voir aJ, AJ, anJ). Par contre, les durées sous l'accent 1 (colonne 1 et 2) sont en général nettement plus longues que sous l'accent 2 (col. 3 et 4); dans la col.5, la voyelle du morphème se trouve aussi sous l'accent 2; en 6, la voyelle n'est pas en position accentuable, mais dans mes exemples elle est en pénultième, ce qui contribue à l'allonger légèrement. En pratique, du

moins dans un premier temps, on pourrait s'en tenir aux moyennes des brèves et des longues, sous l'accent 1 et sous l'accent 2; rien n'empêchera de faire par la suite des règles d'appoint au besoin, par exemple pour les pénultièmes, où tout reste à faire, le corpus ne comportant que des /m/ en début de syllabe finale. Le jeu des rencontres de consonnes est assurément très complexe et influe sur la durée de la voyelle qui précède.

Les consonnes

Syl.:	1	2	3	4	5	6	7
V̆+t m=	19.25	17.5	12.75	11.25	7.25	13.75	13.5
s=	1.5	0.58	1.89	2.87	0.95	2.2	0.71
V̄=t	13	12.2	11.4	10.8	6.6	12.2	11
	2	1.92	1.14	0.84	0.89	0.84	0.7
V̆+s C	21.25	21.8	12.5	13.5	11.5	15	13.5
	3.3	2.17	0.58	2	1.9	0	1.0
V̄+s C	17.25	16.7	13	12.25	10.5	14.75	15.25
	0.5	2.1	1.41	1.5	1.73	1.26	2.2
V̆+v J	11.7	10.25	7.25	6.7	7.0	8.7	8.5
	2.0	2.63	0.95	2.0	1.0	1.15	0.71
V̄+v J	10.7	10.0	8.0	9.0	6.3	7.7	9.0
	3.51	1.0	0.81	1.0	0.58	1.15	0
V̆+b d g	12.5	11.1	7.4	7.6	6.4	9.0	7.2
	2.1	1.55	1.76	1.81	0.78	0.76	0.96
V̄+b d g	10.4	9.6	6.7	6.6	6.4	7.9	8.0
	2.32	2.5	1.67	0.92	1.19	1.64	1.15

Tableau 3. Durées des consonnes

Remarques sur les consonnes

1.Les consonnes longues (les sourdes) qui suivent des voyelles brèves sont très longues sous l'accent primaire (pos. 1 et 2). Leurs durées chutent considérablement à l'intérieur des énoncés. Cette chute est particulièrement marquée dans la colonne 5 où, s'il y a déviation syllabique, la consonne tombe dans une syllabe non accentuable. En 6, la consonne est sous l'accent terminal (empatté, empâté) mais elle ne peut avoir la durée d'une coda puisqu'elle est en position d'attaque dans la rime suivante.

2.Les consonnes longues qui suivent des voyelles longues perdent une partie de leur durée au profit de la voyelle, c'est surtout le cas des constrictives (Santerre 87).

3.Dans les rimes où les voyelles sont longues ou allongées, la durée des codas est nettement plus courte que le noyau vocalique, et elle chute dans les colonnes des accents secondaires.

4.Le programme de synthèse pourra ne tenir compte dans un premier temps que des grandes différences entre les colonnes; c'est ce que je marque par les traits entre des sections de colonnes.

5.La durée de la consonne en position pénultième (7) est en général plus proche de celles de la colonne 6, c'est-à-dire qu'elle paraît affectée par un certain degré d'accentuation.

Conclusions

Les durées liées aux timbres phonologiques sont toujours aussi pertinentes que les timbres eux-mêmes, indépendamment des positions accentuelles, c'est-à-dire que l'analyse phonologique doit obligatoiremenet s'étendre aux morphèmes.

Les durées relatives dans les différentes rimes à vopyelles phonologiquement brèves ou longues (ou allongées), entravées par des consonnes phonétiquement longues ou brèves, abrégeantes ou allongeantes, sont nettement soumises sous l'accent primaire à une systématique qu'on peut résumer par la prédominance du noyau ou de la coda. L'écart positif (en faveur de la voyelle) peut être plus ou moins marqué, de même l'écart négatif en faveur de la consonne. Cette étude confirme Santerre 87.

La durée phonologique des voyelles continue à s'imposer dans les morphèmes sous l'accent secondaire, et même en dehors de tout accent. Les durées consonantiques ne sont pas phonologiques mais elles contribuent à la distinction des morphèmes longs ou brefs dans les diverses positions; la durée phonétique des consonnes compose avec la durée phonologique des voyelles pour créer la systématique des durées dans les rimes.

Le programme de synthèse des durées devra tenir compte des lois grapho-phonémiques du système vocalique et avoir accès à un dictionnaire d'exceptions lexicales d'usage. Ce travail est déjà fait pour le québécois (Bernardi 1986 [4]).

Il ne sera pas possible avant longtemps de se dispenser d'introduire dans la graphie de l'énoncé à synthétiser de nombreux marqueurs analytiques des niveaux syntaxique, morphologique, phonologique et métrique. Il faudra dans l'avenir apprendre à écrire pour l'ordinateur comme on apprend à écrire aujourd'hui selon le code grammatical.

Références

SANTERRE, L. (1987), "Systématique des durées segmentales dans les rimes syllabiques à voyelles longues et brèves par nature" Actes du Congrès international des Sciences phonétiques. Tallinn, URSS.

DELL, F. "L'accentuation dans les phrases du français", in Dell, Hirst et Vergnaud (1984), Forme sonore du langage, Hermann, Paris.

NISHINUMA, Y., et DUEZ, D. (1986). "Réception d'une phrase rythmiquement perturbée", 15 e J.E.P., Aix-en-Provence, pp. 87-88.

BERNARDI, D, (1986), "La synthèse par ordinateur du français montréalais", M.A., Dept. of Electrical Engineering, McGill University, Montréal.

TIMINGS INTERSEGMENTAL ET INTRASEGMENTAL EN FRANCAIS

SOCK R. OLLILA L. DELATTRE C. ZILLIOX C. ZOHAIR L.

Institut de la Communication Parlée
Institut de Phonétique de Grenoble
Université III
38400 St Martin d'Hères.

ABSTRACT

Presented in this study is an examination, on the acoustic level and across rate conditions, of intersegmental and intrasegmental phasings in close phonetic classes : VC vs. VCC. We are adopting here the psychomotoric recognition pattern paradigm, which provides us with a means of inferring motor programs, separate or generalized, from the various types of phasing structures observed. Another thrust of this study is also to compare our acoustic data with related results in movement studies, especially in the intrasegmental domain, for which data is less available.

INTRODUCTION

Les études portant sur le timing de la parole au niveau articulatoire, montrent une constance relative, en dépit des variations d'accent et de vitesse d'élocution, des gestes associés à des segments adjacents [1] et des gestes associés aux événements d'un seul segment [2]. De tels résultats obtenus pour le comportement d'articulateurs oraux ("intersegmentalement") et pour celui de coordinations orolaryngées ("intrasegmentalement") suggèrent la possibilité d'une invariance motrice généralisée.

A la lumière de cette littérature, nous examinerons pour le français, sur le plan acoustique, quelques phasages intersegmentaux (VC & VCC) et intrasegmentaux (VOT & VTT) à travers deux conditions de débits. Nous adoptons un paradigme de psychomotricité emprunté aux études basées sur le phasage, - par ex. celles sur la locomotion [3].

L'activité de production dans la parole - pour ce qu'elle a de cyclique - nous permet de déterminer, sur le plan acoustique, à partir d'événements répétitifs un cycle de détente (d'un relâchement à l'autre) et de retenir à l'intérieur de ce cycle quatre phases : VOT, D VOC, VTT et Silence. Nos résultats se concentrerons sur les phasages vocaliques et ceux du VTT et du VOT. Nous tâcherons de démontrer qu'au lieu de rechercher simplement une constance de timing relatif ou bien une invariance de phase, nous devons être plutôt attentif à une restructuration globale des patrons de phasage et ceci en passant d'une catégorie phonétique (d'une tâche) à une autre, comme cela se produit en fait de la marche au jogging(pour une confirmation de cette opinion dans une revue générale du modèle dit "proportionnel" cf. [4]). Et c'est à partir de ces indices que l'on pourrait inférer de véritables changements de programmes moteurs [5]. Sur le plan intrasegmental, nous examinerons les comportements du VOT [6] et du VTT (Voice Termination Time, cf. [7] pour les mesures; et [8] pour sa validation perceptive) ; et ceci dans ce même cycle détente puisqu'ils sont tous deux liés au timing du geste glottique en relation avec le relâchement et la closion supraglottique, respectivement. Si le timing intrasegmental du VOT est relativement bien connu (pour le français, cf. [9]), nous ne disposons pas de suffisamment d'information sur le rôle que pourrait jouer le VTT dans les coordinations consonantiques. Nous situerons nos résultats, par conséquent, dans la perspective du contrôle du geste glottique en relation avec les deux commandes de closion et de détente responsables de la tenue consonantique. Replacé dans le contexte des recherches actuelles, préoccupées par l'invariance temporelle relative des commandes articulatoires [10], cette étude doit permettre une comparaison des manifestations acoustiques de notre VTT avec les résultats obtenus par [2] examinant la production d'une coordination intrasegmentale.

LES PARADIGMES LINGUISTIQUES

Nous avons utilisé le corpus de verbes français suivant : empatter / empâter, téter / étêter, coter / écôter, égoutter / goûter. Dans certains contextes, ces verbes donnent de véritables paires minimales, permettant de tester les effets de la gémination consonantique sur la longueur vocalique (avec des paires du type : "nous l'empattons ? / nous l'empâtons ?" vs. "nous l'empatt't-on ? / nous l'empât't-on ?" Cf. [11] pour une description détaillée de ce corpus. Nous avons obtenu 12 répétitions de chaque item pour chaque locuteur sous deux conditions de débits, normal et rapide. Ce qui donne un total de 768 items par locuteur. Pour rendre compte du phasing inter/intrasegmental entre les classes VC et VCC, nous limiterons notre exposé aux voyelles extrêmes du corpus ([a], [a:], [u], [u:]), et ceci pour les premiers résultats sur 5 locuteurs appartenant à une base de données plus importante. Notre analyse se focalisera sur 3 d'entre eux : le locuteur J.P. pratique systématiquement l'opposition de quantité (fait rare en français d'aujourd'hui...); le locuteur R.L. est représentatif de l'ensemble des résultats obtenus dans cette étude ; le locuteur C.F. présente des structures de phasages nettement particulières.

Les signaux numérisés (à 16 KHz sur 12 bits) ont été étiquetés manuellement en événements à l'aide d'un éditeur de signal [12]. Un total de 8.898 événements a été ainsi détectés par locuteur.

Nous avons retenus 5 paramètres dans le champ VC : la phase vocalique (ou phase VOT plus phase D VOC (cette dernière présentant une structure formantique claire)) ; la durée de la tenue consonantique ; VOT ; VTT (du début de la closion à l'arrêt du voisement) ; la phase silencieuse ; le cycle détente. Pour plus de détails sur la procédure de mesures, cf. [13].

RESULTATS ET DISCUSSION

1. PHASAGES INTERSEGMENTAUX

. Locuteur J.P. (avec quantité vocalique)

En examinant les effets de la vitesse d'élocution sur le timing acoustique relatif de nos différentes classes phonétiques, nous avons observé, dans le cycle détente, une nette différence de phase vocalique en pourcentage (autour de 17%) entre les classes VC,V:C et les classes VCC,V:CC. Cette distinction, quoique bien significative (t=13.76) n'est pas pour nous surprendre puisqu'on peut s'attendre légitiment à une différence de pourcentage pour une voyelle suivie d'une consonne simple et une voyelle suivie d'une consonne double. Ce qui est intéressant, par contre, est le fait que les phases vocaliques pour les catégories VC et VCC demeurent relativement constantes dans le cycle détente, à travers des conditions de quantité et de vitesse d'élocution. La séparation reste efficace de telle sorte que les classes VCC en vitesse d'élocution rapide ne se confondent jamais avec les classes VC en condition de vitesse d'élocution normale. Nous avons observé aussi que les classes différaient nettement sur le continuum du cycle de détente. Le résultat est donc bien deux patrons temporels nettement différents pour les deux entités linguistiques.

. Locuteur R.L. (sans quantité vocalique)

Même si ce locuteur possède un système vocalique différent du locuteur J.P., la structure de phasage consonantique global est identique. La Figure 4 est une illustration adéquate de la stratégie intersegmentale adoptée aussi par le locuteur précédent. La stabilité relative à travers les conditions, évoquée plus haut, y est aussi bien illustrée, avec une différence en pourcentage de phase d'environ 20% (t=15.42). Cette différence moyenne en pourcentage de phase, couplée avec celle évidente du cycle (autour de 97 ms ; t=13.86) assure une maintenance de deux patrons de phases bien définis pour des tâches phonologiques différentes. Des manoeuvres similaires ont été systématiquement attestés avec 8 voyelles ([a], [a:], [E], [E:], [O], [O:], [u], [u:]) pour le français régional de Savoie [14].

.Locuteur C.F. (sans quantité vocalique)

En étudiant les résultats de ce locuteur, nous avons remarqué une transition, en quelque sorte linéaire, systématique entre les classes VC et VCC, transition due essentiellement à une non invariance de phase à l'intérieur de la classe VC, du débit normal au débit rapide. Les classes phonétiques ne sont pas non plus distinctes chez ce locuteur sur la dimension du cycle. Ce type de comportement, "déviant" par rapport aux autres locuteurs, nous permet d'observer une permanence toute relative de la différence entre programmes moteurs (lesquels vraisemblablement ne sont pas uniquement basés sur des relations de phases). Pour une discussion sur ce phénomène de transition de phase (réactualisant de vieilles notions en phonétique de la motricité, avec une étude du "doubling" et du "singling", proche de notre paradigme [15]), dans les perspectives nouvelles apportées par la synergétique, cf. [16].

2. PHASAGES INTRASEGMENTAUX

2.1. LES PHASAGES DU VOT DANS LE CYCLE DETENTE

Le rôle que joue le VOT dans le cycle détente semble être assez différent de celui de la phase vocalique et de celui du VTT (cf. infra). Comme pour le premier, il n'est pas surprenant que des dispersions semblables entre les deux classes en valeurs absolues produisent des pourcentages différents dans des cycles aussi différents que VC et VCC. Noter (par ex. pour le locuteur R.L., Figure 2) que les différences de pourcentages en valeurs moyennes sont statistiquement significatives (t=6.45). Une analyse plus détaillée des deux classes nous montre que la phase du VOT diminue d'abord de façon radicale avec l'augmentation du cycle, pour atteindre ensuite des valeurs relativement stables (autour de 240 ms). Il semble donc que les programmes pour les cycles longs tendent vers une meilleure stabilité.

2.2. LES PHASAGES DU VTT DANS LE CYCLE DETENTE

La tendance générale est de réduire le VTT en proportion avec l'augmentation du cycle détente. La Figure 3 donne la phase du VTT en fonction du cycle détente pour des productions du même locuteur R.L. Les coefficients de corrélation sont significativement négatifs aussi bien pour les consonnes simples que pour les doubles (r=-.83 et -.61 respectivement ; r=0.51 à p=0.10); la pente de régression étant significativement plus prononcée pour les consonnes simples que pour les géminées. Entre les deux classes segmentales, la différence de phase moyenne est négligeable (tout juste significative, t=2.90). De manière générale, on peut poser d'après ces résultats, qu'en ce qui concerne la phase du VTT et le cycle détente, la paramétrisation pour ce phasage intrasegmental est semblable pour les deux classes phonétiques, impliquant ainsi un programme moteur généralisé [5].

2.3. LES PHASAGES DU VTT ET LA TENUE

Partant du fait que LOFQVIST[17] a pu réaffirmer récemment que la phase de l'ouverture glottique restait relativement constante par rapport à la closion/constriction consonantique, à travers des variations de débit, nous avons examiné le pourcentage du VTT par rapport aux variations de la tenue dans nos deux conditions de vitesse d'élocution. Nous avons trouvé que la phase du VTT était négativement corrélée à la durée de la tenue. Cependant, une augmentation de la durée de la tenue pour la classe des géminées a moins d'effet réducteur en proportion sur la phase du VTT (la pente de regression pour la classe des consonnes simples étant généralement plus prononcée que celle des doubles). La Figure 1 est une illustration de cette stratégie. Il semble, en d'autres termes, qu'il y aurait plus de tendance à une constance de phase du VTT lorsqu'on augmente la durée de la tenue consonantique. Ces résultats se rapprochent de ceux obtenus par [18] dans une étude interlocuteurs. Noter que la phase du VTT tend à augmenter rapidement lorsqu'on diminue la tenue, sans que l'on atteigne pour autant, avec un maximum de 40% pour environ 75 ms de tenue, le seuil perceptif de voisement (environ 50%, LISKER, comm. pers.)

CONCLUSION

La problématique que nous avons soulevée au début de cette étude était celle d'une reformulation de la conception de l'invariance temporelle dans la parole. Réfutant les recherches sur l'invariance du timing relatif pour une seule phase dans un cycle [19], qui fournissent des résultats peu concluants [20](y compris dans le plan de phase, cf. [21]), nous avons préféré examiné les différences de structuration pluri phasique, comme cela apparaît dans la locomotion. De tels résultats montrent que le re-patterning global est le résultat de plusieurs types de comportements à l'intérieur d'une même phase. Quelques phases sont constantes (comme E2 dans le cycle de Philippson) ; quelques unes évoluent de façon continue (comme E1) ; enfin, peu de phases présentent de véritables discontinuités (comme F). De la même façon en parole, sur le plan acoustique, nous avons trouvé (Figure 5) que la seule phase avec une contribution decisivement discontinue aux différences de structuration, liée à des tâches linguistiques proches, était la phase dite vocalique (VOT + D VOC ou D VOC seul). Les

contributions du VOT sont du type E1, comme pour la phase silencieuse (plus proche de E3). Ceci pour nos mesures acoustiques. Mais il faut souligner ici que les études en timing intrasegmental du mouvement, devraient donner des résultats semblables. En ce qui concerne le VTT en particulier, il est intéressant de noter que les travaux de [17], peuvent être réintégrés dans de tels paradigmes de recherche. L'invariance relative de l'ouverture glottique maximale dans la phase de l'occlusion/constriction est, bien entendu, différente de ce que nous obtenons avec le VTT acoustique, qui lui varie inversement avec une augmentation de la tenue (Figure 1). Les résultats en mouvement, replacés dans le cycle de détente complet, donneront ainsi plus de différences en patrons que ceux trouvés pour notre VTT acoustique : des valeurs autour de 50 à 60% pour le mouvement d'ouverture glottique dans la tenue aboutiront à des valeurs nettement différentes dans les tâches VC contre VCC (disons 25% contre 40%). Il semble donc que ce phasing articulatoire soit plus discontinu que notre VTT, fournissant ainsi une contribution plus substantielle au changement de patrons de phases. Ceci reste à être quantifié afin d'évaluer la participation des phasages intrasegmentaux dans les cycles de parole. Une question pertinente à poser dès maintenant est la suivante : comment peut-on révéler de la façon la plus évidente la dépendance du timing laryngé des coordinations supraglottiques dans un programme moteur global ?

REMERCIEMENT : A Christian ABRY, qui a suivi ce travail de près, pour ses commentaires utiles.

REFERENCES BIBLIOGRAPHIQUES

[1]TULLER B. KELSO J.A.S. & HARRIS K.S. (1982) Interarticulator Phasing as an Index of Temporal Regularity in Speech. J. Exp. Psychol. HPP 8, 460-472.
[2]LOFQVIST A. & YOSHIOKA H. (1981) Interarticulator Programming in Obstruent Production. Phonetica 38, 21-34.
[3]SHAPIRO D.C. ZERNICKE R.F. GREGOR R.J. & DIESTAL J.D. (1981) Evidence for Generalized Motor Programs using Gait Pattern Analysis. J. Motor Behav. 13, 33-47.
[4]GENTNER D.R. (1987) Timing of Skilled Motor Performance : Tests of the Proportional Duration Model. Psychological Review 94, 255-276.
[5]SCHMIDT R.A. (1975) A Schema Theory of Discrete Motor Skill Learning. Psychological Review 82, 225-260.
[6]KLATT D.H. (1975) Voice Onset Time, Frication and Aspiration in Word-Initial Consonant Clusters. J. Speech Hearing Res. 18, 686-706.
[7]AGNELLO J.G. (1975) Measurement and Analysis of Visible Speech. in : Measurement Procedures in Speech, Hearing and Language. S. SINGH ed. 379-397.
[8]BERG van den R.J.H. (1986) The Effect of Varying Voice and Noise Parameters on the Perception of Voicing in Dutch Two-Obstruent Sequences. Speech Com. 5, 355-367.
[9]WAJSKOP M. (1979) Segmental Durations of French Intervocalic Plosives. in : Frontiers of Speech Communication Research, LINDBLOM & OHMAN eds. 109-123.
[10]FOWLER C. RUBIN P. REMEZ R.E. & TURVEY M.T. (1980) Implications for Speech Production of a General Theory of Action. in : Language Production 1, B. BUTTERWORTH ed., 373-420.
[11]ABRY C. SOCK R. BOE L.J. OLLILA L. DOUBLIER D. DELATTRE C. & ZILLIOX C. (1986) L'Organisation Temporelle des Voyelles et des Consonnes du Français. Durée Phonologique et Vitesse d'Elocution. Rapport CNET LANNION.
[12]BENOIT C. (1984) EDISIG : Encore un Editeur de Signal ?! 13èmes JEP du GCP du GALF, 211-213.
[13]ABRY C. BENOIT C. BOE L.J. & SOCK R. (1985) Un Choix d'Evénements pour l'Organisation Temporelle du Signal de Parole. 14èmes JEP du GCP du GALF, 133-137.
[14]DOUBLIER D. (1986) La Résistivité de l'Organisation Temporelle des Oppositions de Quantité dans le Français de la Chapelle d'Abondance (Hte-Savoie) face aux Variations de la Vitesse d'Elocution. TER de Sci. du Lang. Grenoble III, 87p. (dir. C. ABRY).
[15]STETSON R.H. (1951) Motor Phonetics : a Study of Speech Movements in Action. Amsterdam : North Holland.
[16]KELSO J.A.S. SALTZMAN E.L. & TULLER B. (1986) The Dynamical Perspective on Speech Production : Data and Theory. J. of Phonetics 14, 29-59.
[17]LOFQVIST A. & YOSHIOKA H. (1984) Intrasegmental Timing : Laryngeal-Oral Coordination in Voiceless Consonant Production. Speech Com. 3, 279-289.
[18]SOCK R. & BENOIT C. (1986) VOTs et VTT en Français. 15èmes JEP du GCP du GALF, 307-310.
[19]TULLER B. & KELSO J.A.S. (1984) The Timing of Articulatory Gestures : Evidence for Relational Invariants. J. Acoust. Soc. Am. 76, 1030-1036.
[20]BENOIT C. & ABRY C. (1986) Vowel-Consonant Timing Across Speakers. 12th Int. Congr. Acoust. A6-1.
[21]NITTROUER S. MUNHALL K. KELSO J.A.S. TULLER B. & HARRIS S.H. (1986) Patterns of Interarticulator Phasing Relations. 112th Meeting of Acoust. Soc. Am. Dec. 8-12.

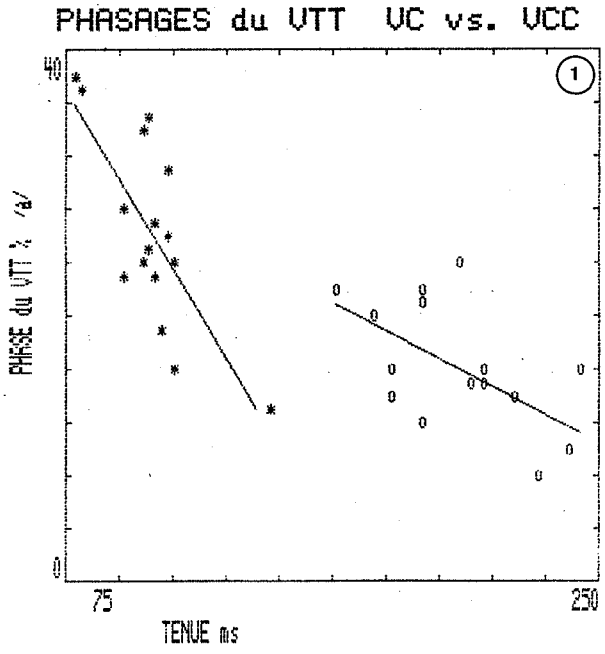


Fig. 1

Le phasage du VTT en fonction de la tenue (Mêmes conditions et même locuteur que pour les Fig. 2 - 5 ci-dessous).

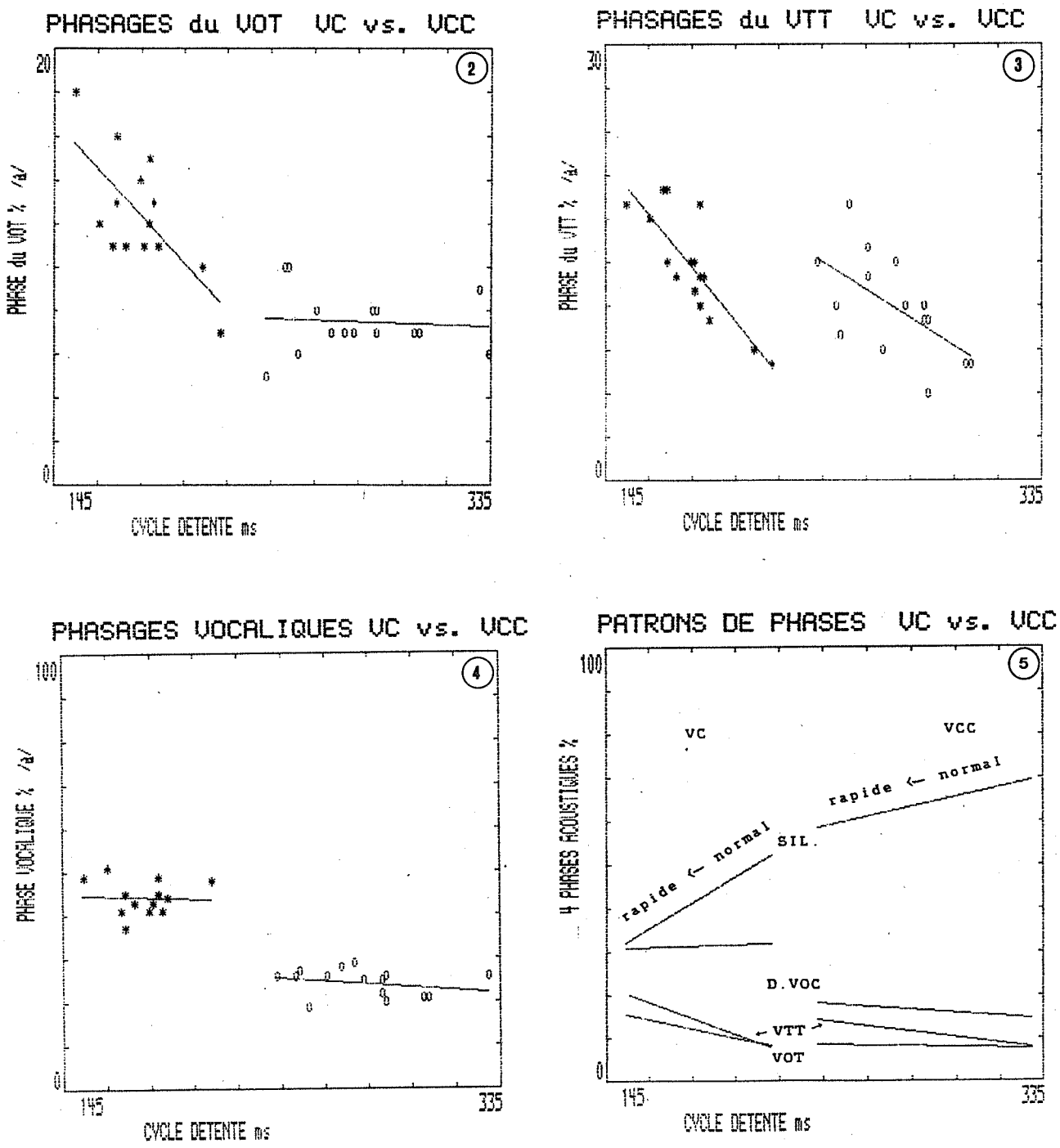


Fig. 2 - 5

Les phases du VOT (Fig. 2), du Voice Termination Time ou VTT (Fig. 3), la phase ouverte dite VOCalique (Fig. 4), déterminée de détente à closion (incluant donc la phase VOT et la phase d'émission d'une structure formantique définie ou D. VOC), sont données dans le cycle détente (de détente à détente) - avec les structures de phasage de l'ensemble (VOT, D. VOC, VTT, SILence) - pour : "...nous l'empattons ?" (VC → *) contre "...nous l'empatt't-on ?" (VCC → 0), sous deux conditions de débit (normal et rapide). Loc. R.L. (sans opposition de quantité vocalique).

DETECTEUR MORPHOLOGIQUE DU PITCH

A. BEN SLIMANE ET E.SELLAMI

IRSIT - ENIT (TUNISIE)

SUMMARY

This paper presents a new method for instantaneous pitch detection in the temporal domain. The method is original because it applies mathematical morphology to the speech signal in order to determine the variation of fundamental frequencies with time.

The speech signal is first filtered in the frequency band of its fundamental. Then morphological transformations are applied to detect the extrema which signify the periodicity of the speech signal. Finaly, the voiced-unvoiced decision is made using a simple decision schema.

The strength of this approach is that the determination of the instantaneous pitch is independent of all variations in the signal amplitude.

INTRODUCTION

La détermination de la fréquence du fondamental ou pitch est une opération assez importante vu le rôle qu'elle joue dans les applications de la parole (codage, synthèse, segmentation, . . .). Sa détection est assez délicate à cause de la variabilité des formes d'ondes du signal de parole [1].

Divers algorithmes de dètection du pitch ont été proposés. Certains opérent dans le domaine temporel se basant sur les méthodes de détection des pics, des vallées, des passages par zéro, et l'autocorrélation [2], [3], [4]. D'autres méthodes opérent dans le domaine fréquentiel comme le cepstre [5]. Des méthodes hybrides qui opérent aussi bien dans le domaine temporel que fréquentiel tel que le SIFT [6] et le LPC [7].

Dans cette communication, une nouvelle approche de détection du pitch basée sur les techniques de la morphologie mathématique est présentée. Cette méthode s'intéresse uniquement à la forme du signal de parole et sera classée parmi les méthodes temporelles. Ce traitement permet de donner les valeurs instantanées du pitch. N'utilisant pas des paramétres relatifs aux passages par zéro, aux seuillages, aux codages des extréma, la méthode morphologique est indépendante des variations de l'amplitude ainsi que des variations locales du signal temporel.

Aprés avoir introduit les opérateurs morphologiques nécessaires pour l'élaboration de la méthode à savoir l'érosion, la dilatation et la transformation par les extréma, le principe d' un détecteur morphologique du pitch est présenté et l'évaluation de la méthode est donnée.

I- TRAITEMENTS MORPHOLOGIQUES

La morphologie mathématique est utilisée dans l'analyse d'images, elle a pris naissance au milieu des années soixante avec les travaux de G. Matheron et J. Serra [8], [9], [10]. Rappelons que les traitements de la morphologie mathématique sont de type non linéaires, en ce sens que les transformations qu'elle utilise sont des transformations en tout ou rien à savoir d'une part les transformations ensemblistes (réunion, intersection, translation, . . .) et d'autre part les transformations par élément structurant (érosion, dilatation, . . .).

1- LES TRANSFORMATIONS MORPHOLOGIQUES DE BASE : EROSION ET DILATATION

Pour effectuer une transformation morphologique sur un signal de parole, nous définissons comme élément structurant une fenêtre temporelle de largeur P et d'origine J [11]. Nous faisons déplacer cette fenêtre sur le signal au pas de l'échantillonnage et nous affectons, dans chacune de ces positions, au point J la valeur maximale ou minimale trouvée dans la fenêtre. L'ensemble des valeurs relatives aux points J et auxquelles nous rajoutons (P-1) valeurs nulles placées aux extrémités forme le nouveau signal transformé. Ceci conduit à deux types de transformations. Nous obtenons une érosion dans le cas d'une transformation par le minimum et une dilatation dans le cas d'une transformation par le maximum. Le choix de la position de l'origine permet d'obtenir une transformation de type symétrique ou asymétrique (fig. 1).

Nous pouvons conclure d'aprés la figure, que l'érosion réduit les pics du signal et affecte des palliers de taille P aux vallées et que la dilatation crée des palliers de taille P autour des pics du signal et réduit ses minima.

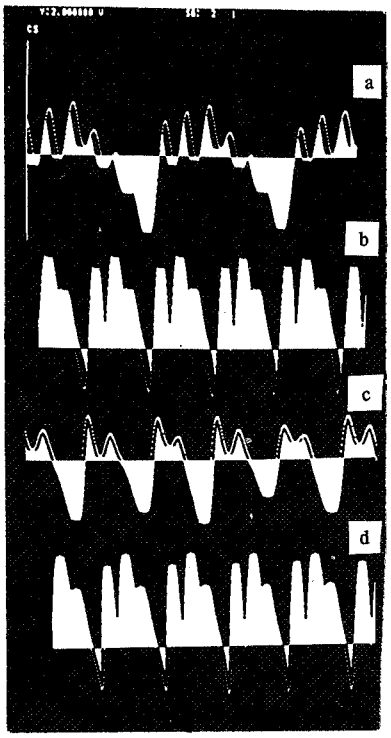


fig.1: a- Erosion asymétrique
b- Dilatation asymétrique
c- Erosion symétrique
d- Dilatation symétrique

2- TRANSFORMATION PAR LES EXTREMA :

Nous prenons une fenêtre symétrique de taille P=2 que nous faisons déplacer le long du signal au pas de l'échantillonnage. A chaque position de la fenêtre, l'origine est affectée de la valeur du premier minimum rencontré. Cette valeur sera gardée jusqu'au prochain extrémum. Le signal résultant de cette transformation est un signal multiniveaux dont les paliers correspondent à tous les extréma du signal (fig. 2).

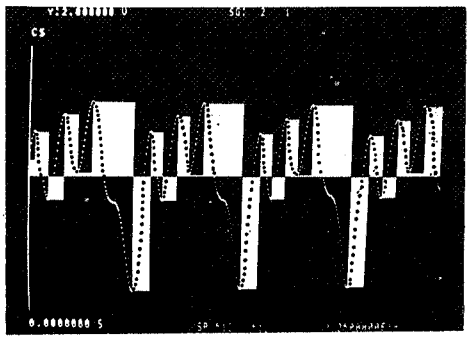


fig.2: Transformation par les extréma

II- DETECTEUR MORPHOLOGIQUE DU PITCH

1- PRINCIPE :

La méthode nécessite en premier lieu, un filtrage du signal de parole dans la bande de fréquence (80 - 600) HZ. Cette bande est choisie

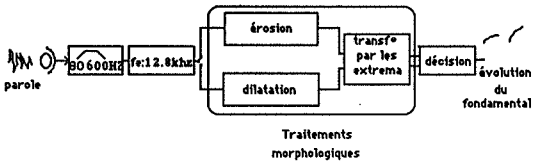


fig.3: Détecteur morphologique du pitch

de façon à couvrir aussi bien le pitch relatif à une voix masculine, féminine, ou celle de l'enfant. Le signal ainsi filtré est échantillonné à la fréquence de 12 KHZ pour donner une précision suffisante de la valeur instantanée du pitch (fig. 4a). Suivant la polarité du micro utilisé, nous opérons sur le signal filtré soit une érosion soit une dilatation de taille fixée en fonction du type de locuteur de façon à extraire soit les vallées dominantes soit les pics dominants (fig. 4b). Nous appliquons par la suite, une transformation par les extréma sur le signal érodé ou sur le signal dilaté et nous mesurons l'intervalle de temps séparant deux fronts descendants du signal multiniveaux relatif àl'érosion ou l'intervalle de temps séparant deuxfronts montants du signal multiniveaux relatif à la dilatation (fig. 4c). Nous obtenons ainsi une suite de valeurs qui seront traités à l'aide d'un système simple de décision qui sera décrit par la suite.

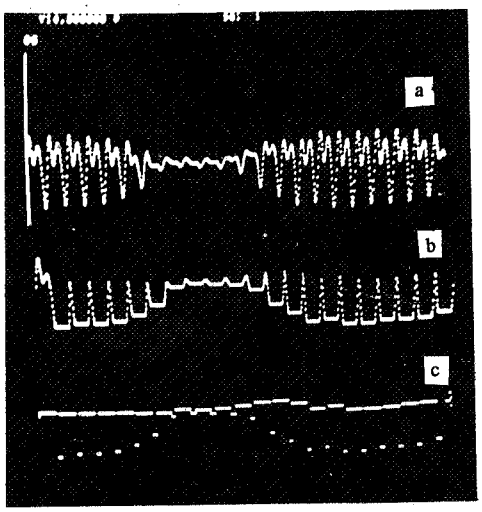


fig.4: a- Signal de parole filtré
b- Erosion
c- Transformation par les extrema

2- CHOIX DE LA TAILLE :

Le choix de la taille dépend de deux critères:

_ Le premier critère consiste à choisir une taille N qui permet d'araser les oscillations harmoniques (fig. 5a). Pour cela la taille doit être supérieure ou égale à la plus grande largeur des pics des harmoniques.

$$N >= Max (N1, N2, \ldots)$$

Dans le cas contraire, quand la taille choisie est inférieure au Max (N1, N2, . . .), (fig. 5b), les pics des oscillations harmoniques subsisteront et

la mesure sera donc faussée.

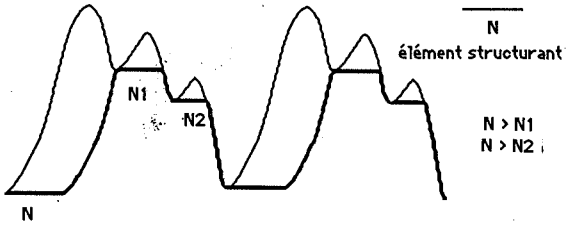


fig.5. a: Erosion de taille N >= Max(N1,N2,...)

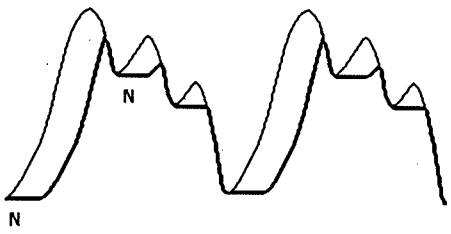


fig.5. b: Erosion de taille N < Max(N1,N2,...)

_ Le second critère consiste à choisir une taille N qui ne doit pas dépasser la limite supérieure de la bande du pitch du locuteur. La figure 5 montre la courbe de correspondance entre période P et taille N. La taille N est la taille minimale de l'érosion ou de la dilatation qui permet d'éliminer un signal périodique de période P et échantillonné à une cadence de 12 KHZ.

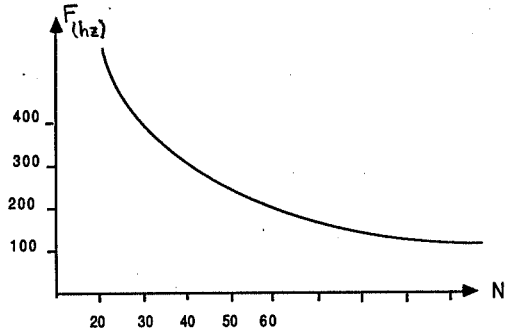


fig.6: Evolution de la période en fonction de la taille

La zone de variation du fondamental est supposée ne pas dépasser 200 HZ pour un locuteur homme,350 HZ pour un locuteur femme et 500 HZ pour un locuteur enfant. La courbe F = f (N) montre qu'il faut choisir une taille inférieure à 60 pour une voix d'homme, une taille inférieure à 34 pour une
voix de femme et une taille inférieure à 24 pour une voix d'enfant.

Compte tenu des limites obtenues par la fréquence d'échantillonnage, les valeurs suivantes ont été choisies :

Locuteur	Homme	Femme	Enfant
Taille	50	25	18

3- SYSTEME DE DECISION :

La transformation par les extrema appliquée au signal érodé / dilaté permet par comparaison entre deux valeurs successives de délimiter les zones voisées en utilisant un seuil de rejet; celui-ci étant obtenu de façon expérimentale. Dans les zones voisées, les valeurs obtenues suite à la transformation par les extréma, constituent les valeurs instantanées du pitch.

L'évolution du pitch est obtenue en faisant une moyenne des valeurs instantanées toutes les 10 ms et est représentée par les figures suivantes :

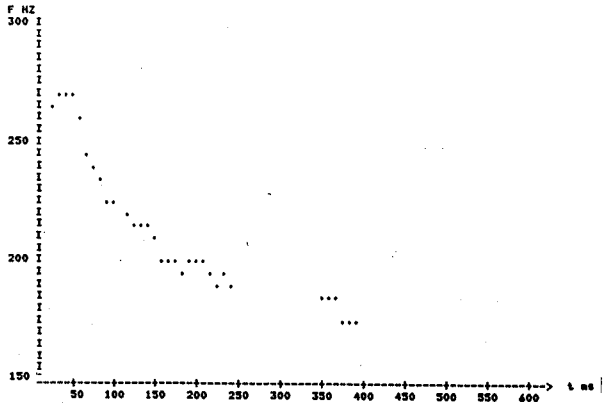


fig.7. a: Evolution du pitch pour l'élocution "CHAMATON"

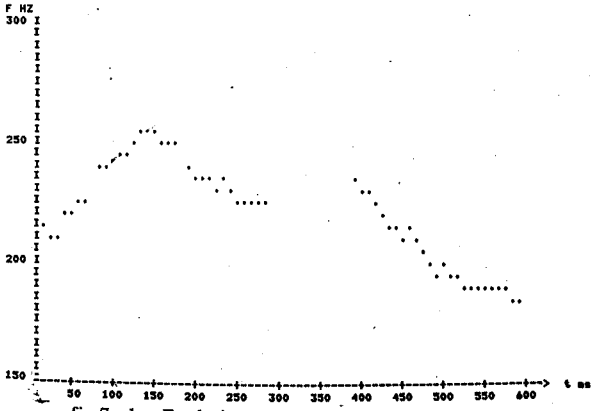


fig.7. b: Evolution du pitch pour l'élocution "HAMAMATAINI"

CONCLUSION

La méthode proposée constitue à notre sens une méthode originale et simple de mise en oeuvre. L'algorithme trés facile à implanter, permet de classer la méthode parmi les méthodes les plus simples et les plus précises de détection du pitch.

Cette méthode a l'avantage de déterminer les instants quasipériodiques d'excitation des cordes vocales; cette détermination est indispensable pour effectuer une analyse fréquentielle du signal en synchronisme avec le fondamental, ce qui apporte une amélioration

sensible aux méthodes de traitement du signal de parole.

Contrairement aux autres systèmes de détection, le détecteur morphologique du pitch ne nécessite pas de fenêtre de pondération, car la méthode s'attache à la détermination de la valeur instantanée du pitch.

La méthode présente une limitation qui réside dans l'identification du type du locuteur pour choisir la taille correspondante. Cette limitation peut être surmontée à l'aide d'un apprentissage qui permettra l'évaluation d'une taille moyenne d'analyse. Cette taille pourra être rectifiée en fonction de l'évolution du pitch.

REFERENCES

[1] J.S. Liénard, "Les processus de la communication parlée", Masson, 1977.

[2] B. Gold and L.R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain", J. A. S. A, Vol.42, pp. 442-448, Aug.1969.

[3] N.J. Miller, "Pitch detection by data reduction", IEEE Trans. Acoust. Speech, Signal processing, Vol. ASSP-23, pp. 72-73, Feb 1975.

[4] M.J. Ross, H.L. Shaffer, A. Cohen, R. Frendberg and H.J. Manley, "Average magnitude difference function pitch extractor", IEEE Trans. Acoust. Speech, Signal processing, Vol. ASSP-22, pp. 353-362, Oct. 1974.

[5] A.M. Noll, "Cepstrum pitch determination", J. A. S. A, Vol. 41, 1967.

[6] J.D. Markel, "The Sift Algorithm for Fondamental Frequency Estimation", IEEE, Audio Electroacoust, Vol. AU-22, 1972.

[7] J. Makhoul, "Linear Prediction : A Tutorial Review", Proc. IEEE, Vol. 63, pp. 561-580, Apr. 1975.

[8] G. Matheron, "Les variables régionalisées et leur estimation", Masson, 1967.

[9] G. Matheron, "Eléments pour une théorie des milieux poreux", Masson, 1967.

[10] J. Serra, "Introduction à la morphologie mathématique", Cahiers du Centre de morphologie mathématique, Ecole des Mines, Fontainebleau, N°3, 1969.

[11] B. Zouabi, N. Ellouze et A. Ben Slimane, "Traitement morphologique de signaux unidimensionnels", Onziéme Colloque GRETSI, Nice Juin 1987.

[12] A. Ben Slimane, "Contribution au traitement du signal de parole par des techniques de morphologie mathématique", D.E.A, ENIT, Tunisie,1986.

[13] M.J. Cheng, "A comparative performance study of several pitch detection algorithms", M.S. Thesis, Mass. Inst. Technol., Cambridge, June 1975.

[14] E. Sellami, "Assistance visuelle à la rééducation vocale prosodique", D.E.A, FST,Tunisie 1985.

DEFINITION D'UNE SYNTAXE DES STRUCTURES PHRASTIQUES ADAPTEE A LA PROSODIE

G. CAELEN-HAUMONT

LABORATOIRE DE LA COMMUNICATION PARLEE
46, AVENUE F. VIALLET 38031 GRENOBLE CEDEX

ABSTRACT

A syntax which fits prosodic information has been developped. This syntax may be considered in 2 parts : the first one lays upon the evaluation of the syntactical distance between constituents (hierarchy, level of syntactical cohesion, transformation of the structure by permutation of units, etc ...); the second one upon a morphosyntactical (i.e. syllabic) description of the different constituents which may offer a specific prosodical behaviour : the first syllable of the sentence / syntagma / word levels; syllables pertaining to a lexical/ independent syntactical / dependent syntactical unit; belonging to monosyllabic / plurisyllabic unit; syllables in the initial / medium / final position. This description is suited to any text with a simple or complex syntax.

These various syntactical informations with semantic, phonetical and prosodical other ones, are coded, first for the morphosyntactical analysis, through 22 labels which arrange oppositions, and then for the analysis of the distance, through a quantification. These codes are included in a data base (36 readings of a 50 word text by 12 speakers).

INTRODUCTION

La caractérisation de l'analyse syntaxique qui est proposée dans cet article, est d'être appropriée aux informations que peut délivrer la prosodie : elle présentera donc, autant que faire se peut, un degré de précision conforme à l'objet qu'elle tente de décrire. Conformément aux études antérieures (1), nous pensons en effet que la prosodie peut fournir des indications sur la classe des mots (lexicaux/non lexicaux), sur le degré d'autonomie syntaxique (des mots grammaticaux), sur la structuration des unités linguistiques (monosyllabes/plurisyllabes), sur la structure constituante de la phrase, sur la nature de certaines relations syntaxiques en termes de distance, mais vraisemblablement pas en français, sur l'identité des constituants en termes de micro-classes (ex: noms/verbes) ou de fonctions syntaxiques associées aux catégories (ex: SN1/SP, SN1/SN2...).

Dans cette optique, nous avons défini deux types d'analyse syntaxique complémentaires, l'une de type morphosyntaxique, l'autre reposant sur des critères de distance. L'une et l'autre constituent une liste d'étiquettes intégrées dans une base de données, posées de manière manuelle. Pour la première, les étiquettes sont à double caractère alphabétique, pour la seconde, à simple caractère numérique. Ce système de codes, issu d'une connaissance a posteriori des structures prosodiques, est destiné à être testé statistiquement dans une phase ultérieure du travail.

ANALYSE MORPHOSYNTAXIQUE

Pour ce type d'analyse, deux unités fondamentales ont été retenues : d'une part la syllabe, d'autre part le "constituant minimal". La justification du choix de la syllabe repose sur le fait qu'elle constitue une unité sur les plans syntaxique et prosodique. Sur le plan prosodique, elle se définit essentiellement aux niveaux temporel et mélodique : outre les voyelles, il est bien connu par exemple qu'en fin de groupe prosodique, les consonnes subissent un allongement notable, ou encore que les consonnes voisées environnant la voyelle accentuée peuvent encadrer les variations de Fo. Ce choix d'ailleurs n'est pas contredit au niveau structural puisque l'unité linguistique minimale, parfois réduite jusqu'à un simple phonème vocalique, est le monosyllabe, qu'il soit lexical ou grammatical.

Si l'unité élémentaire est la syllabe, l'unité de référence, comme dans l'analyse de distance, est le "constituant minimal" (ou CM). Par ce terme, nous entendons une unité de nature syntactico-prosodique. Sur le plan syntaxique, elle correspond à une unité supérieure au "mot" (mot ou unité linguistique entre 2 blancs) conçue comme point d'aboutissement de l'analyse en constituants. Cette unité dominant directement le niveau superficiel, a un niveau quelconque de profondeur : ainsi des constituants majeurs comme par exemple des SN1 sans complément côtoient d'autres constituants de niveau beaucoup moins profond comme les SP issus de SP ... Sur le plan prosodique, ces CM sont définis comme des unités susceptibles de former les groupes prosodiques minimaux repérables à leur frontière de droite par une proéminence prosodique, associant au minimum une variation remarquable des niveaux temporel et mélodique. Ces groupes prosodiques se constituent d'autant plus facilement qu'ils sont plus longs et introduits par des déterminants ou prépositions qui se comportent comme des articulateurs prosodiques. La distinction entre la profondeur des constituants est introduite par la suite dans le calcul des distances.

L'analyse morphosyntaxique distinguant plusieurs niveaux permet de représenter tout texte avec des phrases complexes avec le degré de précision ajusté à l'objet syntactico-prosodique décrit.

FRONTIERES DE CM

Au niveau formel, une distinction est faite entre les constituants minimaux situés en début de phrase et les autres. Plus précisément, on distingue la syllabe initiale (ou un monosyllabe) du CM à l'initiale de phrase des autres syllabes initiales de CM. Prosodiquement en effet, la syllabe initiale de phrase est marquée aux niveaux énergétique, mélodique et temporel. L'information est délivrée par la première étiquette du code : P pour le premier cas, G dans le second.

CLASSES DE MOTS

Au niveau inférieur à celui des CM (i.e. celui du mot), 3 classes de mots aux propriétés prosodiques spécifiques, ont été retenues :

a/ Classe lexicale

Cette classe de mots réunit sans doute les causes de variabilité prosodique les plus nombreuses : les mots lexicaux sont les véhicules privilégiés du sens aux niveaux objectif, subjectif et affectif; ils sont par ailleurs le noyau de l'organisation syntaxique de la phrase; ils présentent au niveau formel, une structuration syllabique très diverse. Le symbole correspondant est le deuxième caractère du code, M ou L selon que les lexèmes sont des mono- ou des plurisyllabes.

b/ Classe des "mot-outils"

Selon A. SAUVAGEOT (2), il existe 4 types de mot-outils : les "déterminatifs nominaux" (articles, démonstratifs, possessifs...), les "déterminatifs verbaux" (pronoms), les "particules rectives" (prépositions) et les "éléments articulatoires" (conjonctions). Et de fait, les mot-outils n'ont pas tous un comportement uniforme au regard de la prosodie. En excluant certaines situations de parole spécifiques comme par exemple la lecture en mot à mot, on remarque en effet que mots lexicaux et mots non lexicaux présentent tantôt une tendance à la cohésion par attraction du noyau lexical, tantôt une tendance à l'indépendance. Les déterminatifs nominaux et verbaux, les particules rectives manifestent une tendance à la cohésion, au contraire, les éléments articulatoires se particularisent par le fait qu'ils ont la possibilité de se rendre indépendants.

La première catégorie constitue une classe de mot-outils non autonomes, de symbole N ou O selon leur nombre de syllabes (cf supra). Elle se distingue généralement par un rythme plus accéléré, une énergie moindre, un Fo d'allure transitoire ou dans un registre moyen-bas, l'absence de proéminence et de pause subséquente. Les mot-outils autonomes que nous appelons "relatants", de symbole R ou S, peuvent se comporter de manière semblable ou inversement peuvent se caractériser par un rythme comparativement plus lent, une énergie et Fo variables, supportant éventuellement des maxima, une possible accentuation suivie au besoin d'une pause, en somme un comportement prosodique proche de celui des mots lexicaux. Cette liberté de comportement prosodique ne se retrouve pas (sauf contraintes exceptionnelles de parole) chez les mot-outils non autonomes.

STRUCTURATION SYLLABIQUE

Des différences dans le nombre des syllabes induisent également des oppositions au niveau prosodique et plus spécifiquement au niveau temporel. Sont essentiellement concernés, les mots lexicaux : les monosyllabes lexicaux sont en effet, toutes choses égales par ailleurs, à la fois plus longs que les monosyllabes grammaticaux et qu'une syllabe quelconque (mais non finale accentuée) d'un plurisyllabe. Le symbole correspondant est M, en premier caractère.

Concernant le décompte des syllabes et les problèmes relatifs au /ə/ final de mot, nous avons adopté une mesure de simplification et d'uniformisation : la syllabe finale des mots à /ə/ final est souvent bifide, la (les) consonne(s) initiale(s) subissant un allongement au contact de la voyelle accentuée, et le /ə/ final étant, s'il n'est pas absent, plus ou moins bien réalisé. Ces considérations prosodiques nous ont amenée à considérer les plurisyllabes à x syllabes comme des plurisyllabes à x-1 syllabes + V (éventuellement si x = 2, des monosyllabes), soit l'écriture suivante du plurisyllabe :

```
/aR  ti  Klə/  ---> /aR  tikl  ə/
 IL  SL  FL          IL   FL    V
```

avec IL : syllabe initiale de mot lexical
SL : syllabe intermédiaire "
FL : syllabe finale "
V : /ə/ final

Cette méthode a l'avantage de neutraliser les différences de localisation de l'accentuation en fonction des mots traditionnellement appelés masculins et féminins et de rendre leur structuration syllabique compatible. Sur un plan pratique, elle offre en outre l'interêt de simplifier les calculs de quantification concernant les phonèmes sous l'accent.

A l'intérieur des plurisyllabes, pour des questions de démarcation de frontières de mots, de structure accentuelle, ou de composition lexicale, il est utile de distinguer syllabe initiale, intermédiaire(s) et finale, dont les symboles sont respectivement I, S, F en premier caractère. Les oppositions peuvent reposer sur les variations temporelles, énergétiques et/ou temporelles.

4/ Description par traits

Parmi plusieurs représentations possibles de cette grammaire que définissent ces différentes unités au comportement prosodique spécifique, nous avons retenu, pour sa simplicité de lecture, une description par matrice de traits (cf Tableau 1).

En résumé les 2 caractères des symboles ont pour signification :

a/ premier caractère

- P : constituant minimal à l'initiale de phrase
- G : " non initial
- I, S, F : syllabe initiale, intermédiaire, finale de plurisyllabe
- M : monosyllabe

I et M correspondent à des unités inférieures au CM.

b/ deuxième caractère

- L, M : classe lexicale (plurisyllabe, monosyllabe) au niveau du constituant minimal (GL, GM), parfois à l'initiale de phrase (PL, PM)
- O, N : classe des mot-outils non autonomes dans les mêmes conditions que précédemment
- S, R : classe des mot-outils autonomes ou "relatants" dans les mêmes conditions que précédemment.

FRONTIERES DE CONSTITUANTS			STRUCTURE SYLLABIQUE				CLASSES SYNTAXIQUES			
INITIALE DE PHRASE	NON INITIALE DE PHRASE		MONO SYLLABE	PLURI-SYLLABE			LEXICAL	NON LEXICAL		
	INITIALE DE CM	NON INITIALE DE CM		INITIALE	INTERMEDIAIRE	FINALE		AUTONOME	NON AUTONOME	
+	−	−	−	+	−	−	+	−	−	PL
+	−	−	+	−	−	−	+	−	−	PM
+	−	−	−	+	−	−	−	+	−	PS
+	−	−	+	−	−	−	−	+	−	PR
+	−	−	−	+	−	−	−	−	+	PO
+	−	−	+	−	−	−	−	−	+	PN
−	+	−	−	+	−	−	+	−	−	GL
−	+	−	+	−	−	−	+	−	−	GM
−	+	−	−	+	−	−	−	+	−	GS
−	+	−	+	−	−	−	−	+	−	GR
−	+	−	−	+	−	−	−	−	+	GO
−	+	−	+	−	−	−	−	−	+	GN
−	−	+	−	+	−	−	+	−	−	IL
−	−	+	−	−	+	−	+	−	−	SL
−	−	+	−	−	−	+	+	−	−	FL
−	−	+	+	−	−	−	+	−	−	ML
−	−	+	−	+	−	−	−	+	−	IS
−	−	+	−	−	+	−	−	+	−	SO
−	−	+	−	−	+	−	−	−	+	
−	−	+	−	−	−	+	−	+	−	FO
−	−	+	−	−	−	+	−	−	+	
−	−	+	+	−	−	−	−	+	−	MR
−	−	+	−	+	−	−	−	−	+	IO
−	−	+	+	−	−	−	−	−	+	MO

TABLEAU 1 : MATRICE DE TRAITS MORPHOSYNTAXIQUE

Dans le contexte P ou G, par suite de neutralisation d'oppositions, L signifie syllabe initiale de mot lexical; dans les autres contextes I, S, F, M, simplement mot lexical. Parallèlement, dans le contexte P ou G, O et S signifient syllabe initiale de mot-outil respectivement non autonome et autonome; dans les autres contextes O et S signifient seulement mot-outil respectivement non autonome et autonome avec une restriction de contextes pour S (un seul contexte I). En d'autres termes, comme on peut le constater sur la matrice de traits, il y a neutralisation des oppositions dès la deuxième syllabe (ou syllabe intermédiaire selon le terme générique) 1° entre les catégories de constituants 2° entre les mot-outils autonomes et non autonomes.

DISTANCE SYNTAXIQUE

Parmi les méthodes d'évaluation de la distance des constituants, celle de COOPER et alii (3) qui repose sur la prise en compte de la hauteur du noeud dans une perspective dynamique, avec quantification différente selon qu'il s'agit d'une dépendance à droite ou à gauche, est séduisante. Cependant elle ne prend en compte ni la nature du lien syntaxique entre constituants ni les considérations de nature rythmique. Le type d'analyse que nous présentons intégre ces informations (qui se traduisent au niveau prosodique par des variations importantes) : le terme de "distance" est donc à prendre dans un sens large.

Le type d'analyse que nous proposons offre une description complémentaire à l'analyse morphosyntaxique à plusieurs égards. Tout d'abord, il s'agit d'une quantification, c'est-à-dire d'une appréciation numérique de la distance des constituants. D'autre part, ce type

d'analyse restaure les oppositions de profondeur des constituants jusqu'au niveau superficiel non compris : c'est que d'un point de vue prosodique, les oppositions syntaxiques plus profondes que celles du niveau superficiel relevant donc au mimimum des CM, déterminent généralement des variations prosodiques majeures (Fo, durée, énergie), tandis que les variations prosodiques au sein des CM relèvent plus des problèmes de démarcation au sein d'une cohésion d'ensemble, que de l'actualisation d'une distance entre mots : les variations nécessairement mineures, ne concernent généralement pas les 3 paramètres à la fois.

En ce qui concerne le cas des déterminants élidés, prosodiquement ils ne sont pas différenciables d'une syllabe initiale d'un mot débutant par une consonne. Un code permet de ne pas les inclure dans le calcul de la distance.

Le calcul de la distance résulte d'une analyse en 4 points délivrant chacun un poids que l'on additionne successivement.

HIERARCHIE DES CONSTITUANTS

Une première analyse repose sur l'analyse en constituants immédiats en fonction de leur niveau dans la hiérarchie, le niveau le plus superficiel recevant le poids 1 (niveau mots), le pas étant de +1 à chaque changement de niveau vers les structures profondes. Les autres analyses concernent désormais les constituants de niveau supérieur aux mots.

LIEN SYNTAXIQUE DES CONSTITUANTS

Une nouvelle pondération est attribuée en fonction de la nature du lien syntaxique des CM. Plus fort est le lien syntaxique, moindre est le poids.

La fin absolue de constituant (ni subordination ni coordination avec le CM suivant) reçoit le poids +3; en finale de phrase, elle correspond au poids +5. Un CM suivi d'un CM coordonné reçoit le poids +2, mais suivi d'un CM subordonné, le poids +1.

Un cas particulier concerne les proclitiques qui dans la structure constituante ont statut de CM alors que prosodiquement ils se comportent comme des unités grammaticales non autonomes : ils sont traités comme des constituants de niveau superficiel avec attribution du poids définitif +1.

DEPLACEMENT DES CM DANS LA STRUCTURE

Une transformation de la structure, par rapport à une organisation normative, par permutation des constituants, est également prise en compte dans le calcul de la distance.

Il faut souligner que l'on prend en considération la structure de phrase telle qu'elle est réalisée et non normée, même si le niveau des constituants dans la hiérarchie est modifié par rapport à la structure standard. Sur le plan prosodique en effet, l'ordre des constituants est d'une importance primordiale.

Selon notre convention, le déplacement correspond à un poids de +2.

RYTHME

Il est bien connu que des critères de rythme ont également une influence sur la structure prosodique des constituants. Nous inspirant des travaux de F. DELL (4) sur l'eurythmie, et dans un objectif de simplification, nous proposons une pondération par niveau successif des constituants. La pondération s'effectue en parcourant la suite des CM de gauche à droite.

Lors du premier passage, on attribue un poids de +2 aux CM possédant un nombre de syllabes supérieur à 4. Il semble en effet que 5 syllabes constituent en effet une "substance" phonique suffisante pour susciter la formation d'un groupe prosodique. Au deuxième passage, on compare les CM restants sous l'angle de leur nombre de syllabes : lorsque le CM de droite possède au minimum 2 syllabes de plus que le CM de gauche, on affecte alors le poids +2 au CM de droite.

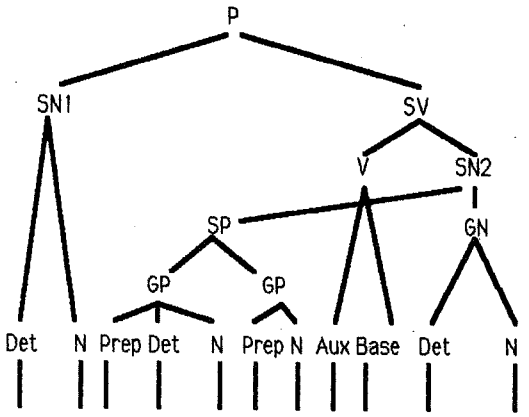
Si à ce stade, les CM ont le même nombre de syllabes (à une près), en remontant niveau par niveau dans la structure profonde, on regroupe 2 à 2 les CM en procédant de la droite vers la gauche. Les CM ainsi concaténés se voient attribuer le poids +2. Voici un exemple détaillant le processus:

nb de syll.	2	5	2	2	2	2

le chat de l'arrière grand-père a vu le bec du merle de Paul

Niveau CM	-/	+2 /	-/	-/	- /	- /
				<----------		
Niveau +1				+2		

Un autre exemple permet de récapituler l'analyse de la distance en ses 4 points et d'éclairer le procédé de la quantification :



	L'oiseau d'	**une**	**tasse**	**de**	**lait**	**a**	**pris**	**trois**	**gorgées**
HIERARCHIE	**4**	**1**	**2**	**1**	**3**	**1**	**4**	**1**	**6**
LIEN SYNT.	**3**		**1**		**3**		**1**		**5**
RYTHME					**2**				
DEPLACEMENT					**2**				
TOTAL	**7**	**1**	**3**	**1**	**10**	**1**	**5**	**1**	**11**

Cette quantification, certes issue d'une approximation subjective, a pour but de s'ajuster aux observations. Les différents critères qui entrent dans la définition de la distance et que l'on a retenus déterminent en effet les causes d'une variabilité importante des paramètres prosodiques : plus la distance est grande entre les CM, plus la pause tend à être respectée et longue, plus le ralentissement à la finale du CM est manifeste, et la variation de Fo importante.

CONCLUSION

Cette syntaxe opère donc sur tout texte quelle que soit sa complexité syntaxique. Elle offre un niveau de description ajusté à la quantité d'infomation que peut délivrer la prosodie. Elle tente également de quantifier la "distance" qui sépare deux CM en tenant compte de leur hiérarchie dans la structure, du type de leur lien syntaxique, du rythme et d'un éventuel déplacement dans la structure.

Cette phase de l'étude est complétée par une expérimentation : cette expérimentation de nature statistique (analyse des corrélations en particulier) doit permettre, sur une base de données (36 lectures par 12 locuteurs avec 3 répétitions) de valider ce modèle, comme pour le modèle sémantique proposé précédemment [5] [6], et si besoin d'ajuster les seuils.

REFERENCES BIBLIOGRAPHIQUES

[1] G. CAELEN-HAUMONT
Structures prosodiques de la phrase énonciative simple et étendue, Thèse de 3ème cycle, Hamburger Phonetische Beiträge, Band 34, Hamburg Buske, 1981.

[2] A. SAUVAGEOT
Français écrit, Français parlé, Larousse, 1962, Nouvelle éd. 1976.

[3] W.E. COOPER, J. PACCIA-COOPER
Syntax and Speech, Harvard University Press, Cambridge, 1980.

[4] F. DELL
L'accentuation dans les phrases en français in Forme sonore du langage, Hermann, 1984, pp 65-122.

[5] G. CAELEN-HAUMONT
Propositions pour un modèle sémantique simplifié de la complexité des signifiés, Actes des 15èmes JEP, GALF-CNRS, Aix-en-Provence, 1986, pp. 201-205.

[6] G. CAELEN-HAUMONT
Grammatical components and macro-prosody : quantitative analysis toward statistical correlations, Proceedings of Montreal Symposium on Speech Recognition, Montreal, Canada, 1986, pp 82-84.

REMERCIEMENTS
Je remercie G. PERENNOU pour ses conseils ayant porté sur le choix de la représentation de la grammaire morphosyntaxique.

GENERATION AUTOMATIQUE DE SCHEMAS MACROPROSODIQUES EN ITALIEN, A PARTIR D'UN TEXTE ECRIT

Michel CONTINI & Olga PROFILI

Institut de Phonétique Grenoble
Institut de la Communication Parlée
Université de Grenoble III
BP 25 X - 38040 Grenoble Cédex

St-Hugh's College
Oxford OX2 6LE
Grande-Bretagne

ABSTRACT

This paper describes a system of syntactic analysis which will contribute to the automatic prediction of Italian sentence prosody. It implies that orthophonetic transcription rules have already been set and that automatic word stress rules have been elaborated. The system is mainly based on the phenomenon of final vowel redundancy which characterizes the Italian language. Simple rules are applied in order to predict the boundaries of each constituant where a specific melodic contour will be located automatically. By inspecting the local context (left and right), these rules reduce the ambiguity created by homographs which are very common in Italian. A restricted list of words has also been stored in order to disambiguate a large number of sentences.

Notre étude se situe dans le cadre plus vaste de la réalisation d'un système d'analyse syntaxique permettant de générer automatiquement le schéma prosodique de la phrase italienne. Elle se situe en aval de la transcription orthographe-phonétique (TOPh) et de l'élaboration de règles morphologiques aboutissant à la prévisibilité automatique de la place de l'accent, déjà réalisées dans le cadre de la synthèse par diphones de l'italien entreprise à l'Institut de Phonétique de Grenoble (projet SYNTALIT).

Dans un premier temps la méthode que nous proposons apparaît comme peu "coûteuse", ne nécessitant pas un stockage en mémoire d'un grand nombre de données.

Outre les règles morphologiques de l'accent, elle exploite surtout la redondance du vocalisme final, très grande en italien, dans les accords du genre et du nombre. Elle utilise néanmoins le stockage en mémoire d'un certain nombre de mots-outils indicateurs soit d'un syntagme nominal (SN), soit d'un syntagme verbal (SV), soit d'un syntagme prépositionnel (SP) et en outre une liste d'exceptions lexicales.

Une telle analyse peut permettre facilement de définir les constituants syntaxiques majeurs de la phrase, de localiser la frontière (F_1) entre le SN et le SV, et par là, de prévoir les contours prosodiques à générer à cet endroit compte tenu des structures accentuelles possibles. Ces contours ont été préalablement décrits à partir d'un corpus de plusieurs centaines de phrases. Une application à la synthèse par diphones réalisée dans le cadre du projet SYNTALIT permettrait d'éliminer ainsi le recours aux marqueurs manuels relatifs aux frontières majeures et mineures.

La méthode que nous proposons peut s'adapter très bien à l'analyse de phrases simples du type :

(il bambino) ((mangia (una mela)),

pouvant également atteindre un degré plus grand de complexité par expansion du SN ou du SV avec des SP du type :

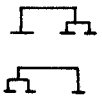
[((il bambino) (di mio fratello)) ((mangia (una mela)) (nel giardino))]

ou encore à des phrases plus complexes comportant par exemple des subordonnées du type :

se il bambino mangia una mela gliela sbuccio.

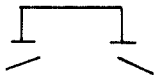
Pour la génération de macrosegments prosodiques nous avons tenu compte des phrases comportant au minimum trois mots prosodiques (MARTIN, 1978; PROFILI, 1987), permettant d'envisager au moins deux découpages différents :

a. il cavallo mangia l'erba
b. il cavallo bianco pascola



Pour ce qui concerne les phrases à deux mots prosodiques, telles que :

il cavallo pascola
ou
dorme tranquillamente
ou
ha acquistato la casa,



nous considérons que le schéma prosodique a toujours une allure circonflexe (montant-descendant).

Les règles qui permettent de rendre compte du découpage de la phrase en constituants syntaxiques majeurs et mineurs sont de type syntagmatique. La grammaire utilisée est décrite ci-dessous : partie gauche → partie droite.

Description de la partie gauche :

\# indique le début et la fin de chaque phrase,
|| délimitent les unités correspondant à un mot graphique. Ces unités seront caractérisées soit par leur valeur morphologique (mots-outils : |D(m)|, c'est-à-dire Déterminant masculin singulier), soit par leurs terminaisons vocaliques (|-a| : bambina),
\+ indique la concaténation des unités délimitées par les lignes verticales | |.

Nous proposons de stocker comme cela a été dit précédemment des listes de mots-outils dans la mesure où ils fonctionnent comme indicateurs du SN, du SV, du SP ou d'une proposition subordonnée (PS). La référence à ces listes sera exprimée par les abréviations suivantes :

1. DETERMINANTS

1.1. Déterminants du SN

|Df| = déterminant féminin singulier : mia, questa, una, ciascuna, etc.
|Dfp| = déterminant féminin pluriel : mie, quelle, tre, alcune, etc.
|Dm| = déterminant masculin singulier : il, quel, un, uno, questo, ogni, etc.
|Dmp| = déterminant masculin pluriel : i, quei, certi, alcuni, cinque, questi, etc.

Naturellement dans l'analyse de la phrase on peut envisager la succession de deux déterminants. Exemple : le loro macchine, questa tua ossessione, etc.

1.2. Déterminants du SV

|Dv| = pronoms personnels io, tu, noi, mi, ti, ci, vi, ne, li, etc.

De même que pour le SN, nous pouvons envisager la succession de plusieurs déterminants du verbe. Exemple : me lo dice, tu glielo dai, etc.

1.3. Déterminants ambigus

Dans cette catégorie |A| nous classons des éléments qui peuvent introduire aussi bien des SN que des SV, puisqu'il s'agit d'articles et de pronoms parfaitement homophones :

|Af| = déterminants ambigus féminins singulier : la, l'
|Afp| = déterminants féminin pluriel : le
|Am| = " masculins singulier : lo, l'
|Amp| = " masculin pluriel : gli ou pronom singulien pluriel.

2. PREPOSITIONS

Ces éléments introduisent les SP

|P| = di, a, da, in, con, su, per, fra, tra, del, dello, al, ai, sugli, coi, etc.

3. CONJONCTIONS

3.1. |Cc| = conjonctions de coordination : e, ma, o, né, etc.

3.2. |Cs| = conjonctions de subordination qui introduisent une subordonnée : se, che, cui, quando, finchè, benchè, prima, dopo, siccome, perchè, affinchè, etc.

4. ELEMENTS LEXICAUX

|L1| = substantifs et adjectifs masculins en |-a| : dentista, poeta, idiota, belga, etc.
|L2| = substantifs féminins en |-o| : moto, foto, etc.

Un lexique de ces exceptions nous permet de désambiguiser un certain nombre de phrases.

ex_1 : il ragazzo spara lo struzzo
il ragazzo belga lo strozzo

ex_2 : la fanciulla compra la moto
la fanciulla snella la guardo

|L3| = la flexion verbale régulière

|L4| = formes adverbiales qui peuvent fonctionner comme des syntagmes autonomes : forse, domani, oggi, ancora, vicino, lontano, adverbes en -mente, etc.

5. FORMES DES VERBES ETRE ET AVOIR

|ES| , |AV| = sono, fui, ebbi, ho, hai, hanno...

En outre les noms propres qui n'admettent pas de déterminant, sauf dans des cas exceptionnels (la Callas, le Puglie) seront reconnus par une marque introduite dans la TOPh.

Sur la base de cette grammaire, une phrase simple de trois mots prosodiques, telle que :

il bambino mangia una mela
1 2 3 4 5

peut être analysée comme suit :

|Dm| + |-o| + |-a| + |Df| + |-a|
1 2 3 4 5

Règle d'accord

Dans une lecture gauche → droite, on rencontre successivement :

Début de phrase
(1) Un déterminant du nom masculin singulier il
(2) Une unité se terminant par -o, obligatoirement en accord (au masculin donc) avec il.
(3) Une unité se terminant par -a. Cette unité ne peut pas s'accorder avec la précédente car -a est la terminaison du féminin ou d'un verbe à la troisième personne du singulier. Elle sera donc reconnue comme verbe du SV qui commence donc avec cette unité. Par conséquent la frontière entre SN et SV va être identifiée à la fin de l'unité précédente qui termine le SN. Une ambigüité pourrait apparaître en supposant une forme masculine en -a qui s'accorderait avec le substantif il bambino.

Ex. * il bambino idiota una mela

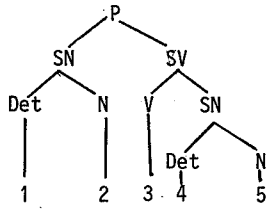
Mais cette phrase est inacceptable car elle n'a pas de SV.

(4) Un déterminant du nom féminin singulier : una, qui confirme que l'unité précédente est bien la forme verbale.
(5) La dernière unité en -a s'accorde avec le Déterminant qui la précède.
Fin de phrase.

On en arrive au texte complet de la règle.

R : # |D(m)| + |-o| + |-a| + |D(f)| + |-a| #
1 2 3 4 5

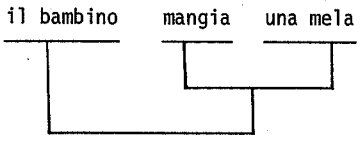
donne



d'où D > SN > F1 > SV > FØ

où D = début de phrase
F1 = frontière majeure
FØ = frontière finale de phrase

A cette structure syntaxique correspond une structure prosodique congruente :



Il ne restera plus qu'à affecter à chaque mot prosodique un contour type, tout en tenant compte de sa structure accentuelle.

Les exemples suivants permettront de mieux illustrer notre démarche.

a. la bambina mangia una mela
1 2 3 4 5

Cette phrase aura la règle :

|Af| + |-a| + |-a| + |Df| + |-a|
1 2 3 4 5

(3) sera interprété comme V et non pas rattaché à (2) malgré sa terminaison. L'unité (2) sera interprétée comme fin de syntagme nominal et sera suivi d'une frontière majeure F1.

b1. la bambina buona mangia una mela
1 2 3 4 5 6

aura la règle :

R : # |Af| + |-a| + |-a| + |-a| + |Df| + |-a| #

Cette règle serait la même pour la phrase suivante :

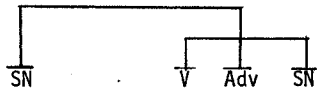
b2. la bambina mangia ancora una mela
1 2 3 4 5 6

Or le découpage syntaxique et la structure prosodique des deux phrases divergent. Pour la première F1 se situerait après (3) et pour la seconde après (2). Nous sommes ainsi amenés à affiner la règle précédente avec une caractérisation de l'unité 4.

R1 : # |Af| + |-a| + |-a| + | -a sauf L4 | + |Df| + |-a| #
1 2 3 4 5 6

R2 : # |Af| + |-a| + |-a| + |L4| + |Df| + |-a| #
1 2 3 4 5 6

Cette dernière aura la structure prosodique :



Ajoutons que la phrase b1 négative ne serait pas ambigüe car :

b3. la bambina buona non mangia una mela

ne pourrait être confondue avec

b4.* la bambina mangia non ancora una mela,

qui serait inacceptable.

c1. I ragazzi italiani studiano la lingua francese
1 2 3 4 5 6 7

aura la règle :

|Dmp| + |-i| + |-i| + |-o| + |Af| + |-a| + |-e|
1 2 3 4 5 6 7

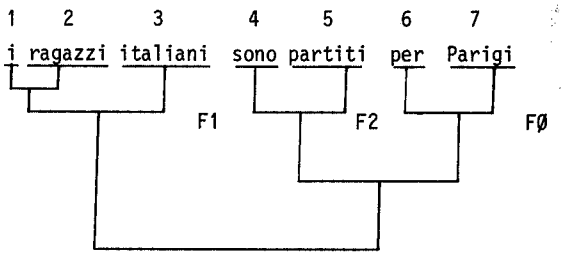
Le SN sera identifié facilement par l'accord de (2) et (3) avec (1) (masculin pluriel). Le début du SV sera identifié par la terminaison |-o| de (4) qui ne peut pas s'accorder avec les précédentes et l'unité |Af| sera reconnue comme un déterminant du nom (article) introduisant le SN2. F1 se situera donc après (3).

c2. I ragazzi italiani sono partiti per Parigi
1 2 3 4 5 6 7

aura la règle

|Dmp| + |-i| + |-i| + |ES| + |-i| + |P| + |-i|
1 2 3 4 5 6 7

tout en se terminant par -i, (5) ne peut pas s'accorder avec (1), (2) et (3) (masculin pluriel) dont il est séparé par une forme du verbe être (ES). La phrase c2 correspond à la structure prosodique suivante :



avec une frontière majeure F1 après (3) et une frontière mineure F2 après (5) entre la forme verbale et le syntagme prépositionnel.

Subsistent malgré tout des cas d'ambiguité difficiles sinon impossibles à résoudre sans un lexique complet de toutes les formes lexicales de l'italien.

Une phrase comme :

d1. la gatta mangia la carne
1 2 3 4 5

sera interprétée en effet de la même façon que :

d2. la gatta magra la vede
1 2 3 4 5

à savoir :

|Af| + |-a| + |-a| + |Af| + |-e|

Or les deux phrases ont une structure prosodique différente, avec un F1 après (2) pour d1 et après (3) pour d2. Cela est dû à l'homophonie entre l'article défini la et le pronom personnel féminin la. A titre d'exemple nous présentons les schémas prosodiques (variations de Fo) des deux structures prosodiques différentes correspondant au découpage syntaxique des phrases d1 et d2.

Ajoutons que l'ambiguité ne persiste plus si la phrase est à la forme négative (voir b3 et b1) ou si le SV comporte certains types d'expansion.

d3. la gatta mangia la carne dei suoi gattini.

Une phrase comme d2 ne pourrait pas admettre le même type d'expansion.

La méthode proposée semble avoir un bon coefficient de prévisibilité dans le cas des phrases énonciatives simples en recourant seulement à des lexiques restreints et spécifiques. Une analyse plus approfondie devrait permettre d'étendre la prévisibilité de macrosegments prosodiques dans des phrases plus complexes. Certes des ambiguités persisteront toujours et elles ne peuvent pas être enlevées sans le recours à un analyseur syntaxique complexe.

Dans un premier temps nous pensons que les cas d'ambiguité pourraient être affectés d'une intonation neutre qui, par ailleurs, connaîtra un mouvement mélodique certain par le simple jeu des variations mélodiques lié à la réalisation et à la grande mobilité de l'accent de mot caractéristiques de l'italien.

BIBLIOGRAPHIE

[1] CHOMSKY N. (1957), Syntactic Structures, Mouton & Co., La Haye.

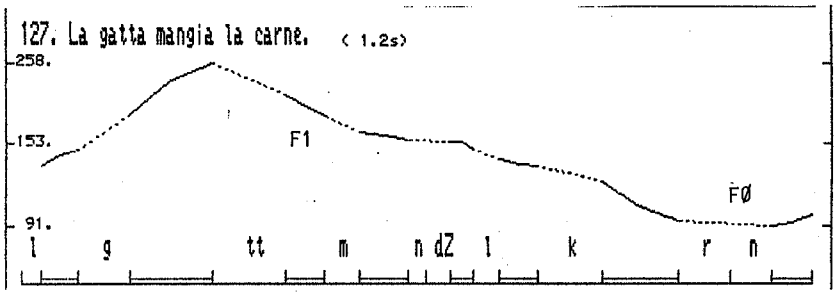
[2] DELMONTE R. (1986), A Computational Model for a Text-to-Speech Translator in Italian, Revue Informatique et Statistique dans les Sciences Humaines. XXII. 1-4. p. 23-65.

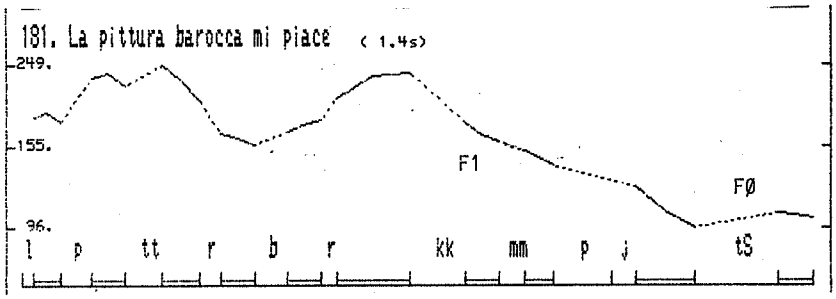
[3] MARTIN Ph. (1978), L'intonation de la phrase en italien. Studi di Grammatica Italiana, VIII, p. 395-417.

[4] MARTIN Ph., PROFILI O. (1987), Accent de mot et structure syntaxique en italien. Information, Communication. Presses de l'Université de Toronto.

[5] PROFILI O. (1987), Acoustic Investigation of Intonation in two Regional Varieties of Italian : Preliminary Results. Progress Reports from Oxford Phonetics, II, p. 47-64.

[6] TEKAVČIĆ P. (1972), Grammatica storica dell'italiano. Vol. II : Morfosintassi. Il Mulino, Bologna.





L'EVOLUTION DE LA VOIX CHEZ L'ENFANT ENTENDANT ET CHEZ LE DEFICIENT AUDITIF

KONOPCZYNSKI Gabrielle, VINTER Shirley

Laboratoire de Phonétique, Université de Besançon, 25030 Besançon-Cedex

The evolution of the voice of babies between 8 and 24 months, with normal and abnormal hearing was studied. It is demonstrated that for both groups the usual Fo is quite stable and lower than generally supposed. The hearing child has a quite different attitude when playing alone or interacting with an adult; in the first case, he practices exploratory voice activities, utilizing a very large range going from low creaking to extremely high squealing; on the other hand, when interacting, he places his voice in the medium zone. On the contrary, the deaf child does not babble when alone and has no exploratory activities. He only produces sounds when he is interacting, and the range is reduced. It can be concluded that there is not just an evolution in the voice, but there is also a building up of the voice with the building up of language and socialization

Si les études sur la voix de l'adulte connaissent un vif intérêt, les travaux sur la voix de l'enfant, et notamment sur son évolution pendant les premières années de sa vie, sont peu nombreux. Ceux concernant l'enfant déficient auditif sont pratiquement inexistants. Seules deux tranches d'âge ont retenu l'attention des chercheurs : la période néo-natale où les paramètres du cri ont été étudiés par des équipes médicales en raison des indications qu'ils apportent pour le dépistage de certaines pathologies (bilan in 1 et 2) et, à l'autre extrémité, la période de puberté qui a également donné lieu à un grand nombre de travaux. Quelques recherches portent sur la période pré-scolaire ou scolaire (bilan in 2). En revanche, la période de babillage (6-24 mois) a été quasiment ignorée, alors que c'est le moment où l'enfant va devoir plier sa voix et ses larges possibilités, dont jusque là les seules contraintes étaient physiologiques, à des contraintes de type linguistique ou social. Certes la littérature pédolinguistique apporte quelques vagues renseignements sur la hauteur de la voix enfantine à cette période, mais les chiffres donnés affichent des divergences appréciables. Pour certains (3,4,) vers 6 mois, le Fo est élevé, supérieur à 400 Hz. Pour d'autres (5) il y a chute progressive de la voix depuis la naissance, et à 6 mois le Fo se situerait aux environs de 350Hz. Pour d'autres encore(6) il y a stabilité du Fo jusque vers 12 mois au moins.(cf. bilan in 2). Mais la plupart des travaux pose de réels problèmes méthodologiques (2), notamment en raison du fait que cris et vocalisations sont souvent englobés dans un même ensemble et que seul le concept de Fo moyen est retenu. Or, il nous est apparu qu'**il y a lieu de différencier**, dès le plus jeune âge, la notion de **Fo usuel** (Fo-u) de celle de **Fo moyen** (Fo-m). Le Fo-u est la **hauteur à laquelle la voix se place naturellement, sans qu' aucune intention particulière ne soit réalisée.** Il représente la dynamique de base d'un locuteur. **Au contraire le Fo-m représente, comme son nom l'indique, la moyenne des fréquences de la voix enfantine, compte- tenu de l'ensemble de ses performances vocales.**

La présente étude est consacrée aux caractéristiques de la voix enfantine lors de la longue période de babillage, c'est-à-dire entre 8 et 24 mois environ. Pour les enfants entendants elle porte sur un sujet de base, de sexe féminin, et sur une demi-douzaine de sujets de contrôle, des deux sexes. Le sujet de base est suivi semaine par semaine entre la fin du mois 8; et celle du mois 10; (plus de 2.000 énoncés traités), puis de façon plus lâche, bi-mensuelle ou mensuelle. Les autres sujets sont suivis de façon assez irrégulière, mais au moins mensuelle. Les sujets mal entendants, au nombre de 5, 1F, 4M. Nous avons écarté de l'analyse cris, pleurs et divers signes d'inconfort pour ne retenir que les autres émissions vocales. Dans ce vaste ensemble, nous avons pu montrer (2), qu'il y a lieu de faire, **pour les sujets entendants, une distinction entre les émissions de jeu vocal apparaissant en situation de production solitaire, appelées JASIS, celles produites en interaction avec un adulte, que nous appelons PROTO- ou PSEUDO-LANGAGE** (abrégé en PL) et celles produites dans des situations variables, par exemple lorsque l'enfant joue avec un substitut d'animé, appelée "catégorie intermédiaire", non étudiée ici. Chacune de ces catégories possède une structuration syllabique, temporelle et mélodique (2) qui lui sont propres. Pour les sujets sourds, il n' y a pratiquement pas de jasis solitaire, seules existent les émissions produites en interaction, mais elles n'ont aucune fonction de type proto-langage. Notre **objectif** ici est de montrer que **l'enfant se sert également de registres vocaux différenciés** dans diverses situations, et de suivre l'évolution de sa voix au moment où il commence à construire son langage et son image sociale. Nous nous limiterons à l'étude de la voix prise dans le sens restrictif de HOLLIEN (1974-1982), référant à la seule fonction glottique, sans prendre en compte les aspects de fonction de transfert du tractus vocal et tous les aspects articulatoires qui, selon les travaux de LAVER (7), peuvent caractériser une voix .

1. LE FONDAMENTAL USUEL(Fo- u)(entendant et sourd).

Calculé pour les adultes sur la base des syllabes inaccentuées (G. FAURE) ou sur la moyenne des énonciatives (technique la plus usitée) ou encore sur le " euh d' hésitation " (P. LEON) , ce Fo-u se repère très aisément chez l'enfant, qu'il soit entendant ou non. Il émet, en effet, un nombre important de productions vocales quasi inconscientes et neutres, qui ont la forme de vocoïdes mi-ouverts, brefs (pour les sourds, M= 10cs, pour les entendants M=22cs, Δ= 6,9 cs, limites : 6-36 cs.), monotones, d'intensité faible Ils sont totalement différents des vocalisations fluctuantes du jeu vocal ou des émissions de **PL**. Malgré leur fréquence d'occurrence élevée, ces [ə] restent totalement ignorés, et par l'entourage, et par les chercheurs, en raison de leur vacuité. Ils présentent pourtant une caractéristique pour le moins étonnante à ce stade d'évolution langagière : en effet, alors que, à tous les niveaux, l'instabilité des émissions règne en maître, ce Fo-u, tel que nous l'avons défini, affiche une remarquable stabilité, comme le montrent les tableaux 1(sujet entendant) et 2(sujet sourd) ainsi que la fig. 1(histogramme type).

AGE (mois; sem.)	NOMBRE D'ENONCES	NOMBRE D'ECHANTIL-LONS(1)	Fo-u EN HERTZ	ECART-TYPE (Δ)EN HZ	ZONE TONALE LA PLUS EMPLOYEE
8;	100	2980	324	78	220-450 : 82 %
9;1	50	1200	340	74	270-430 : 85 %
9;2	50	565	340	76	260-460 : 85 %
9;3 (2)	14	./.	335	76	270-380 : ??
9;4	50	653	335	76	260-400 : 75 %
10;1+2 (3)	20	221	324	38	270-380 : 95 %
10;3+4 (4)	15	205	334	40	240-380 : 80 %

LEGENDE : (1) Traitement réalisé sur P.M. : 20 échant./sec.
(2) La plupart des [ə] de cette semaine sont de type "*creak*" et présentent un mauvais rapport signal/bruit. L'analyse a été effectuée manuellement. Même remarque pour les notes 3,4.

TABLEAU 1 : FONDAMENTAL USUEL DURANT LA PERIODE CHARNIERE
(sujet de base, entendant)

Mois	Fo-u en Hz	Δ en Hz
7;	335	28
8;	330	27
9;	330	25

TABLEAU 2 : Fo-USUEL (sujet sourd)

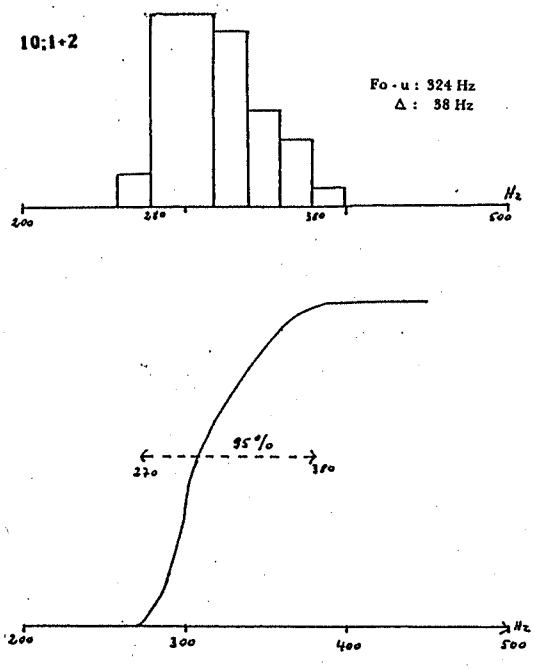


FIG.1 : Fo-U TYPE (entendant et sourd)

L'analyse de la fig.1 (histogramme et courbe cumulative fait en effet ressortir le peu de dispersion du Fo-u. Ainsi, on remarquera que la zone tonale la plus employée est celle comprise entre 300 et 400 Hz,; la courbe cumulative présente une pente raide suivie d'un plateau; ceci montre que 75 % de la tessiture se situe entre 260 et 400 Hz, les zones en deçà de 240 ou au-delà de 460 Hz n'étant pratiquement jamais utilisées pour ce type d'émissions, dont aucune ne dépasse d'ailleurs 500 Hz.

Outre la remarquable stabilité du Fo-u autour de 340 Hz, donc dans le médium inférieur, on notera que l'enfant domine de mieux en mieux sa voix, puisque les écarts-types, relativement élevés aux mois 8; et 9; se réduisent de moitié au dixième mois; la tessiture, d'abord comprise à75 % entre 260 et 400 Hz, soit sept demi-tons, avec des extrêmes allant de 200 à 480 Hz, mais ne dépassant jamais 500 Hz, se resserre de plus en plus pour être à dix mois comprise à 95 % entre 270 et 380 Hz, aucun [ə] ne dépassant plus les 420 Hz. Après 12 mois, le Fo-u n'évolue plus et se situe chez tous nos sujets entre 330 et 350Hz, avec des écarts-types encore réduits, ne dépassant guère 30Hz et même moins pour les sujets sourds.

Soulignons également que l'intensité des [ə] est faible, et surtout qu'elle ne change pas lorsque le niveau sonore environnant augmente; ceci montre que l'enfant est dans un état neutre, qu'il ne cherche ni à attirer l'attention, ni à communiquer. Ce paramètre n'évolue pas avec l'âge; l'intensité moyenne ne dépasse jamais 28 dB et l'écart-type est toujours inférieur à 6 dB .

2.ENFANT ENTENDANT:
2.1.FONDAMENTAL MOYEN DANS LES VOCALISATIONS DU JASIS.

Comme pour le Fo-u, nous étudions les vocoïdes semaine par semaine, afin de détecter une éventuelle évolution. Leur nombre étant très élévé, nous avons sélectionné au hasard 100 énoncés par semaine. Les résultats sont présentés dans le tableau 2.

AGE (mois; sem.)	NOMBRE D'ENONCES	NOMBRE D'ECHANTIL-LONS(1)	Fo-m EN HERTZ	ECART-TYPE (Δ)EN HZ
8;	100	2532	410	76
9;1	100	2778	380	88
9;2	100	1210	401	90
9;3 (2)	50	247	368	86
9;4	100	2198	402	112
10;1+2	100	1839	359	96
10;3+4	100	2718	341	58

NOTE: 1) cf.tableau1.
2) Beaucoup de vocoïdes ici présentent un mauvais rapport signal/bruit, d'où le nombre inférieur d'énoncés traités.

TABLEAU 2 : FONDAMENTAL MOYEN DES VOCOIDES DU JASIS
(sujet de base, entendant)

Comparé au Fo-u, **le Fo-m se distingue par son élévation, sa dispersion et son instabilité.** Les histogramme et courbe cumulative (Fig. 2) révèlent clairement ce comportement tout à fait différent du Fo-m : nous ne trouvons plus un bloc de fréquences massivement centré autour de la moyenne; ici, aucune gamme de fréquence n'est majoritaire, les variations de voix se répartissent inégalement sur toute l'échelle comprise entre 200 et 500 Hz (cf. pente douce des courbes); en outre, 30 à 40 % des énoncés dépassent 500 Hz. Nous avons établi toutes les courbes pour la période 9 à 11 mois, et avons pu remarquer qu'il y a souvent deux classes d'occupation pratiquement identiques, ce qui se voit aussi bien sur les histogrammes qui présentent deux pics que sur les courbes qui affichent deux segments de pente plus raide que le reste. Les deux zones sont assez proches, et le Fo-m se situe de préférence entre elles. Il semblerait que dans ces vocoïdes, l'enfant, essayant ses possibilités, hésite entre divers niveaux de hauteur.Les chiffres de DELACK (5) et DIESTELMANN (6) confirment les nôtres; ils trouvent respectivement un Fo-m se situant entre 340 et 390 Hz pour le premier, et 378 et 412Hz pour le second.

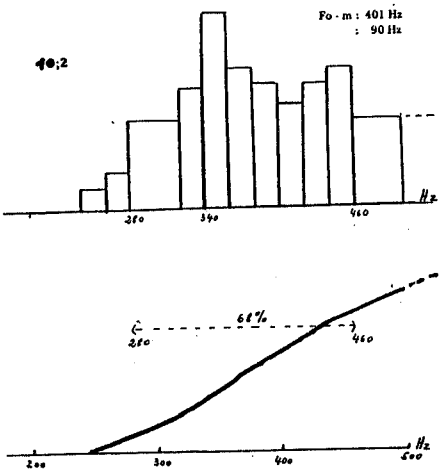


FIG.2 Fo-m JASIS

Il faut remarquer également que si le Fo-u a su atteindre sa stabilité dès le début du neuvième mois, il y a au contraire évolution dans le Fo-m. Une étude semaine par semaine (Tableau 3) montre que la fréquence moyenne baisse progressivement pour se rapprocher de la voix de base à la fin du dixième mois; la dispersion diminue également, tout en restant bien supérieure à celle notée pour le Fo-u, puisque, à la fin de la période soumise à examen, 15 % de la voix dépasse encore 500 Hz; donc la dynamique reste large .Nous sommes donc en désaccord avec DELACK et DIESTELMANN qui estiment que le Fo est stable entre six et douze mois.

Quant à l'intensité de ces vocoïdes, elle montre des caractéristiques semblables à celles de leur Fo-m, à savoir instabilité (Δ toujours supérieur à 6.5 dB). Ils sont tous nettement plus intenses que les vocoïdes ayant servi à déterminer le Fo-u (Intensité moyenne supérieure à 30db) et souvent il y a saturation de l'enregistrement. Hauteur et intensité évoluent donc de pair.

Entre un et deux ans, bien que l'enfant entre progressivement dans le langage, il continue à pratiquer les activités exploratoires dont il était coutumier; leur Fo-m garde l'essentiel de ses caractéristiques antérieures,à savoir élévation(Fo-m >400Hz) et forte dispersion(Δ ≧ 1ooHz). Tout au plus notera-t-on que les fréquences suraiguës diminuent légèrement au fur et à mesure que l'enfant avance en âge, et que l'intensité devient un peu moins instable.

Les résultats obtenus respectivement par le Fo-u et le Fo-m montrent que **dès huit mois, l'enfant, qu'il soit entendant ou sourd, possède une voix de base " stable et plus grave** que ce que la littérature pédophonétique a laissé entendre jusqu'à présent. En outre, les allégations sur l'évolution du Fo au cours des six premiers mois de vie sont à revoir en fonction de nos résultats. Il semblerait en effet que les [∂] soient présents dès la période de *cooing*. L'enfant aurait-il sa " voix de base " rapidement, les progrès développementaux consistant essentiellement en un élargissement de la tessiture? Mais il est alors surprenant que l'accroissement de longueur des cordes vocales qui se fait au cours de la première année (un accroissement de 100 % pratiquement, cf. 8,9) n'ait pas plus d'effet sur le Fo-u. De même, chez le sourd, l'apport d'informations acoustiques à l'aide d'un appareillage n'a pas d'incidence directe sur son Fo-u. Rappelons que celui-ci coexiste avec le Fo-m quand le sujet sourd bénéficie d'une aide acoustique apportée par les prothèses, mais qu'il reste le seul mode de production quand l'enfant n'est pas appareillé. Malheureusement ces productions, trop insignifiantes, ne sont pas captées par l'entourage et ne trouvent donc aucun écho. Faut-il attribuer les divers résultats concernant la voix de base au fait que l'enfant, même sourd, domine la régularité des mouvements laryngés dès trois semaines (10)? Il agirait donc essentiellement sur la tension des cordes vocales, et compenserait par là leur augmentation en longueur et en volume. Tout ce point demande encore investigation, avec d'autres techniques que celles que nous avons pu utiliser jusqu'à présent.

2.2. LA TESSITURE.

Fo-u et Fo-m, avec leurs dynamiques respectives, ne suffisent pas à décrire l'ensemble des possibilités vocales de l'enfant, c'est-à-dire sa tessiture, souvent appelée, à tort, son registre En effet, un certain nombre d'énoncés se situent en deçà de l'utilisation dite normale de la voix. Il s'agit essentiellement de deux sortes de productions vocales, le "*creak*" ou *growling*" ou " craquement vocal", très grave, caractérisé par une instabilité extrême du mouvement glottique et le "couinement sur-suraigu" ou "*squealing*", pouvant atteindre 1800 Hz.. Il est très peu intense, et de durée généralement très brève, comparé au creak qui est au contraire long et puissant. Les émissions produites avec ce type de voix sont en règle générale des vocoïdes isolés; ils apparaissent en nombre élevé, mais uniquement dans le Jasis.

Notons que ces deux modes phonatoires qui caractérisent la tessiture de l'enfant dans son ensemble, sont en même temps deux registres de voix qui, si l'on suit HOLLIEN nécessitent des comportements laryngés complexes, notamment sur le plan du controle de la hauteur. Ceci dénote chez l'enfant des possibilités vocales au moins aussi variées que chez l'adulte.

Enfin, il convient de signaler qu'au contraire du Fo-m qui n'évolue guère ultérieurement, ces deux comportements laryngés disparaissent peu à peu; la voix se normalise donc progressivement, après 18 mois, l'usage des extrêmes de la tessiture étant peu à peu abandonné.

2.3. LA VOIX DANS LES INTERACTIONS.

L'analyse a permis de dégager une nette différence dans cette situation : en effet, l'enfant n'utilise plus les extrêmes de la tessiture, ni les divers comportements exploratoires. Creaks, énoncés sur-suraigus, brusques variations de hauteur disparaissent. **Seule la voix modale est employée**. Un Fo donné est lié à une modalité linguistique donnée et les extrêmes de la tessiture se situent entre 260 et 760 Hz, selon le type d'énoncé; 85% des émissions sont dans une zone fréquentielle inférieure à 500 Hz, et la majorité des énoncés se situe dans le médium. En outre, dans le **PL** l'enfant utilise un mode vocal absent du **JASIS** : il s'agit du **chuchotement**. A huit mois et au début du neuvième mois, les bébés ne savent pas répondre à la voix chuchotée de l'adulte sur le même mode. Généralement ils se taisent ou gazouillent faiblement. Le chuchotement apparaît vers 9;3 et se développe rapidement. . Les énoncés chuchotés sont tous brefs .

Nous pouvons conclure de l'ensemble de ces résultats que **le jeune enfant sait exploiter sa tessiture avec une dynamique la plus large possible; il pratique ainsi une série d'intenses activités exploratoires, où il utilise ses capacités phonatoires jusque dans leurs extrêmes, comme pour en tester les limites. Mais par ailleurs, il a déjà stabilisé certains paramètres et trouvé sa voix naturelle. Enfin, la restriction de la dynamique globale de la voix et l'acquisition du chuchotement dans les interactions témoignent de la socialisation très rapide de l'enfant**. Celui-ci a appris à "ranger" sa dynamique vocale et à s'adapter à la situation d'émission. En même temps qu'il produit les premiers énoncés interprétables linguistiquement bien que la couche verbale en soit encore absente, il commence donc à construire sa voix.

3. ENFANT DEFICIENT AUDITIF

Par rapport à l'enfant entendant, de notables différences existent chez l'enfant sourd. En effet celui-ci ne babille pratiquement pas en situation de jeu solitaire. Chez le sourd profond les activités de type exploratoire sont absentes; en revanche, l' enfant appareillé les possède, mais leur fréquence ne dépasse guère 9ooHz. De façon générale, **seules**

existent les émissions produites en interaction(11). Leur Fo-m est se situe aux environs de 350 à 450Hz, que l'enfant soit appareillé ou non. En revanche, l'appareillage a une incidence sur la dynamique vocale générale. Les écarts-types avant appareillage sont généralement inférieurs à 30Hz, ils augmentent nettement après appareillage, sans cependant atteindre la grande dynamique observée chez les entendants. La tessiture , réduite quand l'enfant ne dispose pas d'informations acoustiques, augmente, principalement dans les aigus (Fig3). En revanche, les couinements suraigus, ainsi que les fréquences graves et les creaks restent absents. A fortiori, même en cas d'appareillage précoce, le chuchotement, impossible à percevoir, n'apparaît que tardivement.

4. CONSTRUCTION DE SA VOIX PAR L'ENFANT ET SOCIALISATION

De l'étude des caractéristiques vocales de l'enfant entre 9 et 24 mois, et de leur utilisation selon les situations, un point ressort clairement : **la voix évolue, mais beaucoup moins qu'on ne serait en droit de s'y attendre, étant donné la maturation physiologique du larynx** Notre démonstration a mis en relief l'existence d'un Fo-u, très stable, et celle d'un Fo-m, totalement différent. L'enfant possède très tôt sa voix de base naturelle, due très certainement à des caractéristiques innées et à un remarquable contrôle du travail de ses cordes vocales.
Mais à ces caractéristiques individuelles s'ajoutent des caractères acquis, tels que l'emploi de certaines modalités vocales pour certaines situations ou certains types de structures linguistiques (2). Ces caractères se structurent progressivement. On peut donc dire que **l'enfant construit sa voix, en même temps qu'il construit son langage.** Pour le sujet sourd, le même processus de retrouve, décalé dans le temps. Rappelons en effet qu'à partir de 12/13 mois, l'enfant entendant entre dans la période du langage articulé et qu'il émet les premières holophrases. Au niveau vocal, il commence à éliminer les registres trop marqués, d'un usage inadéquat dans la communication sociale. Il restreint sa dynamique quand il est en interrelation avec un adulte. Certes, contrairement à ce qui se passe dans l'acquisition des éléments segmentaux ou prosodiques, la structuration de la voix ne se fait pas vers une "voix-cible " , car il n'y a pas en ce domaine de cible précise, à valeur fonctionnelle, la voix de chaque individu étant différente pour des raisons anatomiques et physiologiques. Mais dans les interactions, tout se passe comme si la "modalité" de la voix faisait office de cible. Donc, si l'individualité de la voix est déterminée dès la naissance (rappelons que les mères ou le personnel hospitalier reconnaissent souvent les nouveaux-nés à leurs cris), sa structuration se fait progressivement, en parallèle à celle du langage . Il n'est alors pas étonnant que chez les enfants sourds , retard de langage et non construction de la voix aillent de pair.

Ces considérations sur la construction de la voix et sur les problèmes qu'ont les enfants sourds permettent de mener plus loin nos réflexions au niveau théorique. Il apparaît en effet que l'enfant développe peu à peu son système en concordance avec celui de l'adulte, car au cours des interactions sociales, il est amené à revoir et à ré-organiser son système initial. Ayant déjà eu l'opportunité de rencontrer diverses situations sociales et d'être confronté à diverses tâches, il organise son système en correspondance avec celui de son partenaire, donc il se socialise peu à peu et se "culturalise" par et durant ce processus. Et si l'enfant sourd, surtout quand il n'est pas appareillé et pris en charge précocement, présente souvent des problèmes de socialisation, n'est-ce pas dû partiellement à l'impossibilité de s'adapter ainsi, donc de construire sa voix, base du langage, lui-même facteur essentiel de la communication ?

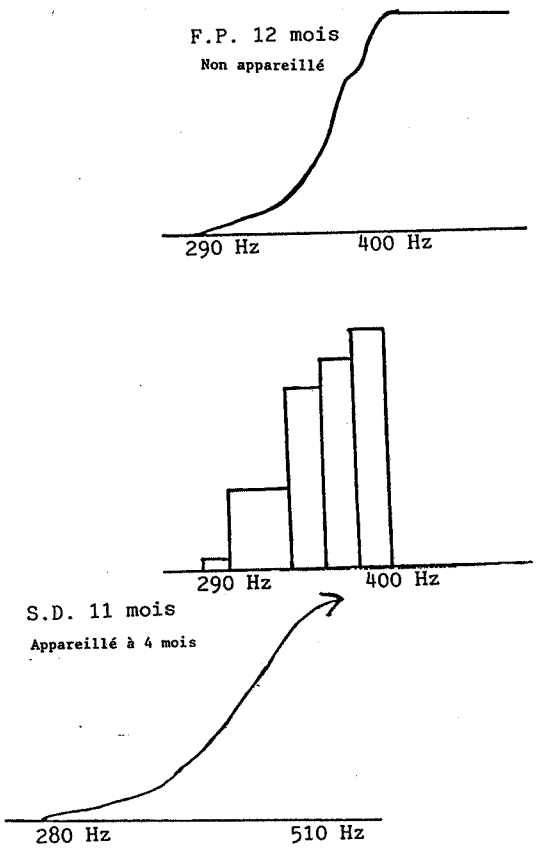


FIG.3:Fo-m 2 sujets sourds type de surdité profonde III)

1)KASKINEN H.,MICHELSSON K.(1982)
The history and development of cry analysis.
Helsinki: Voces Amicorum.. in Honorem Sovijärvi.153-167.
2)KONOPCZYNSKI G.(1986)
Du Prélangage au Langage: Acquisition de la Structuration Prosodique. Thèse Etat, Univ.StrasbourgII
3)BRUGIROUX M.C.,CASSOU A.,QUINCEZ C.(1980)
Etude de la mélodie de la voix de l'enfant sourd.
Bulletin d'Audiophonologie 11/1,89-95.
4)RODRIGUEZ D.(1985)
Contribution à l'étude ontogénétique des séquences de comportement et des vocalisations chez le jeune enfant. Univ. de Besançon,Diplôme de Docteur en Sc. Vie .
5)DELACK J. (1978)
Aspects of infant development in the first year of life.
in BLOOM: Readings in Language Development 94-114.
6) DIESTELMANN M. (1982)
Aspekte der verbalen Kommunikation in der präverbalen Phase. Univ. Munich: Doctoral Dissertation.
7)LAVER J. (1975)
Individual Features in Voice Qualities.
Edimbourgh: Ph.D.
8) CHARACHON D. (1971)
Connaissances actuelles sur la physiologie de la phonation. Journal Français d'O.R.L. XX/2,403-408.
9)CHARACHON R. (1971)
Croissance du larynx. Journal Fr.d'O.R.L. XX/2, 397-400.
10) FOURCIN A.(1978)
Acoustic patterns and speech acquisition.
in Waterson & Snow : The Develop. of Communication.47-72.
11) VINTER S. (1985)
Voix et mélodie de l'enfant sourd.
Bulletin d'Audiophonologie,N.S.I/1-2,219-236.

STRUCTURES INTONATIVES ET SYNTAXIQUES EN GREC

Theodoros MALAVAKIS

Institut de la Communication Parlée
Institut de Phonétique de Grenoble
L.A. CNRS 368
Université de Grenoble III
38040 GRENOBLE CEDEX

ABSTRACT

The aim of our analysis is to examine relationships existing between syntax and intonation in Greek. The links between syntactic units and the melodic structure of an utterance are studied, using a corpus containing affirmative and interrogative modality phrases in different structures. The place of stress in the words is taken into consideration also.

Results confirm our initial hypothesis : the function of each syntactic element and the structure to whom it belongs, plays a predominant role on its contour. As a matter of fact, a nominal syntagm as a subject, has a particular contour which can be predicted from its acoustic structure and the type of syntactic structure.

INTRODUCTION

A partir d'un corpus constitué d'énoncés de modalité affirmative et interrogative, nous avons essayé d'extraire des indices qui nous permettent une classification des contoues prosodiques relativement à la fonction syntaxique du mot et à sa structure acoustique. Des mots oxytons, paroxytons et proparoxytons ayant des différentes structures acoustiques et fonctions syntaxiques ont été insérés dans des phrases appartenant à des structures différentes.

Il n'a pas été étudié l'influence des paramètres extra-linguistiques tels que contexte, idiosyncrasie etc.

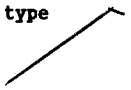
PROCEDURE D'ANALYSE

Six modèles de structures [1] ont été établis; tous les énoncés du corpus - qu'ils soient affirmatifs ou interrogatifs - se classent par rapport à ces modèles (voir fig. 1). Les constituants internes des énoncés appartiennent à toutes les catégories grammaticales et assument diverses fonctions syntaxiques (Syntagmes Nominaux-sujets, Syntagmes Nominaux-compléments d'objet, Syntagmes Verbaux etc). D'autre part, des propositions subordonnées ont été intégrées dans des principales, afin d'étudier le comportement des constituants internes dans celles-là.

RESULTATS

Les Syntagmes Nominaux à Fonction Sujet.

Placés dans des positions pré- ou post - verbales (mais jamais en finale absolue*) ils ont constamment le même contour, à une exception près. Il est étonnant de constater qu'indépen - damment de la modalité de l'énoncé les SN-sujets gardent le même contour. Dans les mots proparoxytons l'influence de l'accent sur le F0 est minime.

Ce contour est du type 

Les Syntagmes Nominaux à Fonction Complément d'Objet (Directs ou Indirects).

L'uniformité de ces contours ne surprend pas, dans ce sens qu'ils se trouvent tous en position intérieure dans l'énoncé. Cependant, leur contour plat ———— propre aux SN-compléments d'objet directs ne se rencontre pas dès lors qu'il s'agit d'un mot paroxyton; comme nous l'avons signalé dans un article précédent [4], les paroxytons ont un comportement spécifique. En effet, leur contour est toujours descendant et ceci indépendamment de la modalité de l'énoncé.

Le Rôle des Structures

Nous n'allons pas donner des résultats pour toutes les structures du corpus, mais analyser le comportement de deux éléments à fonction syntaxique identique, à l'intérieur de deux structures. Il s'agit de deux SN-sujets en position post-verbale, proparoxytons et comportant cinq syllabes (voir fig. 2).

Examinés de près, les contours sont certes de la même famille scalaire (du type ⟋⟍) mais la pente du contour du SN de la structure B est moins prononcée par rapport à celle du contour de la structure A. C'est une constatation régulière, valable pour les six structures; le niveau de dépendance syntaxique du mot influe directement sur le gradient de la pente. Plus le mot se trouve impliqué dans des rapports de dépendance à niveau supérieur, plus sa pente devient abrupte.

REMARQUES DE CONCLUSION

S'il est certain que les contours des SN - à fonction Sujet ou Complément d'Objet - sont stables en grec, il n'en reste pas moins certain que le degré de "profondeur" du mot dans la structure conditionne le gradient de la pente; on se trouve pour ainsi dire, en présence d'un "besoin d'affirmation" de la part du contour; plus l'unité pouvant recevoir le contour sera enchâssée, plus la pente de ce dernier sera raide. Il s'avère donc nécessaire de procéder à une sous-catégorisation des contours, selon la structure à laquelle ils appartiennent.

---

* Il est extrêmement rare en grec de trouver le sujet d'un énoncé rejeté à la fin.

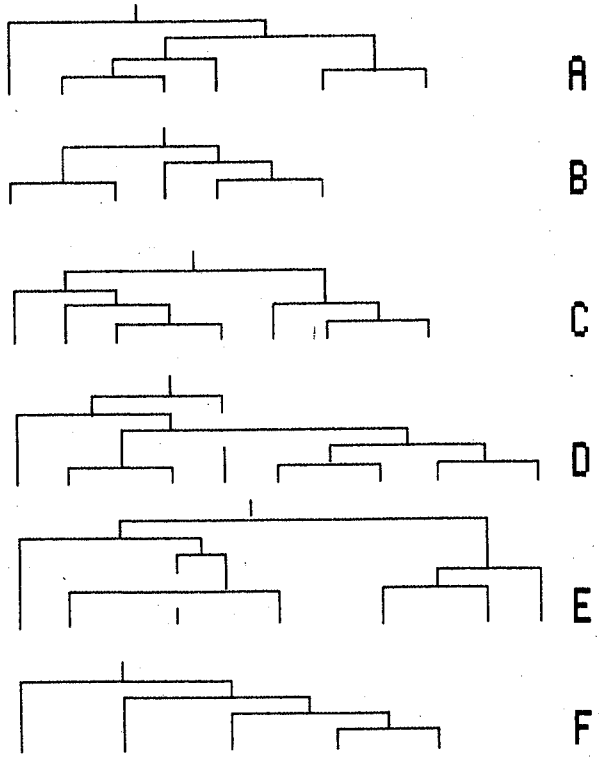
REFERENCES BIBLOGRAPHIQUES

[1] MARTIN Ph. (1977) Résumé d'une Théorie de l'Intonation. Bulletin de l'Institut de Phonétique de Grenoble, Vol. VI, 57-58.

[2] MARTIN Ph. (1978) L'Intonation de la Phrase en Italien. In Studi di Grammatica Italiana, Accademia della Crusca, Firenze, 395-417.

[3] MALAVAKIS Th. (1985) Syntaktika ke epitonika phenomena (Phénomènes Syntaxiques et Intonatifs). Actes du VI^{e} Colloque de Linguistique Grecque, Thessaloniki, 41-56.

[4] MALAVAKIS Th. (1987) Intonation Patterns in Greek. A paraître in Proceedings of the 11th Congress of Phonetic Sciences, Tallinn, URSS.



LES SIX STRUCTURES PRINCIPALES-0

Fig. 1

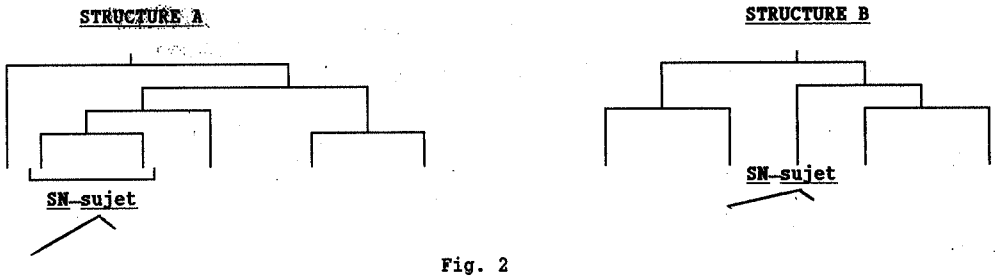


Fig. 2

STRUCTURE RYTHMIQUE DE LA PHRASE FRANCAISE

STATUT THEORIQUE ET DONNEES EXPERIMENTALES

Philippe Martin
Experimental Phonetics Laboratory
University of Toronto
300 Huron St.
Toronto, Ont.
Canada M5R 2L4

ABSTRACT

Temporal aspects of prosodic words can be considered a priori as independent of the prosodic and the syntactic structures of the sentence.

Experimental data show that for sequences of 3 and 4 prosodic units (sequences of syllables ending with a stressed syllable), the temporal patterns are in fact independent of the syntactic and melodic characteristics of the sentence, except for the length of the final (stressed) syllable.

The average durations of the syllables inside the prosodic words tend to be either constant among the consecutive prosodic words of the sentence in cases where the prosodic and syntaxic structures are not congruent, or to be modulated in order to compensate for the unequal number of syllables in the prosodic words where prosody and syntax are congruent.

INTRODUCTION

Dans le cadre d'un modèle phonosyntaxique de l'intonation, les mouvements mélodiques nécéssaires à la synthèse peuvent être dérivés soit de la structure syntaxique de la phrase, soit d'un principe d'eurythmie [1]. Cette dernière approche permet du reste de faire l'économie d'un analyseur syntaxique pour la génération de contours prosodiques à partir du texte [2],[3].

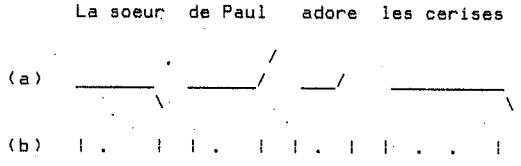
Cependant, les patrons rythmiques des mots prosodiques, tout aussi importants que les faits mélodiques pour obtenir une synthèse de bonne qualité, sont moins bien connus. En particulier, leur statut théorique par rapport aux structures mélodique et syntaxique de la phrase devrait être établi pour mieux interpréter les données expérimentales de plus en plus nombreuses qui apparaissent dans la littérature.

HYPOTHESES

Le développement récent des recherches sur le rythme de la phrase [4][5] [6][7][8], pose le problème du rapport qui existerait entre les faits de rythme et de mélodie. Les données expérimentales disponibles, parfois contradictoires, ne permettent pas toujours d'établir clairement le statut des faits temporels. En effet, alors que beaucoup d'aspects semblent indiquer une corrélation avec la syntaxe (dans les cas de levée d'ambiguité par insertion de pauses par exemple), les observations phonétiques mènent au contraire à conférer au rythme une valeur uniquement prosodique.

De même que l'existence d'une structure prosodique peut être posée indépendamment de la structure syntaxique de la phrase, on peut imaginer qu'il existe une structure rythmique (SR), a priori indépendante des structures prosodique (SP) et syntaxique (SS).

Cette structure rythmique est définie comme une organisation hiérarchique d'unités minimales rythmiques, ou mots rythmiques. Comme la structure prosodique, elle est planaire et connexe. Un mot rythmique est alors la plus petite unité rythmique contenant une et une seule syllabe accentuée (qui ne porte ni un accent d'insistance ni un accent emphatique), et apparaît comme une succession de durées syllabiques abstraites. Le mot rythmique correspond donc au mot prosodique.



(a): séquence de mots prosodiques
(b): séquence de mots rythmiques

En tant que durée abstraite, chaque élément d'un mot rythmique constitue un coefficient multiplicateur de la durée intrinsèque de chaque syllabe, reflétant la complexité des mouvements articulatoires nécessaires à leur accomplissement [8],[9].

Le problème est alors d'établir s'il existe un rapport entre les réalisations des mots rythmiques et d'autres éléments d'une phrase donnée qui en conditionneraient soit les unités successives (durées totales de chaque mot rythmique), soit la composition de chaque unité (le motif des durées relatives à l'intérieur d'un mot rythmique).

Trois situations peuvent être envisagées pour définir les rapports entre la structure rythmique et la structure prosodique:

Hyp. 1: la structure rythmique est indépendante à la fois de la structure prosodique et de la structure syntaxique;

Hyp. 2: la structure rythmique dépend de la structure prosodique, mais est indépendante de la structure syntaxique;

Hyp. 3: la structure rythmique est indépendante de la structure prosodique mais dépend de la structure syntaxique.

L'hypothèse ou la structure rythmique dépendrait des deux autres structures à la fois est exclue du fait que la structure prosodique est indépendante a priori de la structure syntaxique.

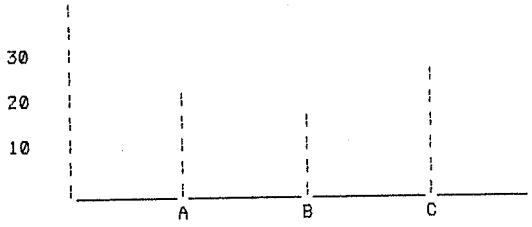
MOT RYTHMIQUE

Par définition, le mot rythmique se termine en français par une syllabe accentuée. Dans sa réalisation, cette syllabe se différencie des syllabes inaccentuées qui précèdent par un changement de fréquence fondamentale, une augmentation ou une diminution de durée, ou une combinaison des deux.

Les données décrites dans [10], relatives à des mots rythmiques de 2 à 7 syllabes de configurations diverses et organisés dans des structures à trois mots prosodiques, montrent que l'augmentation de durée dans des structures prosodiques à trois éléments est proportionnelle à l'importance de la frontière syntaxique correspondante, donc au contour mélodique indiquant cette frontière.

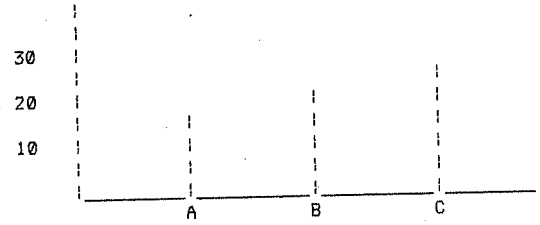
Structure (A (B C))

Durée [ms]



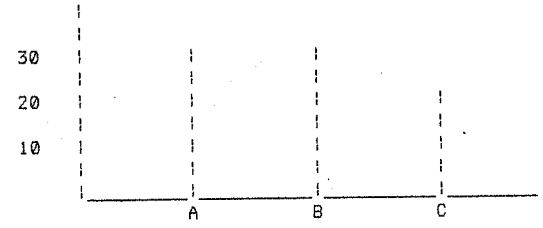
Structure ((A B) C)

Durée [ms]



Structure (A)(B)(C)

Durée [ms]



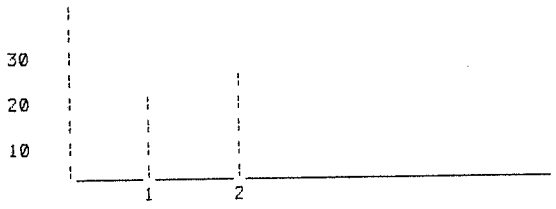
Durée syllabique moyenne (type CV) de la syllabe accentuée pour 3 structures prosodiques.
(3 locuteurs, 47 phrases)
D'après Y-M. Park [10].

Les caractéristiques des syllabes accentuées de structures de type énumératif à trois ou quatre éléments [11] font donc apparaître une diminution de durée des syllabes accentuées finales.

Contrairement aux observations portant sur une alternance dans le mot rythmique entre des durées courtes et brèves [5], ces deux études font état d'une augmentation graduelle de durée de la première à la dernière syllabe du même mot rythmique, avec une possible diminution de longueur de l'avant dernière syllabe lorsque le nombre d'éléments atteint 4 ou 5. Ici encore, ces configurations semblent indépendantes de la position du mot rythmique dans la phrase, ainsi que de sa structure syntaxique.

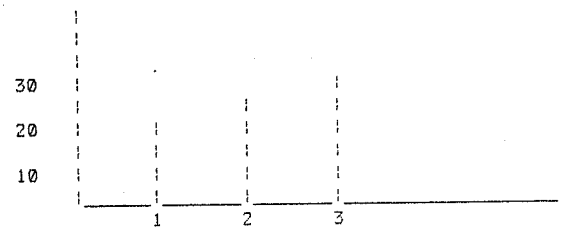
Motif rythmique de séquences de 2 syllabes

Durée [ms]



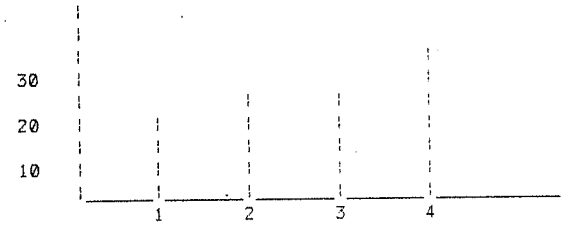
Motif rythmique de séquences de 3 syllabes

Durée [ms]



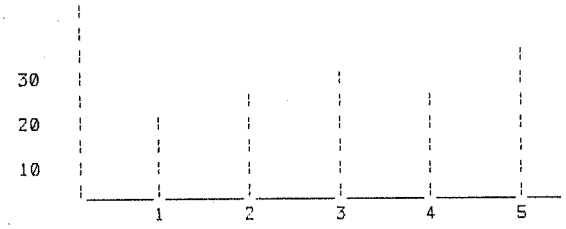
Motif rythmique de séquences de 4 syllabes

Durée [ms]



Motif rythmique de séquences de 5 syllabes

Durée [ms]



Structure temporelle (motif) de mots rythmiques de 2 à 5 syllabes pour 3 structures prosodiques différentes. (3 locuteurs, 47 phrases, moyenne des structures syllabiques V, CV, CVC, CCV et CVV)
(D'après Y-M. Park [10]).

STRUCTURE RYTHMIQUE ET STRUCTURE PROSODIQUE

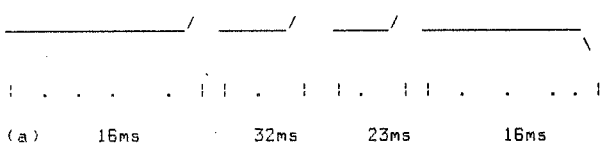
Le patron temporel des mots rythmiques semblant indépendant des structures syntaxique et prosodique, on peut se demander si une corrélation peut être établie avec les durées globales des mots rythmiques, plutôt qu'avec leur motif.

Les observations faites par Wenck et Wiolland [7] montrent par exemple que la durée moyenne des syllabes d'un mot rythmique ont tendance à diminuer lorsque l'intervalle entre deux syllabes accentuées successives est grand, de manière à égaliser la durée des mots rythmiques, et réaliser ainsi une meilleure eurythmie au sens défini dans [1]. On constate alors une modulation des durées des mots rythmiques tendant à l'isochronie.

Les mesures rapportées dans [11] relatent également ce phénomène pour des structures présentant des différences importantes du nombre de syllabes dans des mots rythmiques successifs, en parallèle avec des réalisations moins nombreuses ou la compensation temporelle ne se fait pas.

Ainsi l'exemple suivant, de structure 5.2.2.5, tend à l'isochronie:

Ces petits chatons très tôt du thon ils ont dégusté

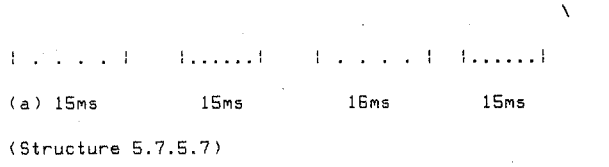


(a) 16ms 32ms 23ms 16ms

(a): Durée syllabique moyenne

Cependant, parmi les exemples décrits dans [11], certaines réalisations, prononçées par le même locuteur, sont plutôt isosyllabiques:

Ces petits chatons, très tôt dans l'après-midi, des cynhodorons, ils en avaient dégustés



(a) 15ms 15ms 16ms 15ms

(Structure 5.7.5.7)

Ces réalisations isosyllabiques sont toutefois plus rares dans les conditions d'énonciations de cette étude (environ 12% pour 27 phrases de 4 mots rythmiques de 2, 5 et 7 syllabes prononçées par 2 locuteurs).

Il semble donc que deux mécanismes modulant la durée des syllabes des mots rythmiques coexistent:

- l'un tendant à l'isochronisme entre les syllabes accentuées, donc à l'isochronisme des mots rythmiques, en diminuant la durée des mots rythmiques longs et en augmentant celle des mots courts;

- l'autre tendant à conserver une durée égale pour toutes les syllabes quelle que soit leur nombre dans le mot rythmique (isosyllabisme).

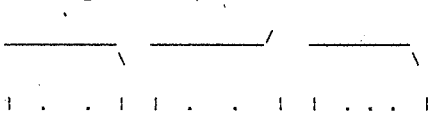
Si l'on rapproche ces constatations du principe d'eurythmie présenté dans [1], selon lequel les structures prosodiques non conguentes avec la structure syntaxique choisies par les locuteurs sont celles qui réduisent la dirythmicité en égalisant le nombre de syllabes des mots prosodiques, on est amené à conclure que la solution non syntaxique ne sera réalisée qu'en cas d'isosyllabisme. Les cas observés dans [11] font du reste état de conditions sémantiques (mots rares) et phonologiques (mots peu courants de 7 syllabes) particulières, qui montrent que les réalisations des locuteurs sont dans ces exemples clairement non syntaxiques.

Autrement dit, l'eurithmicité pourra être obtenue même en cas de congruence syntaxique de structures asymétriques au prix d'une compensation isochronique tendant à ramener l'eurythmie par une modulation de la durée des mots prosodiques.

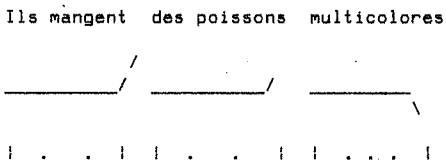
Au contraire, une solution isosyllabique entraine la non congruence syntaxique des structures asymétriques par le choix de structures prosodiques eurythmiques.

Les solutions isochronique, congruente avec la structure syntaxique, et isosyllabique, non congruente avec la syntaxe apparaissent dans l'exemple suivant:

Solution isochronique:



Solution isosyllabique:



CONCLUSION

Les données expérimentales relatives à des structures prosodiques de 3 et 4 unités montrent que les motifs temporels des mots rythmiques associés aux mots prosodiques ne semblent pas dépendre de la hiérarchie particulière des structures syntaxique ou prosodique de la phrase (sauf pour la durée de la syllabe accentuée, lié au contour mélodique indiquant la position de l'unité dans la structure prosodique).

Par contre, les durées moyennes des syllabes accentuées qui les composent sont modulées de manière à réaliser des séquences eurythmiques lorsqu'il y a congruence de la structure prosodique avec la syntaxe.

Si cette congruence n'est pas réalisée, les mots rythmiques ont tendance à être isosyllabiques. L'eurythmie est alors obtenue par le choix d'une structure égalisant le nombre des syllabes aux différent niveaux de la hiérarchie prosodique.

REFERENCES

[1] Ph. Martin. "Prosodic and Rhythmic Structures in French, Linguistics, 1987,(sous presse).

[2] Ph. Martin, "Structure prosodique et structure rythmique pour la synthèse", Actes des 15èmes JEP, Aix-en-Provence, 89-91. (1986).

[3] Ch. Sorin, D. Larreur et R. Llorca, "A Rhythm-Based Parser for Text-to-Speech Systems in French", Proc. XIth Int. Congress of Phonetic Sciences,- (1987).

[4] G. Caelen-Haumont, "Le rythme dans la parole: une revue des études portant sur le français", Le Rythme ,avec H. Meschonnic, Colloque d'Albi, I, (1983).

[5] D. Duez, et Y. Nishinuma, "Some evidence on rhythmic patterns of spoken French:, Perilus, 4, Stockolm, 30-40, (1985).

[6] R. LLorca, "Essai de systématisation du rythme de la parole française", Actes Séminaire Prosodie et Reconnaissance, Aix-en-Provence, 165-192. (1982).

[7] B. J. Wenck and F. Wiolland, "Is French really syllable-timed?", Journal of Phonetics, 10, 193-216. (1982).

[8] A. Di Cristo, "La durée intrinsèque des voyelles du français", Travaux de l'Institut de Phonétique d'Aix, 211-235. (1980).

[9] K. Bartkova et Ch. Sorin, "Predictive Model of Segmental Durations in French", JASA 77, suppl. 1, S54. (1985).

[10] I. Guaitella, "Considérations sur le rythme d'une structure prosodique à un seul niveau", Mémoire de DEA, Université de Provence, (1986)

[11] Y-M. Park, "Etude sur le rythme en français", Mémoire de DEA, Université de Provence, (1986).

ANALYSE ACOUSTIQUE DE LA STRUCTURATION RYTHMIQUE DU FRANCAIS ORAL

Valérie Pasdeloup

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

The analysis of rhythm in speech is intricate owing to the fact that the rhythm does not intermingle with its phonic realization, but results from the interaction between different systems that constitute the meaning, the intonative, accentual, syntactic, enunciative, lexical, gestural and attitudinal systems, among others.
Working on the acoustic and syntactic analysis of two corpus, we present a trial of rhythm modelization, which include a three level organization of rhythm (rhythmic macro-sequences, temporal sequences and accentual groups), and a series of rhythmic rules made up of linguistic and phonotactical rules.

INTRODUCTION

Des phénomènes observés dans différentes langues, tels que le principe d'alternance de durée syllabique (1), ou le principe d'alternance rythmique faible/fort (2), ou la récurrence plus ou moins régulière d'accents, de patrons intonatifs (3) ou de séquences temporelles ou accentuelles ne semblent pas avoir une fonction uniquement linguistique, mais semblent résulter d'un compromis entre les activités linguistiques d'une part, et biologiques, psycho-motrices et cognitives d'autre part. En effet, toute réalisation motrice fondée sur des principes récurrents est plus économique sur le plan de la production motrice (5); si de plus ces principes de récurrence se réalisent dans un cadre linguistique où ils remplissent des fonctions linguistiques, ces principes sont à la fois économiques sur le plan de la production motrice et rentables sur le plan linguistique. Comme le dit P.Fraisse (1967:24) : "l'homme est créateur de rythmes en se dégageant peu à peu des contraintes biologiques où ces rythmes prennent naissance, sans qu'il en soit cependant jamais complètement séparés."
Notre hypothèse est la suivante : plus on descend dans la hiérarchie linguistique (syntaxique et énonciative entre autres), moins les principes d'organisation linguistique sont opérants, et plus les principes d'organisation non linguistique tendent à imposer leur structuration.

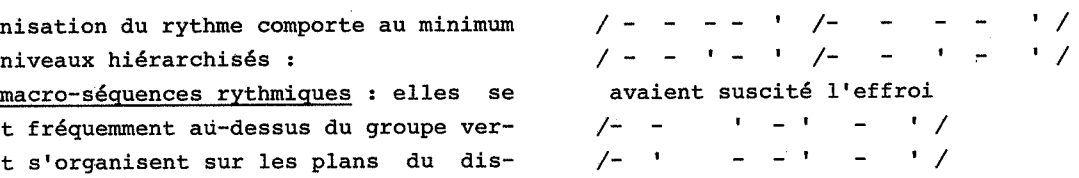
La complexité dans la structuration rythmique réside principalement dans le fait qu'elle résulte de l'interaction des niveaux lexical, syntaxique, accentuel, intonatif, gestuel et attitudinal, niveaux par eux-mêmes déjà porteurs de sens et fonctionnant en relation les uns avec les autres. Pour H.Meschonnic (1982:70) : "Le rythme est organisation du sens dans le discours. S'il est organisation du sens, il n'est plus un niveau distinct, juxtaposé. Le sens se fait dans et par tous les éléments du discours. La hiérarchie du signifié n'en est plus qu'une variable, selon les discours, les situations." La structuration rythmique perçue dans la parole n'est donc pas présente au seul niveau acoustique (7), et ne peut être assimilée à un rythme musical avec le sens en plus. Le décodage des paramètres acoustiques, sans l'appui des autres niveaux où s'élabore le sens, ne semble pas permettre d'en faire l'étude (8). De plus, la perception joue un rôle actif de normalisation, non négligeable dans la structuration rythmique (6); l'auditeur projette vraisemblablement des structures rythmiques sur un continuum acoustique flou, qui peut même être discordant, quand ce n'est pas localement incompatible avec l'interprétation qu'il en fait.
A partir de l'analyse acoustique de deux corpus (spontané et lu), nous proposons un

essai de modélisation du rythme qui comprend une organisation du rythme à trois niveaux, de même qu'une série de règles rythmiques hiérarchisées.

ORGANISATION HIERARCHIQUE DU RYTHME

La syllabe est l'unité rythmique minimale, puisqu'elle semble être une unité tant sur les plans de la production que de la perception (9); d'un point de vue rythmique les syllabes fonctionnent comme des éléments percussifs. En français, la syllabe inaccentuée constitue l'unité temporelle, la valeur de référence dans l'organisation des durées; les variations de débit, c'est à dire les modifications rapides ou progressives de cet valeur étalon, jouent un rôle au niveau macro-rythmique.
La syllabe est non seulement une unité constituante, mais aussi structurante, dans la mesure où des structurations s'établissent par le nombre de syllabe, ce qui semble être tout particulièrement le cas en français (10). Le nombre de syllabe de groupes accentuels, de séquences temporelles, de syntagmes, de groupes de sens, ou de groupements d'autre nature, peut être identique, dans un rapport du simple au double, ou dans un rapport multiple ou graduel.

L'organisation du rythme comporte au minimum trois niveaux hiérarchisés :
- les macro-séquences rythmiques : elles se situent fréquemment au-dessus du groupe verbal, et s'organisent sur les plans du discours et de l'énonciation. Des séquences temporelles, des groupes accentuels, des groupes syntaxiques, syntactico-sémantiques et intonatifs forment des ensembles unis au niveau macroscopique par des principes de récurrence, d'équilibrage, de symétrie-miroir, mais aussi de contraste, de rupture et de gradation (3-11).
- les séquences temporelles : ces séquences sont délimitées par un fort allongement et/ou une pause. Elles correspondent à des groupes de sens qui sont congruents à l'organisation syntaxique ou qui la réorganisent en fonction de l'énonciation ou de l'expressivité (thème/rhème, accent d'insistance etc...), et dont le découpage est à la fois économique sur les plans de la production motrice et du décodage perceptif, et rentable sur le plan linguistique.
Trois principes phono-tactiques opèrent au niveau des séquences temporelles :
- la sur-segmentation de longs syntagmes ou groupes de sens en séquence temporelle;
- la sous-segmentation de courts syntagmes ou groupes de sens en séquence temporelle;
- l'équilibrage syllabique et/ou temporel des séquences temporelles (11).
Un changement de mode dans l'utilisation de ces principes est significatif sur le plan discursif, c'est à dire au niveau des macro-séquences rythmiques.
- les groupes accentuels : ces groupes sont les groupements rythmiques minimaux. Ils sont constitués d'un nombre x de syllabes inaccentuées (habituellement entre un et quatre) suivies d'un accent, et ne correspondent souvent pas à des groupes de sens, mais peuvent parfois coïncider avec des mots. Les groupes accentuels sont donc moins contraints par le linguistique que les séquences temporelles, qui correspondent toujours à des groupes de sens.
Les deux exemples ci-dessous montrent l'organisation en groupe accentuel et en séquence temporelle réalisée par deux sujets sur la même phrase d'un corpus lu (les traits horizontaux désignent les syllabes accentuées, les traits verticaux les inaccentuées, et les barres transversales les limites des séquences temporelles) :



Le tableau 1 illustre l'organisation rythmique (macro-séquences rythmiques et séquences temporelles) et les principaux principes rythmiques d'un court extrait de discours spontané (émission Droit de réponse de Michel Polac, 5 Avril 1986). On remarquera dans ce tableau que le groupe verbal ABC constitue trois séquences temporelles, alors qu'à l'inverse, les trois groupes verbaux D1 D2 D3 ne constituent qu'une seule séquence temporelle (D). Ainsi l'organisation ternaire produite par la récurrence d'une même structure temporelle en ABC, réapparaît ensuite sur le plan syntaxique, dans la récurrence des trois groupes verbaux de la séquence temporelle D; le rapport qui lie la structuration temporelle à la structuration syntaxique est inversé et contribue à créer une rupture sur le plan macro-rythmique.

TABLEAU 1 (U : syllabe brève, - : syllabe longue)

		entre une série de syllabes brèves et CONTRASTE . une (ou deux) syll. longues	
5 SEQUENCES TEMPORELLES	A	U U U U U / et a lo rs_au dé . - - / pa rt_heu	RECURRENCE d'une même structure temporelle (en A B C)
	B	U U U U U U / donc vou s_a vez pas eu . - / d'chance	
	C	U U U / pui sque vous . - / heu	
	D1	U U U U U U U U U U U U / y'a eu deux per sonnes d'la fa mi lle_au cho mâ	RUPTURE RYTHMIQUE entre des structures temporelles différentes (ABC/D et D/E)
	D2	U U U U U U U U U U U U U / ge_et en plus un d'vos fils a fait une mé nin gi	
	D3	U U U U U U / te_et là vou s_a vez cra . - / qué	
	E	U U U U U U U U / vou s_a vez fait une p'tite dé pre . - / ssion	

3 MACRO-SEQUENCES RYTHMIQUES

REGLES RYTHMIQUES

De nombreux travaux tendent à montrer que la place de l'accent en français n'est pas systématiquement à la finale du mot (2-4); l'accent final n'est pas toujours réalisé, ou l'est sur des éléments supposés inaccentuables (modaux, auxiliaires, conjonctions et autres morphèmes grammaticaux), et un accent secondaire peut se manifester en dehors de la syllabe finale des mots de plus de deux syllabes. La question est de savoir si cet accent secondaire est un accent lexical dont le domaine des propriétés est le mot, ou un accent rythmique dont le domaine des propriétés est par exemple, une séquence temporelle, un groupe intonatif ou un groupe de sens.

Les six règles rythmiques présentées ci-dessous comprennent des règles linguistiques et phono-tactiques. Les règles linguistiques rendent compte d'une structure syntaxique ou énonciative -de surface-, hiérarchisée à deux niveaux. Les règles phono-tactiques interviennent ensuite, déterminant de façon non systématique la présence et la position d'accents secondaires (ou ictus mélodiques (12)).

Règles linguistiques :

- Régle 1 : démarcation en séquence temporelle des constituants de premier niveau dans la hiérarchie syntaxique ou énonciative; accent réalisé à la finale par un fort allongement et/ou une pause.

Exemple :

Eric et Pat avaient barboté mes chats

/ / /

-Règle 2 : démarcation en séquence temporelle ou en groupe accentuel des constituants de second niveau dans la hiérarchie syntaxique ou énonciative.

Exemples :

Eric et Pat avaient barboté mes chats

/ / / /

/- ' - ' /- - - - ' - ' /

Règles phono-tactiques :

- Règle 3 : l'accent secondaire peut être situé sur l'antépénultième d'un mot portant déjà un accent final. Cet accent secondaire fait bloc avec l'accent final et le renforce.

Exemples :

Phonéticiennement vôtre : l'hurluberlu

/ - ' - - - ' / - ' - '/

C'était un emberlificoteur

/ - ' - - - - - ' - ' /

La dégradation de ma maladie (...)

/ - - ' - ' / - - ' - ' /

- Régle 4 : l'accent secondaire peut être situé sur la première syllabe d'un mot ne portant pas obligatoirement l'accent final.

Exemples :

La phénoménologie est une philosophie

/ - ' - - - - ' / - - ' - - ' /

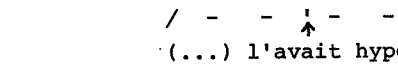
Il pouvait supporter une dose infinitésimale

/- - - ' - ' / - - ' - - - - ' /

- Règle 5 : dans un mot plurimorphémique, l'accent peut être situé à la frontière des morphèmes (à la finale d'un morphème interne); le mot est accentué à la finale, et ne reçoit en général qu'un seul accent secondaire, même s'il est composé de plus de deux morphèmes.

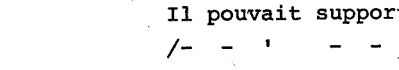
Exemples :

C'est anticonstitutionnel



(...) l'avait hypersensibilisé

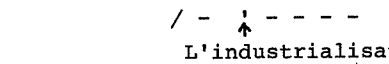
Il pouvait supporter une dose infinitésimale



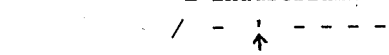
- Règle 6 : règle du premier accent; en début de phrase on se hâte d'accentuer habituellement sur la deuxième syllabe de la phrase (la première syllabe du mot lorsqu'il est précédé d'un article d'une syllabe), même si le mot ne porte pas d'accent final.

Exemples :

La généralisation de ce processus (...)



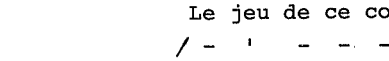
L'industrialisation de la zone Est (...)



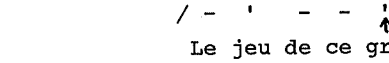
L'usage de l'accent secondaire n'est pas systématique, et certains locuteurs semblent en faire plus usage que d'autres. La probabilité d'apparition des règles rythmiques est fortement contrainte par le contexte accentuel immédiat et par la structuration rythmique de l'ensemble de l'énoncé.

Exemples :

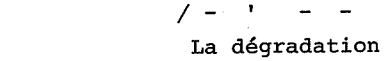
Le jeu de ce comédien (...)



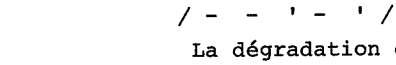
Le jeu de ce gros comédien (...)



La dégradation de ma maladie (...)



La dégradation de ma grave maladie (...)



Suivant les stratégies discursives employées, deux procédés complémentaires d'accentuation s'observent en français :

- la tendance à mettre l'accent à la finale d'un mot ou d'un groupe (séquence temporelle ou syntagme), et optionnellement sur l'antépénultième (décompte syllabique de droite à gauche).
- la tendance à mettre l'accent au début d'un mot ou d'une phrase (décompte syllabique de gauche à droite).

I.Fonagy (4) parle de "tendance centrifuge" lorsque les deux procédés utilisés conjointement forment un "arc accentuel" qui relie le début et la fin d'une unité lexicale ou syntaxique, et de rythme staccato ou en "dent de scie" lorsqu'une seule des deux tendances est réalisée.

CONCLUSION

Le rythme est l'organisation du sens dans le discours. Il inclut essentiellement deux composantes, l'une acoustique et l'autre syntactico-sémantique, qui interagissent. La composante acoustique du rythme se définit comme une structuration temporelle sur laquelle se développe une organisation accentuelle. Un accent secondaire s'observe en plus de l'accent final; il prend place généralement sur l'antépénultième ou sur la première syllabe du mot.

BIBLIOGRAPHIE

(1) DUEZ, D.; NISHINUMA, Y. (1985), Le rythme en français : alternance des durées syllabiques, Travaux de l'Institut de Phonétique d'Aix-en-Provence, 10, 151-169.
(2) VERLUYTEN, S.P. (1983), Phonetic Reality of Linguistic Structures : the Case of (Secondary) Stress in French, Proceedings of the Tenth International Congress of Phonetic Sciences, Utrecht, M.P.R. Van den Broecke et A.Cohen eds., 522-526.
(3) FONAGY, I. et J. (1983), L'intonation et l'organisation du discours, Bulletin de la Société de Linguistique de Paris, 28, Klincksieck, 161-209.
(4) FONAGY, I. (1980), L'accent français : accent probabilitaire, L'accent en français contemporain, Studia Phonetica, 15, Didier, 123-233.
(5) FRAISSE, P. (1967), Psychologie des rythmes humains, Colloque sur les rythmes, Lyon 4 Déc. 1967, Journal Français d'Oto-Rhino-Laryngologie, sup. n°7; SIMEP, 23-33.
(6) FRAISSE, P. (1956), Les structures rythmiques,Studia Psychologica, Publications universitaires de Louvain, 124 p.
(7) MESCHONNIC, H. (1982), Critique du rythme, Verdier, Lagrasse, 713 p.
(8) PASDELOUP, V. (1987), Analyse temporelle et perceptive de la structuration rythmique d'un énoncé oral, Travaux de l'Institut de Phonétique d'Aix-en-Provence, n°11, à paraître.
(9) SEGUI, J. (1983), The syllable : a Basic Perceptual Unit in Speech Processing ?, Attention and Performance : Control of Language Processes, H.Bouma et D.G. Bouwhuis eds., Lawrence Erlbaum Associates, 165-181.
(10) DE CORNULIER, B. (1979), Problèmes de métrique française, Thèse de Doctorat, Université de Provence, Faculté des Sciences de Lumigny.
(11) WIOLAND, F. (1984), Organisation temporelle des structures rythmiques du français parlé, Bulletin des rencontres régionales de linguistique, Lausanne, 293-322.
(12) ROSSI, M. (1985), L'intonation et l'organisation de l'énoncé, Phonetica, 42, 135-153.

ETUDE DE L'INTONATION POUR LA SYNTHESE DE L'ALLEMAND

B. SCHNABEL*, F. EMERARD*

* I.C.P. - I.P.G.
B.P. 25X
38000 GRENOBLE Cedex

* C.N.E.T.
B.P. 40
22301 LANNION Cedex

Intonation models of a given language in speech synthesis are generally based upon one of the following two principles : either to use a model assuming the congruence between intonation patterns and syntactical structures or to determine the intonation patterns using the phonetical segmentation of a given data-base.

The goal of our study is to develop a tool for the generation of intonation based as well on a syntactical analysis as on rythmic and syllabic data. The values for the fundamental frequency and the segmental duration for each element are defined by its position in the main clause, the subordinate or the phrase. Furthermore, the subdivision of the sentences into minimal elements and the detection of vocalic duration maxima are realized by means of orthographical and syllabic criteria.

I INTRODUCTION

Des études systématiques ont été menées afin de déterminer le degré de congruence - s'il en est - entre la syntaxe et l'intonation [1,2,3].

Aussi variées que puissent être les réponses à cette hypothèse, il est toujours possible de décrire les contours prosodiques d'une réalisation particulière en s'appuyant sur une catégorisation syntaxique.

Postulant qu'un modèle intonatif peut être déduit d'une réalisation seule, dans la mesure où celle-ci est perçue comme naturelle, il est loisible de proposer des contours mélodiques à chaque unité syntaxique (pour être appliqués à la synthèse de la parole) sur la base d'un corpus lu par un seul locuteur. De plus une réalisation naturelle fournit toutes les informations concernant la durée segmentale d'un phonème [4].

Cette étude a une double intention (sans exigence sur un plan théorique) : définir l'évolution de la fréquence fondamentale sur des unités plus petites que la phrase, ainsi qu'obtenir les durées segmentales intrinsèques et contextuelles de chaque son, en fonction de paramètres phonétiques, morphologiques et syntaxiques. Par ailleurs, le découpage de la phrase comme la détection des maxima de la durée vocalique est effectué automatiquement à l'aide de critères orthographiques et/ou syllabiques.

Par la suite, ce modèle remplacera le traitement actuel de la prosodie [5], qui s'est avéré trop élémentaire.

II METHODES

Contrairement à la langue française, l'allemand est une langue à accent mobile, ce qui amène une distribution complexe de l'accent au niveau morphologique et syntaxique, comme le montre la figure 1 (d'après DELATTRE [6]).

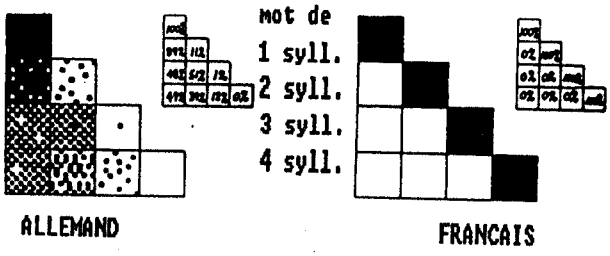


Figure 1 : Comparaison de la place de l'accent en allemand et en français

L'interêt de cette étude est donc centré sur un classement des valeurs de la durée et de la fréquence fondamentale au moyen de critères morphologiques et syntaxiques. De plus, les valeurs divergentes selon le contexte phonétique ont été retenues pour les durées segmentales des consonnes appartenant à des groupes consonantiques.

A) LE CORPUS

Le corpus est constitué de 45 phrases adaptées à une étude de l'intonation, et d'un texte continu d'environ 3 minutes. Les 45 phrases comprennent des énonciatives, interrogatives, exclamatives et impératives. En outre les phrases énonciatives incluent des groupes de continuation (progrédients mineurs et majeurs) ainsi que des énumérations. Le texte représente un résumé de l'histoire "Der Rattenfänger von Hameln" ; "Le joueur de flûte de Hamelin".

L'ensemble du corpus a une durée d'environ 5 mns 1/2. Il a été lu par une locutrice germanophone de naissance, née à ESSEN (R.F.A.) en 1955, et enregistré en chambre sourde sur magnétophone Nagra. Quelques phrases d'entrainement ont précédé l'enregistrement qui a été répété trois fois. La réalisation jugée la plus naturelle à l'écoute a été retenue pour cette étude.

B) EXPLOITATION DU CORPUS

L'étiquetage du corpus a été effectué manuellement sur les tracés spectrographiques [7]. Ont été mésurées les durées de chaque segment phonétique ainsi que trois valeurs de fréquence fondamentale (début, "milieu" , et fin) pour chaque voyelle.

Après la suppression des résultats de mesure jugés trop "déviants", les valeurs moyennes des durées et celles des fréquences fondamentales ont été stockées dans des tableaux de structures différentes reposant sur des catégorisations (à plusieurs niveaux) tant phonétiques (axe des abscisses) que syntaxiques (axe des ordonnées).

B-1 Tableaux des durées

Ils contiennent les valeurs moyennes des durées segmentales mesurées sur les phrases et le texte.

Deux structures de tableaux ont été retenues pour les deux types de sons : vocalique et consonantique. Dans chacun de ces deux cas, trois tableaux regroupent les données suivant qu'elles ont été extraites de phrases énonciatives, interrogatives ou exclamatives.

Les six tableaux ainsi constitués respectent la même décomposition syntaxique et morphologique, comme le représente la figure 2.

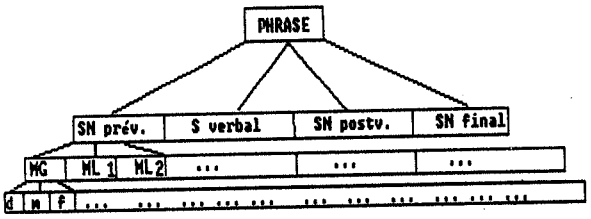


Figure 2 : Le découpage de la phrase

SN = syntagme nominal
prév. = préverbal
postv. = postverbal
MG = mot grammatical
ML = mot lexical (1 avec minuscule, 2 avec majuscule)
d = syllabe initiale du mot
m = syllabe à l'intérieur d'un mot
f = syllabe finale du mot

En revanche, la catégorisation phonétique diffère selon le type vocalique ou consonantique :

- type vocalique

Les voyelles sont catégorisées par leur quantité (ou qualité : longues, brèves, diphthonguées), puis par la nature de la syllabe (ouverte ou fermée) et enfin par leur accentuation (syllabe accentuée ou non ; l'accent secondaire propre à l'allemand n'ayant pas été retenu ici), comme schématisé sur la figure 3.

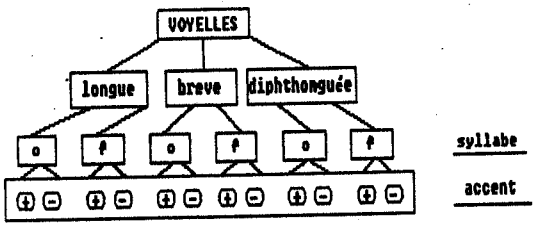


Figure 3 : La catégorisation des voyelles
o = syllabe ouverte
f = syllabe fermée

- type consonantique

Les consonnes sont classées d'après la catégorisation traditionnelle en occlusives /p,t,k,b,d,g/, mi-occlusives /pf,ts/ et fricatives /f,s,ch,x,v,z/ sourdes et sonores, en nasales /m,n,ng/, liquides /l,r/ et aspirantes /h/, ainsi qu'en semi-voyelle /j/ et sibilantes /R,L,M,N/, puis par leur accentuation, comme le montre la figure 4.

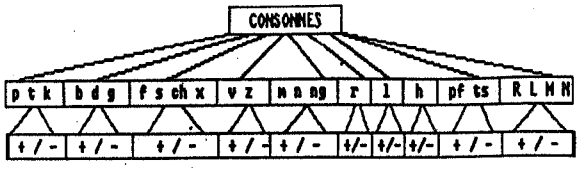


Figure 4 : La classification des consonnes
+ / - = la syllabe porte oui ou non un accent

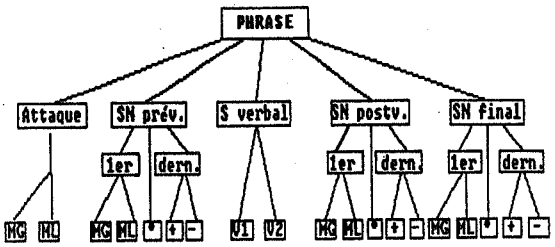


Figure 5 : Catégorisation syntaxique pour le tableau de F0

MG = mot grammatical
ML = mot lexical
V1 = verbe non - séparé
V2 = verbe séparé
* = mot à l'intérieur d'un syntagme
+ = mot suivi d'une pause
- = mot non - suivi d'une pause
1er = premier mot du syntagme
dern.= dernier mot du syntagme

B-2 TABLEAUX DES FREQUENCES FONDAMENTALES

Les tableaux de Fo rassemblent les moyennes de chacune des trois valeurs mesurées sur les voyelles dans les phrases. La catégorisation syntaxique diffère de celle utilisée pour les durées. Ici, une importance prépondérente est accordée à la position de la syllabe. Un cas particulier est donc traité pour les débuts et fins de syntagmes comme le précise la figure 5.

L'axe des abscisses contient, à coté des données phonétiques des informations morphologiques sur le mot. Les mots lexicaux sont regroupés d'après leur nombre de syllabes et la place de l'accent, en supposant que l'accent en allemand ne se situe pas sur la dernière syllabe d'un mot de plus de 4 syllabes (voir aussi figure 1), comme le schématise la figure 6.

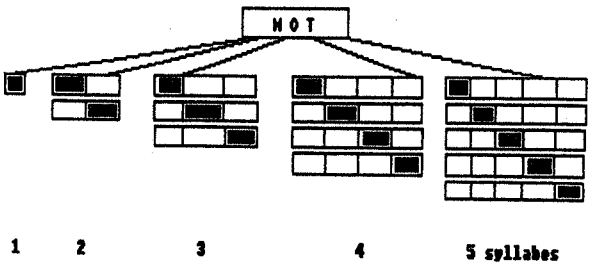


Figure 6 : Subdivision morphologique du mot d'après le nombre de syllabes et la position de l'accent

■ = position de l'accent

Cette démarche permettra par la suite de calculer les durées segmentales et les évolutions de la fréquence fondamentale de manière plus fine :
* pour les durées, chaque son aura une valeur adaptée à sa nature et à sa place en fonction du contexte,
* pour la fréquence fondamentale, la hauteur de chaque voyelle sera définie par la place du mot (dans le syntagme) auquel elle appartient. L'évolution de Fo entre deux voyelles sera calculée par interpolation pour les consonnes sonores.

III OBSERVATION

Au moment de la rédaction, nous ne disposions que de premiers résultats concernant la durée et la fréquence fondamentale. Les résultats définitifs de cette étude seront présentés lors de la communication même.

En résumé, les tableaux présentent les caractéristiques suivantes :

1) **Durées intrinsèques :**

- les maxima de durée se trouvent sur les mi-occlusives et les sibilantes,
- les consonnes sourdes /p,t,k,f,s,ch,x/ sont plus longues que les sonores /b,d,g,v,z/,
- les occlusives et les fricatives ont une durée supérieure à celles des autres consonnes,
- les nasales, liquides, aspirantes et semi-voyelles présentent sensiblement la même durée moyenne.

2) **Durées en contexte :**

- les consonnes et les voyelles présentent des variations similaires de durée dépendant de leur appartenance à une syllabe accentuée ou non,
- la plus grande variation contextuelle de durée est cependant réservée à la syllabe finale du syntagme,
- en fin de syntagme, la durée des voyelles est augmentée d'un facteur 1.5 dans une syllabe non accentuée et d'un facteur 2 dans une syllabe accentuée (indépendamment de la nature de la syllabe) ; celle des consonnes est augmentée de la même manière dans les seules syllabes fermées.

3) **Variations de Fo :**

- elles dépendent de l'accent et de la nature du syntagme,
- dans les syntagmes préverbaux et post-verbaux non finaux, Fo monte sur la syllabe accentuée, descend légèrement, puis remonte jusqu'à la syllabe finale (incluse),
- dans les syntagmes finaux, Fo descend sur l'accent du groupe, monte légèrement, puis descend jusqu'à la fin de la phrase,
- dans ces deux derniers cas, l'évolution de Fo est plus marquée si l'accent du groupe se situe sur la dernière syllabe,
- les syntagmes verbaux présentent des contours assez monotones.

IV REFERENCES

[1] BANNERT R. (1983) "Modellskizze für die deutsche Intonation". in Zeitschrift für Literaturwissenschaft und Linguistik 49, 9-34.
[2] MARTIN Ph. (1981) "Pour une étude de l'intonation". in L'Intonation : de l'acoustique à la sémantique. (ROSSI, ed.), 234 - 271.
[3] BAILLY G. (1986) "Un modèle de congruence relationel pour la synthèse de la prosodie du français". 15e JEP du GALF, Aix-en-Provence, 75 - 78.
[4] EMERARD F. (1977) Synthèse par diphones et traitement de la prosodie. Thèse de IIIe cycle, Grenoble.
[5] ZINGLE H. (1974) Traitement de la prosodie allemande dans un système de synthèse de la parole. Thèse d'Etat, Strasbourg.
[6] DELATTRE P. (1965) Comparing the phonetic features of French, German and Spanish. Heidelberg.
[7] MONNE J. (1983) "Programme de calcul de spectrogramme". in Traitement du signal de parole, ENST - Paris, 275-277.

REGLES D'ACCENTUATION EN GREC MODERNE
(Prévisibilité automatique)

Argyro TSEVA Michel CONTINI

Institut de la Communication Parlée - Institut de Phonétique de Grenoble
Université des Langues et Lettres
B.P. 25 X 38040 GRENOBLE CEDEX - FRANCE

ABSTRACT

The Modern Greek stress pattern, a language with a limited degree of stress possibilities (1), reflects in general the Ancient Greek stress system; stress generally occurs on the same syllable as Ancient Greek.

This study is a contribution to stress rules on the synchronic level (2). It is shown here that determining the position of stress depending on the recognition of the morphological structure of the word isn't satisfactory (P. GARDE, 1968). Stress is possible to predict, on the contrary, when the final phonetism of the word is taken into account. Among 31 734 Greek words examined, the latter method predicts the place of stress in 89 % of occurrences.

POSITION DU PROBLEME

L'apport de P. Garde dans sa classification typologique des langues d'après le rapport accent / morphème a été le point de départ de notre démarche. Cependant l'analyse morphologique des mots s'est avérée insatisfaisante en ce qui concerne la prévisibilité automatique de la place de l'accent en grec moderne.

Nous citons comme exemple les substantifs féminins comportant le morphème [-i-] pouvant présenter les trois possibilités accentuelles; accent sur la syllabe finale, pénultième et antépénultième (la dernière possibilité accentuelle n'est pas très répandue). Comme nous pouvons le constater dans les exemples suivants, tous les substantifs comportent la même désinence; [-a].

[prɔ'spaθ-i-a]	"effort"	dérivé de
[prɔspa'θ-ɔ]	"je m'efforce"	
[la'tr-i-a]	"adoration, culte"	dérivé de
[la'trev-ɔ]	"j'adore"	
[aði'c-i-a]	"injustice"	dérivé de
['aðik-ɔs]	"injuste"	
[pɔta'm-j-a]	"vallée de rivière"	dérivé de
[pɔ'tam-i]	"rivière"	

Pour cette catégorie de mots, la définition de la place de l'accent est liée à l'accentuation du mot d'origine. Ainsi, par exemple, les substantifs dérivés de verbes en [-evɔ] sont accentués sur la syllabe pénultième (ex. : [la'tria] "adoration" dérivé de [la'trevɔ] "j'adore"). Bien que la place de l'accent dépende de la structure morphologique du mot d'origine, on ne peut pas soutenir qu'elle dépend des propriétés accentuelles des morphèmes du mot dérivé. Ce dernier ne présente aucune indication pour la définition de la place de l'accent de manière automatique.

Le même problème soulevé pour les substantifs en [-ia] se retrouve dans les substantifs dérivés dépourvus du morphème dérivatif. Tout en sachant que, pour ces derniers, la délimitation de la place de l'accent est possible si l'on se réfère à des états antérieurs de la langue, il ne faut pas oublier, comme M. Triantaphyllidis le signale dans l'introduction de sa grammaire (1978, p. XXVIII), que les limites entre les anciens et les nouveaux phénomènes dérivationnels ne se distinguent pas toujours facilement, et que souvent elles n'existent même pas.

METHODE ADOPTEE

Nous avons effectué un test préliminaire qui nous a permis de constater que la prévisibilité de la place de l'accent est possible, dans une large mesure, si on tient compte de la structure finale du mot. Cette dernière est la position la plus riche en informations morphologiques puisqu'on y retrouve toujours la désinence et les morphèmes dérivatifs, s'ils en existent. Cette procédure apparaît également satisfaisante pour les mots de formation simple dont les trois possibilités accentuelles sont admises, bien que la désinence (au nominatif singulier) ne permette pas la délimitation de la place de l'accent, à quelques exceptions près.

Exemples :

[aðer'f-i]	"soeur"
[a'γap-i]	"amour"
['zaxar-i]	"sucre"

En prenant en considération un (ou plusieurs) son(s) précédant la désinence, on peut indiquer la place de l'accent dans la majorité des cas et souligner tous les mots faisant exception à la règle formulée.

Nous retenons la méthode présentée, ci-dessus, comme la solution idéale pour formuler les règles d'accentuation. Pour la première fois, ces dernières, ainsi que les exceptions signalées à l'intérieur de chaque règle, incluent tous les mots existants à l'intérieur d'une catégorie grammaticale du vocabulaire grec.

Les règles formulées ont été testées à partir d'un dictionnaire du grec moderne dont tous les mots, ont été stockés sur mini-ordinateur LSI 11/73 de Digital dans un fichier pour édition et traitement, (au total 45 750 mots).

Nous limitons notre champ d'investigation à deux catégories de mots accentogènes à savoir tous les substantifs (formes déclinables au nominatif singulier et formes indéclinables à l'exception de mots d'origine et de noms propres) et tous les verbes (indicatif présent, première personne du singulier) (au total 31 734 mots).

Dans notre démarche, après distinction des mots en verbes ou en substantifs, nous avons reclassé les substantifs en sous-catégories, d'après leur désinence, sans tenir compte du genre. A l'intérieur de chacune de ces dernières, on retrouve les différentes structures syllabiques finales précédant la désinence, le type d'accentuation le plus répandu (accent sur la syllabe finale, pénultième ou antépénultième) ainsi que la liste de toutes les exceptions. Parfois, à l'intérieur des cas exceptionnels, d'autres règles d'accentuation (sous-règles) sont formulées et illustrées généralement avec un exemple.

Ainsi, par exemple, 418 occurrences sur 515 existantes de substantifs se terminant par [-ða] sont accentuées sur la syllabe pénultième (ex. : [ɛfimɛ'riða] "journal"). Parmi les exceptions, 74 occurrences sur 76 se terminant par [-itiða] et tous les mots composés du substantif ['fluða] "écorce" (6 occurrences) sont accentués sur la syllabe antépénultième (ex. : [flɛ'vitiða] "phlébite", [lɛmɔ'nɔfluða] "écorce de citron"). Enfin on retrouve le mot "fée" qui présente deux formes; [nɛ'raiða] et [anɛ'raiða] . En définitive, la place de l'accent est prévisible dans 497 occurrences (418 + 73 + 5 + 1) sur 515.

La règle annoncée, ci-dessus, sera donc présentée comme suit :

[-ða] (418 occurrences sur 515)
> 497 sur 515

Exemple : [ar'kuða] "ours"

Exceptions : Accent sur la syllabe antépénultième

[-iða]

- les substantifs qui se terminent en [-itiða] (76 occur.), ex. : [flɛ'vitiða] "phlébite", [ðina'mitiða] "dynamite" sauf [ci'tiða], [ri'tiða].

- les substantifs [ayri'ɔjiða], ['vɔtriða], [ɛrasi'tɛxniða], ['ɛriða], [i'cɛtiða], ['iriða], [kali'tɛxniða], ['kɔniða] ou [kɔ'niða], [narkɔ'θɛtiða], [(a)nɛ'raiða], [ksi'lɔviða], [para'statiða], ['zmiriða], ['trɔpiða].

[-uða]

les mots composés avec le substantif ['fluða] "écorce" (6 occurrences), ex. : [lɛmɔ'nɔfluða] "écorce de citron".

L'accent frappe, en revanche, la syllabe antépénultième dans la plupart des substantifs se terminant par [-ma] (2433 occurrences sur 2445). Nous aurons ainsi :

[-ma] (2433 occurrences sur 2445)
> 2434 sur 2445

Exemples :

['prɔɣrama] "programme"
['aniɣma] "ouverture"

Exceptions :

Accent sur la syllabe pénultième

[-ama] ['a'nama] (et ['anama]), [θi'mjama] (ou [θi-'miama]), [pi'(t)zama], [rɛ'klama].
[-ima] [padɔ'mima].
[-uma] [ka'luma], [ba'ruma].
[-Cma] [ma'xatma], [plat'ɔrma].

Accent sur la syllabe finale

[-ama] [ma'ma].
[-ɛma] [sinɛ'ma].

RESULTATS

Les résultats montrent que la structure finale du mot permet la délimitation de la place de l'accent dans 80 % des cas; 25 433 occurrences sur 31 734 mots étudiés (substantifs et verbes). Avec l'élaboration de sous règles (complétant les 243 règles principales), appliquées aux cas exceptionnels, on peut atteindre le pourcentage de 89 % (28 329 occurrences). Le 11 % restant est traité sur une liste d'exceptions. Ainsi sur un très vaste lexique, nous arrivons à un système de règles qui fonctionne 9 fois sur 10.

Cette étude se veut une contribution à l'explication du fonctionnement linguistique de l'accent; elle nous semble directement exploitable :

- en linguistique appliquée : apprentissage du grec moderne, langue étrangère,

- en synthèse : exploitation automatique de textes simplifiés (sans marques accentuelles), par exemple : telex...,

- en reconnaissance automatique : récupération de l'accent à partir de la chaîne phonétique.

NOTES :

(1) Il s'agit d'un accent libre, avec toutefois une restriction : il ne remonte pas au-delà de la troisième syllabe en partant de la fin du mot, quel que soit le nombre des syllabes. Pour les problèmes d'ensemble relatifs à l'accent en grec moderne nous renvoyons à A. TSEVA (1987).

(2) Une autre démarche avec notamment des références diachroniques a été présentée par H. TONNET (1984). Cette étude consiste à formuler les règles d'accentuation grâce à des démarches différentes tels le recours à des états antérieurs de la langue, au contenu sémantique, à l'origine du mot, à l'analyse morphologique, etc. Malgré cette solution polyvalente, nous constatons qu'un ensemble de mots échappent aux règles déjà formulées.

REFERENCES BIBLIOGRAPHIQUES :

P. GARDE, L'Accent. Presses Universitaires de France, Paris, 1968. - 172 p.

A. GEORGOPAPADAKOS, Le Grand dictionnaire de la langue néo-hellénique (en grec). Malliaris - Phaidheia, Athènes, 1984. - 1184 p.

H. TONNET, Manuel d'accentuation grecque moderne (démotique). Klincksieck, Paris, 1984. - 112 p.

M. TRIANTAPHYLLIDIS, Grammaire néo-hellénique - démotique (en grec). Institut d'Etudes Néo-helléniques, Université de Thessalonique, 1978. - 448 p.

A. TSEVA, Contribution à l'étude de l'accent en grec moderne. Règles de prévisibilité et analyse instrumentale. Thèse de Doctorat, Université de Grenoble III, 1987. - 435 p.

JOURNEE D'ETUDES FRANCO-ARABES

16e Journée d'Étude sur la Parole - Société Française d'Acoustique - Hammamet, 5-9 oct 1987.

LES INVARIANTS PHONETIQUES EN ARABE EN VUE DE L'ELABORATION AUTOMATIQUE D'UN AUDIOGRAMME VOCAL POUR TOUS PUBLICS ARABOPHONES.

L. Abou Haidar*, E. Lhote*, C. Condé**, F. Lefevre*, J.P. Dupret*

* Lab. de Phonétique - 25030 Besançon Cédex
** Lab. de Math. Inf. Stat. - 25030 Besançon Cédex

This paper proposes an introduction to the automatic recognition of Arabic vocal forms, especially the relation between acoustic forms and phonological structures. First, a matrix of phonetic features in Arabic is worked out on the basis of articulatory properties. Then Correspondance Analysis and Hierarchical Classification are applied in order to extract relations between phonetic features and phonemes considered as points of anchorage for the Arab listener, whatever his own dialect.

Introduction

Le système humain de perception de la parole réagit par rapport à des formes qui, pour chaque langue et chaque individu, résultent d'une interaction entre des propriétés acoustiques, des structures phonologiques (et linguistiques) et des concepts mentaux. Alors que l'association entre le signal sonore et le concept (LAFON, 1964) [1] a retenu l'attention des chercheurs, celle qui existe entre forme acoustique et structure phonologique n'est pas encore très exploitée.

Cette corrélation présente dans la langue arabe un terrain d'étude privilégié en raison de la situation sociolinguistique qui explique le fonctionnement des dialectes par rapport à la langue MERE.

Nous avons choisi d'utiliser la complexité qui se rattache aux différentes dénominations de la langue arabe en considérant que pour tout arabophone il existe une identité de langue appelée [fusha]. Cette langue, qui est comprise par tous les arabophones, est régie par des règles uniques de syntaxe, de morphologie et de phonologie que l'on retrouve dans les réalisations dialectales. Considérée comme standardisée et comme modèle, elle est apprise à l'école et utilisée dans les médias. Mais comme c'est le cas pour toute langue parlée sur des territoires géographiquement très grands, la langue orale "standard" est très influencée par les dialectes. Et ceci explique pourquoi des réalisations acoustiques qui peuvent être très différentes peuvent en même temps être perçues comme la même forme par des auditeurs arabophones.

Nous considérons quant à nous que la référence commune est un comportement d'auditeur arabophone qui, implicitement, ancre son écoute en arabe sur des règles de base de la langue [fusha], qu'il s'agisse de la langue MERE ou d'un dialecte. Nous supposons donc l'existence de repères sur le plan perceptuel, permettant à des sujets arabophones d'origine dialectale différente, de se reconnaître entre eux et de se comprendre par l'intermédiaire de leur langue orale.

Nous nous intéressons ici aux éléments acoustiques communs aux dialectes et à cette langue [fusha] qui fait l'objet d'un consensus chez tous les locuteurs arabophones des différents pays arabes.

Nous nous proposons dans un premier temps de mettre au point, grâce à une approche multidimensionnelle, un test audiométrique pour tout public arabophone en vue d'élaborer ensuite un audiogramme vocal. Nous considérons en effet que si une altération acoustique affecte l'audition d'un locuteur arabe, des traces de la déformation afférente seront décelables dans la perception, par cet individu, des formes acoustiques des structures de base de la langue.

Le modèle d'analyse

Notre objectif est de réussir à prévoir de façon automatique les formes qui présentent un écart par rapport à un modèle préalablement défini. Pour l'élaboration de l'étalon de référence, nous faisons appel au Test de Mots sans signification de DUPRET (1980) [2]. Le choix de logatomes, unités non significatives, de structure CVC et CVCV, est destiné à tester chez l'auditeur la qualité de perception linguistique qui conduit le sujet à établir un lien entre une forme acoustique et une signification. Dans le cas d'un logatome, l'auditeur établit une corrélation entre une forme acoustique et une structure phonologique, sans utiliser en principe la suppléance mentale.

A l'aide du Test de Confrontation Indiciaire de LEFEVRE (1985) [3], on établit ensuite une comparaison entre un stimulus modèle et son imitation par le sujet testé.

Afin d'établir des listes de logatomes correctement équilibrées, dont les éléments phonématiques soient représentatifs des catégories de phonèmes

auxquelles ils appartiennent, nous avons dans un premier temps constitué une matrice des traits distinctifs du système consonantique de l'arabe et traité cette matrice par l'intermédiaire de l'Analyse Factorielle des Correspondances. C'est le fruit de cette étape du travail que nous présentons ici.

Matrice des traits distinctifs

Nous avons retenu les 28 phonèmes consonantiques déjà répertoriés par ailleurs par SIBAWAYHI en l'an 180 de l'Hégire. Ces consonnes sont les constituants de base de la langue [fuṣḥa] pour tout arabophone de toutes origines dialectales :

[t̪] [k] [ʔ] [b] [d] [m] [n] [t̪] [d̪] [q]
[f] [θ] [s] [ʃ] [X] [ħ] [ð̪] [z] [ʒ] [ð]
[ʕ] [h] [l] [ɾ] [s̪] [ð̪] [w] [j]

La version standardisée et communément admise quant à la caractérisation des phonèmes consonantiques de l'arabe tient compte, mises à part la gémination et l'emphase, des traits articulatoires suivants :

- sourdité / sonorité
- occlusion / constriction
- oralité / nasalité
- lieux d'articulation : bilabial, labiodental, apicodental, alvéodental, apicoalvéolaire, prédorsoalvéolaire, alvéopalatal, palatal, vélaire, uvulaire, pharyngal, glottal.

Pour ce qui est de l'identité des consonnes auxquelles nous attribuons le trait d'emphase, nous précisons que cette caractéristique reflète pour nous avant tout un point de vue subjectif, c'est-à-dire la situation de l'auditeur et son mode de perception d'un segment caractérisé par l'emphase. Nous pensons que c'est avant tout à la perception du timbre de l'unité produite que l'on peut indiquer si telle consonne est sous le trait d'emphase ou non. Nous sommes par ailleurs convaincus que l'emphase existe au moins à deux niveaux dans la chaîne sonore:

- le niveau "segmental" dans lequel nous distinguons les consonnes emphatiques [t̪], [d̪], [s̪], [ð̪], [q]. Pour les consonnes antérieures, le trait d'emphase se traduit par la présence d'un second lieu d'articulation, vélaire, qui contribue fortement à donner la sensation d'emphase. L'ajout du caractère vélaire ne s'impose pas dans le cas de [q] qui le comporte intrinsèquement.

- le deuxième niveau dans la chaîne sonore dont les segments pourraient être affectés par le caractère d'emphase se situe pour nous au-delà du phonème : la combinaison de certains phonèmes peut impliquer que tel segment semble porteur d'un facteur d'emphase sans que les sons eux-mêmes soient emphatiques.

Dans la première étape de notre tavail nous n'avons pas tenu compte du trait de gémination qui sera introduit dans un temps ultérieur pour la raison suivante : la gémination n'est pas le dédoublement d'une consonne ou la production de deux consonnes identiques successives. Quel que soit le mode articulatoire relatif à la consonne marquée par ce trait, le tenue de la constriction ou de l'occlusion dure plus longtemps et est plus intense que pour une consonne simple. Le découpage syllabique a toujours lieu cependant entre le segment d'attaque et celui de libération : en arabe une consonne géminée ne se trouve jamais en position initiale dans le mot mais toujours en position médiane ou finale, par conséquent le segment correspondant à l'attaque ne se trouve jamais en position initiale dans la syllabe. Pour mettre en valeur le trait de gémination il est donc préférable de l'étudier en situation intervocalique. C'est pourquoi nous introduisons dans le corpus une structure syllabique CVCV en même temps que le trait de gémination. La figure 1 représente la matrice de traits articulatoires de l'arabe utilisée pour la constitution du corpus.

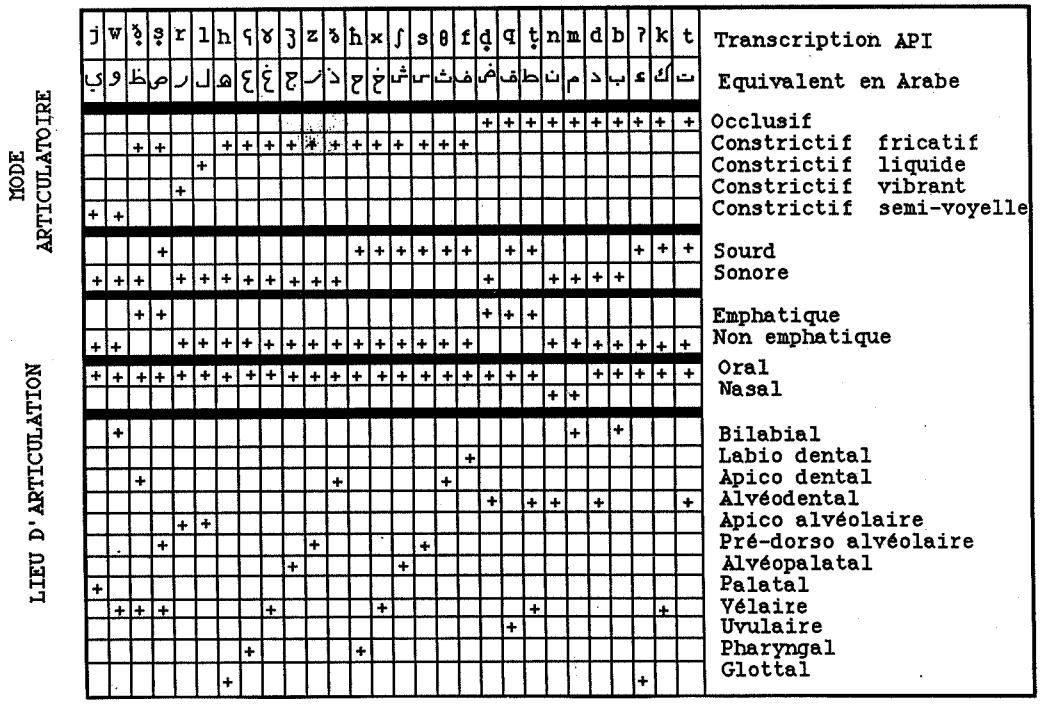


Figure 1
Matrice des traits distinctifs des consonnes de l'arabe

Présentation des méthodes de traitement

Après avoir mis au point le tableau de classification des phonèmes et des traits distinctifs de la langue arabe, nous avons cherché une méthode objective de traitement qui attribue à chacun des traits un poids égal et qui permette d'apprécier l'importance relative des éléments entre eux. Cette méthode est l'Analyse Factorielle des Correspondances [4] qui répond à ces exigences et qui offre en outre l'avantage de visualiser les résultats d'une analyse multidimension- nelle en représentant les forces d'attraction des éléments les uns par rapport aux autres. Il nous est apparu important en effet de prendre de la distance par rapport aux représentations phonétiques habituelles.

Il s'agit de mesurer à l'intérieur d'un espace multidimensionnel, des distances entre des lignes (les phonèmes) et des colonnes (les traits); la métrique utilisée est celle du Chi-Deux.

Les résultats se présentent sous forme de graphes obtenus par projection de l'hyperespace sur un espace à 2 dimensions (plan). L'avantage de la méthode - outre le fait qu'elle n'impose que peu de conditions au tableau traité - réside dans sa capacité à extraire et à 'donner à voir' la structure du tableau de départ et ce, sous une forme facilement accessible (des proximités de points sur un plan).

Cette structure apparaît sous la forme de groupes de points plus ou moins éloignés les uns des autres et permet la constitution d'une typologie des phonèmes.

La méthode de Classification a été utilisée ensuite pour affiner les résultats de l'AFC.

Cette approche des données phonétiques donne à chaque composante une importance égale. Par l'association des traits et des phonèmes qui ont le plus d'éléments en commun, elle met en évidence le champ d'attraction de chacun et permet d'extraire des traits plus forts que d'autres. Nous en arrivons donc à dégager objectivement des relations de force au sein d'un système phonémique.

Traitement par AFC

Le traitement par la méthode d'Analyse Factorielle des Correspondances a été appliqué uniquement sur la matrice des traits distinctifs des consonnes de l'arabe et non sur celle des voyelles en raison de leur faible degré de complexité. Le traitement des traits distinctifs a donné la répartition schématisée dans la figure 2.

Les deux plus grandes oppositions entre traits ayant contribué le plus fortement à la construction du graphe sont:

1- l'opposition (apicoalvéolaire, liquide, vibrante) vs (emphatique).

2- l'opposition (occlusive, alvéodentale, nasale) vs (constrictive fricative).

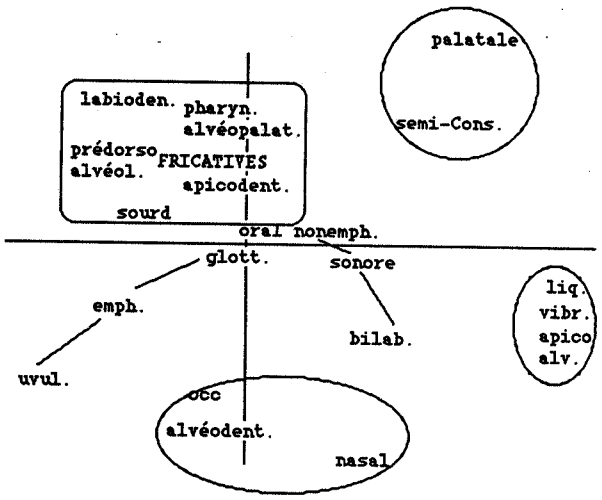


Figure 2 - Graphe des traits distinctifs

Les traits dont la fréquence d'occurrence est la plus importante dans la matrice se trouvent à l'intersection des deux axes : (oral- nasal - non emphatique). En revanche les traits les plus excentrés correspondent à ceux qui sont caractéristiques du plus petit nombre de phonèmes.

Cinq zones se démarquent sur le graphe. Chacune d'elles est caractérisée par la présence de traits dont le comportement dans la matrice est semblable dans la mesure où chaque groupement représente une catégorie de phonèmes.

La répartition des phonèmes sur le graphe (Figure 3) est caractérisée par les deux plus grandes oppositions :

1- ([l], [r]) vs ([ṭ], [q])

2- ([m], [n]) vs ([j])

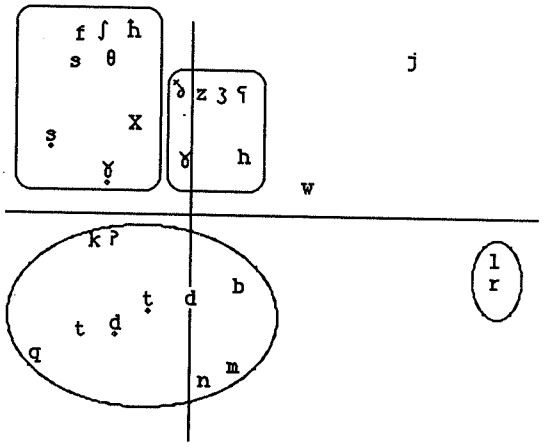


Figure 3 - Graphe des phonèmes

Si nous faisons abstraction des consonnes liquide et vibrante, nous remarquons que l'axe 1 sépare très nettement les occlusives des constrictives. D'autre part, l'axe 2 met en relief la distinction entre consonnes sourdes et sonores.

Trois zones se distinguent sur le graphe, à l'intérieur desquelles les sous-groupes de phonèmes se démarquent d'une manière très pertinente :

- la zone "d'influence" des consonnes occlusives est divisée en trois sous-groupes à l'intérieur desquels la distinction d'oralité et de nasalité est mise en valeur.

- la répartition des constrictives fricatives met en relief la distinction se rapportant au caractère de sourdité.

Classification des éléments

A partir des résultats obtenus par le traitement de l'AFC, la méthode de Classification procède de la manière suivante : elle regroupe successivement les éléments du tableau, en évaluant la distance qui les sépare et en partant des deux éléments jugés les plus proches du système, ceux-là même qui constitueront la première classe. C'est ainsi que nous avons pu observer, au fur et à mesure du dépouillement des résultats du tableau descriptif de l'arbre de classification, les étapes principales mettant en lumière le poids relatif à chaque élément dans l'interaction que les traits et phonèmes entretiennent les uns avec les autres. Les différentes classes obtenues nous révèlent les regroupements de phonèmes suivants :

1- [l] [r]
2- [j]
3- [f] [θ] [s] [ʃ] [X] [ħ] [ð] [z] [ʒ] [ɣ] [h] [k] [ʔ] [s̪] [ð̪] [ʕ]
4- [t̪] [d̪] [t] [q]
5- [m] [n] [b] [d] [w]

Interprétation des résultats

La visualisation des résultats de l'AFC sur un graphe d'une part, l'analyse des différentes étapes de classification d'autre part, offrent une nouvelle approche de certaines relations entretenues par les éléments constitutifs de la matrice des traits distinctifs. Il en ressort les remarques suivantes :

- le statut particulier des consonnes [l] et [r] d'une part, [j] d'autre part, se retrouve aussi bien dans le graphe que dans la classification car ces consonnes se démarquent de l'ensemble.

- la distinction sourdes / sonores est mise en relief dans les deux cas : nous avons pu observer à une certaine étape de la classification que les consonnes constrictives sourdes et sonores constituaient deux catégories bien distinctes.

- la distinction entre modes articulatoires est mise en relief, mis à part le cas des deux occlusives [k] et [ʔ]: l'une des classes est constituée des occlusives sonores ainsi que de la semi-consonne [w], l'autre est constituée des occlusives sourdes emphatiques et de la sonore [d̪]. La répartition des consonnes emphatiques sur le graphe ainsi que dans les différentes classes témoigne de la particularité et du poids de l'emphase par rapport à celui de la sourdité notamment.

Perspectives d'avenir

Parmi les résultats obtenus, certains sont plus ou moins attendus par les phonéticiens, d'autres sont particulièrement inattendus : certaines occlusives se sont retrouvées dans une classe de constrictives ; la semi-voyelle est associée aux occlusives sonores non emphatiques… Le dépouillement des résultats du traitement informatique suggère une reformulation de la matrice de traits.

L'analyse détaillée des différentes étapes de la classification nous a montré en effet deux insuffisances de la matrice de traits articulatoires :

- l'absence de critères acoustiques n'a pas permis le regroupement en une même classe de [r], [l] et [j].

- la description incomplète du trait emphatique, en particulier au plan perceptif, est certainement à l'origine de la présence de deux occlusives dans la classe des constrictives. Le glissement des deux occlusives s'explique, selon nous, par la force d'attraction du lieu d'attraction vélaire (en arabe).

Les insuffisances qui sont apparues dans l'interprétation des résultats de l'AFC nous ont montré le caractère indispensable de l'analyse acoustique et de l'analyse perceptuelle dans l'appréciation par l'auditeur des relations hiérarchisées entre unités phonologiques. La prise en compte des unités perceptuelles est en partie destinée à établir des règles distributionnelles dans l'élaboration des unités CVC et CVCV lors de l'élaboration de listes de test homogènes et équilibrées.

Conclusion

En conclusion, nous pouvons d'ores et déjà augurer de l'intérêt phonétique de ce type d'approche qui, par son caractère multidimensionnel, permet de faire progresser la reconnaissance de formes vocales arabes dans la relation spécifique entre forme acoustique et structure phonologique et, du même coup, la connaissance de la structure phonétique de la langue arabe.

Références

[1] LAFON Jean-Claude (1964) :
Le Test Phonétique et la Mesure de l'Audition
Paris : Centrex.

[2] DUPRET Jean-Pierre (1980) :
Test de Mots sans Signification
Mémoire du Collège National d'Audioprothèse.

[3] LEFEVRE Frank (1985) :
Une Méthode d'Analyse Auditive des Confusions Phonétiques : Confrontation Indiciaire
Thèse de Doctorat, Besançon.

[4] BENZECRI J.-P. et al. (1973) :
Correspondances, volume III
Paris : Dunod.

HIERARCHIE DE SONORITE ET SEGMENTATION SYLLABIQUE DANS LE PARLER ARABE MAROCAIN

BENKIRANE Thami
CAVÉ Christian

U.A. 261 CNRS, INSTITUT DE PHONETIQUE
UNIVERSITE DE PROVENCE - 13621 AIX (FRANCE)

Cette étude a pour but d'examiner l'effet éventuel de la hiérarchie de sonorité (HS) sur le temps de détection de cibles syllabiques ou non dans le parler arabe marocain.

Nous avons retenu une liste de 8 paires de mots dissyllabiques de type C1V1C2V2 et C1V1C2C3V2 pour lesquels la suite initiale CVC est la même pour chaque paire. La consonne C2 est soit + sonant , soit - sonant . Les consonnes + sonant occupent un rang plus élevée dans la HS que les consonnes - sonant. Les séquences cibles sont de type CV ou CVC et correspondent à la partie initiale de chaque mot.

Le temps de réaction (TR) au cibles CV est toujours plus court que le TR aux cibles CVC. La hiérarchie de sonorité exerce un effet significatif sur le TR. Celui-ci est plus court quand la consonne C2 est + sonant que lorsqu'elle est - sonant . Ce résultat, ainsi que l'absence d'interaction entre la sonorité et le type de cible est en contradiction avec l'idée que la HS jouerait un rôle primordial dans la position de la frontière syllabique.

INTRODUCTION

1.Cadre conceptuel.

La syllabe, sa définition, sa nature et ses limites; sujet classique de débats parmi les linguistes; a connu récemment un regain d'intêret dû à des travaux qui l'ont abordée du point de vue du décodage de la parole naturelle. En effet, à la suite des travaux de Savin et Bever (1970) sur le statut perceptif du phonéme, Mehler & al (1981) ont montré que la détection de cibles constituées de deux ou trois phonèmes était plus rapide quand il y avait coincidence entre la cible et la syllabe initiale du mot stimulus. Par exemple, lorsque la cible à détecter est /CAR/, le temps de réaction (TR) est plus court si le mot stimulus est CARTON que s'il s'agit de CAROTTE. Toutefois il existe des cas où la place de la frontière syllabique peut prêter à discussion, par exemple, en français, INSTABLE ou ASPERGE.

En arabe marocain, il existe des mots de structure CVCCV pouvant donner lieu, à priori, à deux segmentations syllabiques :

1 CV/CCV (C = consonne)
2 CVC/CV (V = voyelle)

Les règles de syllabation de l'arabe littéral dissocient systématiquement les deux C intervocaliques (CVC/CV) car la phonologie de l'arabe littéral interdit une séquence biconsonnantique dans une même syllabe. En ce qui concerne le parler arabe du Maroc (PAM), Benkirane (1982) au terme d'une étude acoustique, phonologique et phonotactique, défend l'idée que la structure syllabique de ce type de mot est CV/CCV.

Lors d'une expérience précédente, nous avons utilisé la méthode du temps de réaction pour tenter de valider, d'un point de vue perceptif, cette analyse de la structure syllabique. Pour cela, nous avions effectué une étude sur la détection de cibles en arabe marocain. Les mots stimulus étaient 8 paires de mots dissylabiques de type CVCV et CVCCV, pour lesquels la suite CVC étaient la même pour chaque paire. Les cibles étaient les séquences CV et CVC correspondant à chaque mot. Nos résultats montraient que le TR aux cibles CV était identique quelle que soit la structure du mot stimulus et qu'il était toujours plus court que le TR aux cibles CVC.

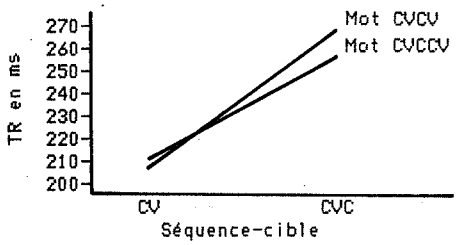


Fig. n°1 Temps de réaction en fonction des cibles et du type de mot.

Ces résultats allaient dans le sens de nos hypothèses et semblaient confirmer une structuration syllabique identique pour les deux types de mots, c-à-d. CV/CV et CV/CCV. Toutefois, l'impossibilité, liée aux caractéristiques phonologiques du PAM, de présenter des mots CVC/CV avec les cibles correspondantes peut laisser un doute quant à l'interprétation des résultats. Afin de surmonter, dans la mesure du possible cette difficulté, nous avons réalisé une deuxième expérience, fondée sur le même principe, mais en contrôlant, cette fois, la hiérarchie de sonorité dans le groupe CC intervocalique afin de nous placer dans la situation la plus favorable à une syllabation de type CVC/CV et donc dans une situation défavorable à notre hypothèse.

2. Hiérarchie de sonorité.

Le phonéticien danois Jespersen (1904) est probablement le premier qui ait fondé la répartition des élements phoniques qui composent la syllabe sur leur degré de sonorité ou d'audibilité. Il s'agit d'une hiérarchisation qui distingue 8 échelons de sonorité depuis les occlusives (minimun de sonorité), jusqu'aux voyelles les plus ouvertes (maximun de sonorité). Cette échelle est la suivante :

Min. de sonorité → Max. de sonorité

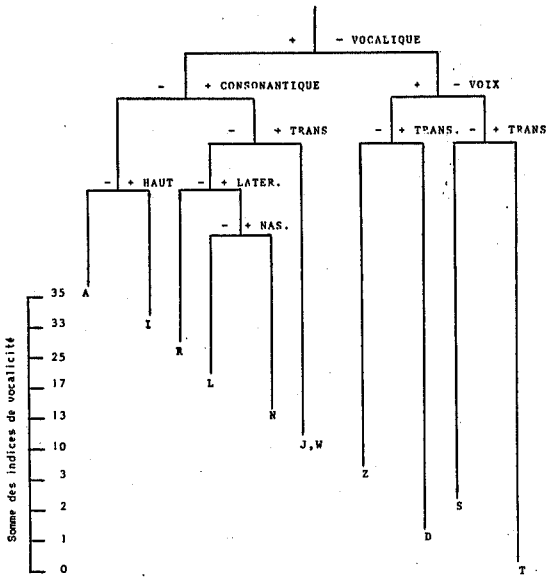
1	2	3	4	5	6	7	8
cons. - voisé (t,f)	occlus. + voisé (b,d,g)	fricat. + voisé (v,z)	later. et nas. (l,n)	vibr. (r)	voyel. fermée (i,u)	voyel. mi fermée (e,o)	voyel. ouv. (a)

Selon Jespersen, dans une syllabe, les consonnes se regroupent autour d'un pic de sonorité en respectant la hiérarchie. La syllabe est alors définie comme la distance entre deux minima de sonorité (par ex. par/tie et pa/trie). Le classement de Jespersen est, de toute évidence, fondé sur le degré d'ouverture des sons. (La théorie articulatoire de l'ouverture sera reprise et developpée par Saussure, 1916). Bien que la notion de sonorité demeure mal définie, il est possible d'y voir un rapport avec la sonie. Dans cette perspective, on peut établir un rapprochement avec la conception acoustique de Zwirner (Cité par Malmberg,1974, p.185) selon laquelle il y aurait autant de syllabes que de maxima d'intensité.

Hàla (1961) considère la sonorité (qualité du phonème par rapport à l'efficacité de la voix laryngienne) comme l'élément fondamental de la syllabe. La sonorité a pour corollaire le degré d'aperture ou de stricture avec lequel sont réalisés les sons. Les segments vocaliques sont dotés d'une grande perceptiblité et pour cette raison, ils sont, parce que les plus sonores, appelés à occuper le centre de la syllabe. Hàla défend l'idée d'un isomorphisme entre les classifications articulatoires et acoustiques des sons. Ainsi, dans une syllabe, les phonèmes sont groupés autour des noyaux vocaliques. Lors de la phase initiale jusqu'au point culminant (la voyelle), on observe un crescendo de l'intensité et un élargissement graduel de l'aperture du minimun vers le maximun. Lors de la phase finale, après le point culminant, on observe un decrescendo de l'intensité et de l'aperture du maximun vers le minimun. Ce phénomène est appelé principe de la syllabe optimale ou encore chez Selkirk (1984 p.116) "sonority sequencing generalization" : "In any syllable, there is a segment constituting a sonority peak that is preceded and/or followed by a sequence of segments with progressively decreasing sonority values".

Pour Hooper (1972,1976), Vennemann (1972), Malmberg (1963, 1974), Harris (1982) et Selkirk (1984) la hiérarchie de sonorité permet de rendre compte des containtes phonotactiques observées dans les différentes langues et donc de délimiter les syllabes. La frontière syllabique est déterminée en fonction du degré de sonorité des éléments phoniques contigus. Selkirk (1984) propose de rendre compte de la structuration interne des syllabes de l'anglais au moyen de "l'indice de sonorité" de chaque segment phonique, plutôt qu'en utilisant les traits majeurs +/- consonantique, +/- sonant et +/- syllabique.

Tout récemment, Rossi (1987) propose pour sortir la notion de sonorité du flou qui l'entoure de la remplacer par le concept de vocalicité qu'il définit au moyen d'indices acoustiques. La somme de ces indices de vocalicité permet de classer et de hiérarchiser les différents segments phoniques d'une langue:



Classification et hiérarchisation des sons en fonction des indices de vocalicité (d'après Rossi 1987)

Par ailleurs ces indices permettent d'établir des règles qui rendent compte de la structure de la syllabe et de ses frontières. Par exemple, une des règles établies pour le français, insère une frontière syllabique devant un minimun de vocalicité (par ex. pa/trie, ar/pen/ter). Ainsi les degrés de vocalicité ou de sonorité rendent possible, d'une façon heuristique, la détermination de la limite syllabique dans une séquence de segments consonantiques située entre deux voyelles. C'est cette possibilité que nous avons exploitée dans l'expérience que nous présentons ici, afin de renforcer une syllabation de type CVC/CV en PAM, en utilisant des mots pour lesquels l'indice de sonorité (I S) de la consonne C2 est minimum. Par exemple, les deux mots marocains suivants

mots	b	a	r	d	a		b	a	d	r	a
I.S.	1	35	25	1	35		1	35	1	25	35

devraient, conformément à la hiérarchie de sonorité, être syllabés en insérant une frontière syllabique devant le minimum de sonorité :

bar/da ba/dra.

C'est en tenant compte de ces critères de hiérarchie de sonorité que nous avons constitué le corpus expérimental.

EXPERIMENTATION

1. Choix du corpus.

Nous avons choisi deux séries de mots dissyllabiques de type C1V1C2V2 et C1V1C2C3V2. La structure syllabique des mots CVCV est clairement de type CV/CV. Pour les mots CVCCV, nous avons contrôlé la hiérarchie de sonorité du groupe consonantique intervocalique, selon les principes exposés ci-dessus. Ce corpus de 16 mots stimulus est complété par 16 mots distracteurs (V annexe n°1).

Tous les items, mots stimulus et mots distracteurs, commencent par une consonne voisée et sont accentués sur la première syllabe. Les mots ont été enregistrés par un locuteur dont la première langue est l'arabe marocain.

2. Protocole expérimental.

A partir des 32 mots retenus, 16 mots expérimentaux et 16 mots distracteurs, on a procédé au tirage aléatoire de 12 ordres de présentation des items qui restent les mêmes pour tous les sujets. De plus, l'ordre de présentation de ces 12 ordres a lui même été déterminé de façon aléatoire pour chaque sujet. Ces précautions ont été prise de façon à éviter les effets de liste et les effets de blocs.

L'ensemble de l'expérience est piloté, en temps réel, par un micro ordinateur qui contrôle le magnétophone présentant les stimulus acoustiques, l'apparition des cibles sur un écran vidéo et enregistre les temps de réaction.

Un bip sonore précédait d' une seconde le début de chaque mot et un intervalle de quatre secondes séparait les mots. Les mots étaient présentés à un niveau moyen de 60 dB.

Quinze sujets de langue maternelle arabe marocain ont passé l'expérience. Leur tâche consistait à appuyer le plus vite possible sur un bouton réponse si le mot entendu commençait par la cible présentée préalablement. Dans le cas contraire ils n'avaient pas à répondre. Bien entendu, la consigne ne donnait aucune information sur les buts de l'expérience ni sur la structure des items expérimentaux.

RESULTATS

Avant de procéder à l'analyse des résultats on a corrigé les fichiers de réponses de chaque sujet en éliminant les TR inférieurs à 100 ms (anticipation) et les TR supérieurs à 1200 ms (réponse tardive) selon la pratique habituelle dans ce genre d'expérience. Ceci nous a conduit à éliminer 1,32% des réponses. L'ensemble des valeurs des temps de réaction (2880 réponses) a été soumis à une analyse de variance, en prenant les sujets comme facteur aléatoire.

Le TR moyen aux cibles CV est de 453.2 ms alors que le TR moyen aux cibles CVC est de 555.5 ms. La différence est nettement significative : $F(1,14) = 11,7$ $p<.004$. Ainsi, quel que soit le mot stimulus présenté, les cibles CV sont détectées beaucoup plus rapidement que les cibles CVC. Le TR au cibles CV n'est que très faiblement influencé par la structure du mot stimulus. En effet le TR est de 458 ms pour les mots CVCV et de 449 ms pour les mots CVCCV. Par contre le TR aux cibles CVC est influencé par la structure du mot stimulus. Il est de 579 ms en contexte CVCV et de 532 ms en contexte CVCCV. Ceci est confirmé par le fait que le type de mot a un effet significatif sur le TR. Indépendamment du type de cibles, le TR est plus court pour les mots CVCCV (490.3 ms) que pour les mots CVCV (518.4 ms) : $F(1,14) = 5.60$ $p< .03$. Enfin l'interaction CT (interaction entre le type de cible et le type de mot) qui teste en fait l'effet syllabique que nous recherchions, n'est pas significative; $F(1,14) = 2.85$ $p<.11$.

La figure 2 résume l'ensemble de ces résultats.

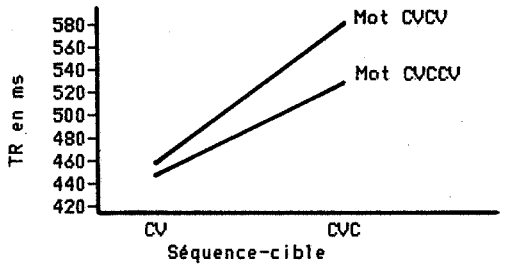


Fig.n°2 Temps de réaction en fonction des cibles et du type de mot.

Enfin, la hiérarchie de sonorité a un effet sur le temps de réaction. Celui-ci est légèrement plus court quand la consonne C2 est + sonant (495.8 ms) que lorsque C2 est - sonant (512.8 ms). La différence, bien que faible en valeur absolue (17 ms) est significative ($F(1,14) = 15,6$ $p<.001$). En fait, ce résultat est plutôt en contradiction avec l'idée que la hiérarchie de sonorité jouerait un rôle déterminant dans la place de la frontière syllabique. En effet si le passage par un minimum de sonorité déterminait la place de la frontière syllabique, on s'attendrait à un TR plus court dans le cas de C2 - sonant. Ceci est encore confirmé par le fait que l'interaction HC (sonorité et type de cible) est non significative ($F<1$).

DISCUSSION CONCLUSION

Pour cette expérience le corpus a été choisi de façon à favoriser au maximun l'apparition d'une syllabation de type CVC en arabe marocain. Les résultats obtenus ne sont pas en accord avec ce type de syllabation.

Comment interpréter l'ensemble de nos données? En fait, plusieurs interprétations semblent possibles.

La première, si l'on prend en compte l'absence de significativité de l'interaction CT, est que ces résultats confirment ceux de notre première expérience et sont en accord avec l'hypothèse d'une syllabation ouverte dans le PAM. Toutefois, comme nous l'avons fait remarquer la situation ne sera jamais conclusive puisque, si cette hypothèse est exacte, il est et il demeurera impossible de présenter des mots dont la structure syllabique corresponde à une cible CVC.

La deuxième interprétation, si l'on considère l'effet de la hiérarchie de sonorité et l'absence d'interaction entre la sonorité et le type de cible, est que la hiérarchie de sonorité ne serait pas un facteur décisif pour déterminer la position d'une frontière syllabique. L'argument phonotactique, en l'occurence le fait qu'une séquence C1C2 (où C1 est + sonant et C2 - sonant) soit permise à l'initiale d'un mot et donc d'une syllabe semblerait plus pertinent. Par exemple, dans le PAM les mots ltam, jtim, rtila, nta, lbas etc...commencent par ce type de séquence, ce qui semble en contradiction avec le principe de syllabe optimale. Cette apparente liberté phonotactique s'explique par l'histoire de la langue : la formation des groupes consonantiques à l'initiale est due à la perte des noyaux syllabiques brefs (par exemple : litaam > ltam; jatiim > jtim).

Enfin, on ne peut pas exclure que la méthode du temps de réaction ne soit pas un paradigme expérimental pertinent pour mettre en évidence un effet syllabique, lorsque la position de la frontière syllabique n'est pas dépourvue de toute ambiguité.

Le débat reste ouvert.

BIBLIOGRAPHIE

BENKIRANE, T. (1982) Etude phonétique et fonction de la syllabe en arabe marocain. Thèse de 3e cycle, Université de Provence, 1982, manuscrit non publié.

BENKIRANE, T.; CAVE, C. (1984) Segmentation syllabique en arabe marocain : étude expérimentale par la méthode du temps de réaction. TIPA 9, 85-99

HALA, B. (1961) La syllabe, sa nature, son origine et ses transformations. ORBIS, 10, pp.69-143

HARRIS, J.W. (1982) Spanish syllable structure and stress : a non-linear analysis. MIT Press, Cambridge, Mass

HOOPER, J. (1972) The syllable in phonology theory. Langage, 48, 3, 524-540

(1976) An introduction to natural generative phonology. Academic Press, New York.

JESPERSEN, O. (1904) Lehrbuch der Phonetic. Leipzig

MALMBERG, B. (1963) Phonetics. Dover publications, New York.

(1974) Manuel de phonétique générale. Editions Picard, Paris.

MEHLER, J.; DOMERGUES, J-Y.; FRAUENFELDER, U.; SEGUI, J. (1981) The syllable's role in speech segmentation. J. Verb. Learning Verb. Behav. 20, 398-305

ROSSI, M. (1987) Cours de DEA. Institut de Phonétique d'Aix

SAUSSURE, F de. (1916) Cours de linguistique générale. Paris

SAVIN, H.B.; BEVER, T.G. (1970) The non-perceptual reality of the phoneme. J. Verb. Learning Verb. Behav. 9, 295-302

SELKIRK, E. (1984) On the major class features and syllable sound structure. In Language sound structure : Studies presented to Morris Halle. ARONOFF M.& OEHRLE R.T. (Eds) MIT Press, Cambridge Mass

VENNEMANN, T. (1972) On the theory of syllabic phonology. Linguistische Berichte, 18, 1-18

VENNEMANN, T.; MURRAY, R.W. Sound changes and syllable structure in Germanic phonology. Language, 59, 3, 514-528

Annexe 1

Mots stimulus	Mots distracteurs
afu	Rula
aji	uma
afja	wina
ajfa	bunja
badi	dima
bari	lawja
badra	majla
barda	mala
zadu	mazja
zani	najma
zadna	wazda
zanda	zajba
Raba	zibu
Raja	zriba
Rabja	zuza
Rajba	zwina

CONVENTIONS LINGUISTIQUES ET "NATUREL" ACOUSTICO-PHYSIOLOGIQUE: PEUT-ON PARLER DE REGLES DE COARTICULATION?

Jean-François P.BONNOT

Laboratoire de Phonétique et Centre de Recherches en Informatique
et Combinatoire, UA CNRS 1099
Université du Maine, BP 535, F-72017 Le Mans Cedex

ABSTRACT

The purpose of this article is to discuss some problems related to the notions of physiological "naturalness" and of linguistic conventional properties.

1. MECANISMES BIOLOGIQUES ET COMPORTEMENT

1.1. Le terme "naturel" est employé avec des guillemets dans le titre de notre communication; de plus, deux qualificatifs viennent en limiter le sens. On ne saurait toutefois nier que les problèmes que nous allons traiter dans une perspective phonétique et phonologique s'inscrivent dans un cadre beaucoup plus vaste. On ne peut évoquer la notion de naturel physiologique sans faire surgir simultanément celle d'universal (Hagège |1|). Comme le souligne Longacre |2|, la recherche est des plus difficiles, car nous sommes dans l'impossibilité de nous situer en un lieu nous permettant d'observer en toute indépendance d'esprit les rapports unissant la "réalité" à la langue: "every fact which we have, even scientific fact, consists essentially of an observation plus discourse"(p.318). Vieux débat, dont nous n'avons ici qu'un avatar. Monod ne déclarait-il pas que "c'est le langage qui aurait créé l'Homme, plutôt que l'Homme le langage". De même, pour le phénoménologue husserlien Pos |3|; "le linguiste est linguiste grâce au fait qu'il est un sujet parlant, et non malgré ce fait (...)" On pourrait voir là une argumentation triviale, marquée au coin du simple bon sens. C'est pourtant à ce niveau que s'articule la problématique. Lors des "Conférences Whidden", Chomsky |4| suggérait que l'état cognitif initial de l'être humain (i.e. avant l'apprentissage) comporte une représentation de la grammaire universelle (GU). Il ajoutait qu'en général, on ne saurait se fonder sur des travaux portant sur la mémoire ou le comportement pour falsifier des hypothèses ayant trait à la nature de GU. Il admettait toutefois que ces données ne doivent pas être écartées a priori (p.51).

1.2. Il est donc clair qu'une telle analyse n'implique pas le rejet des données biologiques. Elle oblige néanmoins à admettre, comme le fait par exemple Mehler dans un récent et peu amène compte-rendu du beau livre de Lieberman, The Biology and Evolution of Language, que la grammaire fait partie des "dons naturels" de l'Homme, bien plus qu'elle n'est le résultat d'un apprentissage |5|. D'autre part, pour Mehler, considérer que la faculté de langage constitue l'aboutissement d'un processus évolutionniste fondé sur la sélection naturelle serait à mettre en parallèle avec la "philosophie" du bon Dr Pangloss: si nous portons des lunettes, c'est que notre nez a été conçu dans ce but... Ce point de vue n'a rien de nouveau; Loucel |6| rapporte que le grammairien arabe Ibn Faris (m.1004) avait consacré, voici près d'un millénaire, une longue réflexion au thème suivant:"la langue arabe est-elle Fixation révélée ou Convention?" (La question était d'ailleurs largement rhétorique, dans la mesure où, à l'époque, la solution la plus "raisonnable" était celle de la Foi.) Chomsky soutient qu'il n'existe pas de système comparable à GU dans les organismes non humains, et qu'il est par conséquent assez peu intéressant de mettre en parallèle l'acquisition des comportements symboliques par les animaux et certaines caractéristiques du langage humain: "en poussant jusqu'au bout ce raisonnement absurde |écrit Chomsky|, on pourrait même soutenir que la distinction entre le saut et le vol est arbitraire et qu'elle n'est qu'une question de degré. Les gens peuvent réellement voler, mais seulement moins bien. Dans le cas du langage, des propositions de ce type ne (...) paraissent pas avoir plus de force ou de sens"(p.55). Lieberman |7| fait valoir qu'il n'a jamais contesté que la faculté de langage trouve son origine dans des mécanismes innés, génétiquement transmis. Mais, pour Lieberman, "some of these mechanisms, like the Hebbian distributed neural network that probably is the basis of the linguistic lexicon, are clearly present in living pongids. Studies of the lexical ability of many chimpanzees and one gorilla, for example, demonstrate that they can use and acquire words. In contrast, these animals lack both the anatomical mechanisms and neural mechanisms that underlie human speech (...) However, homologous neural mechanisms exist in these and lower species."

1.3. Dès lors, les enjeux apparaissent clairement: les anatomistes comme Laitman |8| ou Wind |9| et les biolinguistes comme Lieberman |10| insistent sur le fait qu'il existe une constante interaction entre tous les éléments du comportement humain et animal et les mécanismes biologiques qui les sous-tendent: pour Wind, chez les mammifères, le système respiratoire et laryngé permet des vocalisations variées du point de vue de l'amplitude et de la mélodie; l'ingestion très rapide de nourriture devant être digérée rapidement a entraîné des adaptations de l'oropharynx et des muscles faciaux; la succion a dû engendrer également des évolutions fondamentales. Les linguistes générativistes insistent quant à eux sur le fait qu'il n'existe pas de prédisposition génétique de l'enfant favorisant l'apprentissage d'une langue "X" plutôt que d'une langue "Y"; cependant, le développement de la grammaire a lieu "dans un organisme possédant des contraintes initiales fixes délimitant la classe des grammaires possibles"(Chomsky et Halle |11|). Même si les contraintes s'appliquant à GU sont très générales et que les universaux de forme, comme le soutient Hagège |1|"sont en réalité non des universaux des langues, mais des conditions générales de cohérence de la linguistique"(p.53), on peut tout de même suggérer que ces deux types de dispositions - physiologiques et linguistiques - s'informent et se contraignent mutuellement. A ce propos, il est intéressant d'observer que de nombreux chercheurs admettent que des activités telles que la parole intérieure ou la lecture silencieuse mettent en oeuvre des programmes articulatoires. Certes, les points de vue divergent assez largement quant au rôle de cette composante: certains y voient plutôt une manifestation résiduelle, tandis que d'autres, comme Jakobson et Waugh |12|, soutiennent - à la suite de Sokolov - qu'il existe un rapport génétique entre parole intériorisée et oralisation, la première étant la projection de la seconde. Hochberg |13| pense même que la lecture courante passe toujours par la création d'un tel programme. Haber et Haber |14| ont d'ailleurs réalisé une expérience portant sur des "four-

chelangues" ("tongue-twisters"; terme proposé par Hagège |1|). Ils montrent que les phrases comportant des difficultés nécessitent une durée supérieure à celle des séquences de contrôle, non seulement lorsqu'elles sont lues à voix haute, mais également lorsqu'elles le sont silencieusement. D'autre part, certains travaux de Locke |15| portant sur l'évaluation subjective de la "difficulté articulatoire", semblent révéler - pour les mêmes phonèmes - une assez bonne corrélation entre les jugements portés par des sujets adultes et le degré de maîtrise dont font preuve des enfants de trois ans. Cela ne signifie pas qu'il ne faille tenir compte que de la phase d'exécution: il ressort de recherches utilisant des approches très différentes (Haber et Haber |4|, Smith et coll.|16|, Bonnot |17|, Bonnot et coll.|18|) que la phase de planification est également impliquée. Enfin, il a été démontré que l'activité EMG accompagnant la sous-vocalisation est spécifique, et ne peut être confondue avec une élévation généralisée du niveau de l'activité musculaire (Garrity |19|, MacGuigan et Winstead |20|).

1.4. Il ne semble donc pas possible de proposer une description réaliste de la production de la parole qui tiendrait pour quantité négligeable ce conditionnement réciproque. Nous partageons l'opinion de Keating |21|, qui écrit: "il may be that some physical patterns and movements are preferred over others because of general principles of economy of effort and motor control (e.g.Nelson 1980). The point here is that such principles must be more subtle than absolute mechanical constraints (...) Physical factors clearly influence vowel duration, but they do not control it." Cette conclusion, qui vaut ici pour la durée vocalique, peut être étendue à bien d'autres phénomènes. Nous présenterons quelques illustrations que nous tirerons plus particulièrement de nos travaux.

2. LA CONVENTION ET LE NATUREL

2.1. Hammarberg |22|, de même qu'un certain nombre d'autres phonologues, pose une équivalence entre règles d'assimilation, dont dépendent les phénomènes de coarticulation - on sait que pour Chomsky et Halle il en va autrement puisque, justement, ces faits sont sensés être universels - et les autres règles liées au contexte. Au contraire, Fowler |23| juge nécessaire de maintenir une séparation nette entre faits phonologiques et processus d'assimilation coarticulatoire. Elle se base notamment sur la différence entre l'harmonie vocalique, que l'on n'observe pas dans toutes les langues, et la coarticulation transconsonantique entre voyelles qui, elle, serait universelle. Dans un autre article, Fowler |24| fonde son choix sur ce qu'elle appelle les propriétés systématiques conventionnelles et nécessaires du langage. Elle donne comme exemple des premières la formation du pluriel régulier anglais (|s,z,Iz|). Les oppositions contrastives entre voyelles longues/brèves, ainsi que les couples consonnes emphatiques/consonnes non emphatiques en arabe sont également de bons candidats. Fowler illustre les propriétés nécessaires par l'exemple de fo suivant les occlusives sourdes. Il est tentant de verser dans cette seconde catégorie la différence de durée vocalique mesurée devant consonnes sourde/sonore, qui est attestée pour un grand nombre de langues. Il est cependant possible de trouver des contre-exemples: ainsi, en polonais, Keating |25| a montré qu'il n'existait aucune différence entre |a| précédant |t| et |a| précédant |d|. Ces conclusions valent également pour le tchèque. En arabe saoudien, Port |26| signale que, dans des monosyllabes | ʃa:k| "encercler" vs. |ʃa:g| "difficile", aucun effet des consonnes sourde et sonore ne peut être détecté. Afin de mieux spécifier l'opposition entre "conventionnel" et "naturel", Fowler rapporte une expérience de Lindblom et coll.: les voyelles phonologiquement longues du suédois sont abrégées devant consonnes de la même manière que les voyelles de l'anglais. Ce type de constatation, bien qu'intéressant, n'est nullement universel (cf.Port). D'autre part, il arrive que les nécessités proprement contrastives l'emportent sur les tendances physiologiques: en arabe saoudien, nous avons pu faire apparaître |27| que l'écart entre voyelles longues et brèves s'accroît au contact des géminées et qu'il y a donc un renforcement du contraste. La tendance "naturelle" conduit au contraire à prédire un abrègement, aussi bien pour les longues que pour les brèves. Dans d'autres situations, l'interaction entre les composantes est si complexe qu'il est pratiquement impossible d'établir une hiérarchie des déterminations: dans le système franco-provençal d'Hauteville, décrit par Martinet |28|, les consonnes s'allongent automatiquement après voyelle brève accentuée. Or, dans cette situation, les deux "r" ("faible" et "fort") qui se distinguent en intervocalique, pourraient se confondre, si "r" faible (un battement) s'allongeait: cette liquide ne se modifie donc pas. Si "r" ne s'allonge pas, c'est qu'il s'agit de maintenir une opposition, mais si la liquide est très sélectivement frappée par le phénomène, c'est à cause de sa nature acoustico-articulatoire. Un autre exemple vient étayer la démonstration: Oresnik et Pétursson |29| observent qu'en islandais du sud, les consonnes "longues" ont une durée qui n'est que très légèrement supérieure à celle des "brèves". Il s'agirait donc d'un phénomène compensatoire automatique. Dans le cas de |r|, la différence de durée est plus nette:"consonant quantity is produced only in Northern Icelandic, except for |r| which conserves the quantity in all cases." La situation n'est évidemment pas la même qu'à Hauteville, mais on retiendra qu'il s'agit dans les deux cas d'une liquide |r|.

2.2. A ce point de la discussion, il est intéressant de mettre en relation assimilation d'emphase et harmonie vocalique (dans les langues africaines par exemple). Bothorel |30| remarque que l'emphase et l'harmonie ont assez souvent été décrites à partir des "distinctions concernant les variations du volume de la cavité pharyngale.(Pour couvrir ces deux domaines, Lindau |31| propose le trait |±expanded|). Pour Fowler |23|, "vowel harmony is not unnatural in the sense that it seems to be an exaggeration of a coarticulatory tendency general to languages." Si nous raisonnions de la même façon que Fowler, nous serions amené à conclure que l'assimilation d'emphase est beaucoup moins "naturelle" que l'harmonisation vocalique, puisqu'il ne peut y avoir coproduction vocalique: il serait évidemment absurde de considérer que l'emphatisation des voyelles anté- et postposées est attribuable à un processus où la consonne serait "superposée" à une production vocalique continue.On se trouve dans une situation où, d'une manière totalement exempte d'ambiguïté, la présence du segment emphatique (la consonne) est une condition nécessaire. Nous écrivions en 1976:"Comme les consonnes, mais à un moindre degré, les voyelles |emphatisées| sont caractérisées par une grande stabilité d'un bout à l'autre de leur réalisation (...) Du point de vue des différences, excepté pour l'axe 140° |point fixe-zone uvulaire|, les voyelles entourant |t| et celles entourant |ṭ| se distinguent à tous les niveaux. En effet, les écarts maximums sont toujours plus forts dans le voisinage de la consonne non emphatique. la consonne emphatique agit sur la voyelle précédente ou suivante pour la stabiliser dans une certaine position"(p.360). Nous ne suggérons pas pour autant que les voyelles perdent leur individualité: elles acquièrent un trait configuratif, non sous l'effet de contraintes purement mécaniques - au demeurant cela n'est pas possible pour V1 - mais en raison d'une adaptation contrôlée par la microstructure temporelle (phasing). Cette situation n'est pas en contradiction avec des dispositions naturelles du tractus vocal. En ce qui concerne le russe, Purcell |31| et Öhman lui-même |32| ont montré que la coarticulation transconsonantique entre voyelles était rare. Harris |33| est obligée de reconnaître que "|Fowler's model| does not deal with competing articulation - the circumstance in which the articulators are constrained during consonant production so that free vowel-to-vowel coarticulation cannot take place." De plus, la production des consonnes emphatiques, qui met à contribution, sinon l'ensemble, du moins une partie importante du

tractus, ne constitue pas un cas de figure rare. Nous en voulons pour preuve le fait que, dans de nombreux parlers arabes, l'emphase est un phénomène très dynamique, qui s'étend bien au-delà des classiques "mutbaqa". A El-Ḥâmma de Gabès, Cantineau |34| relève que presque toutes les consonnes ayant leur lieu d'articulation en avant de la zone vélaire, ont une correspondante emphatique. Dans ce même parler, il est intéressant de noter qu'il existe au moins des "traces notables d'harmonisation vocalique, certaines voyelles ayant leur timbre conditionné par celui des voyelles des syllabes voisines"(p.225). Dans ce cas, on voit (a) que l'emphase s'étend à un grand nombre de consonnes , (b) que l'harmonie vocalique est attestée au sein du même parler, (c) que l'on a des exemples de la coexistence des deux phénomènes!

2.3. Par ailleurs, la recherche d'universaux acoustico-physiologiques, pour louable qu'elle soit, conduit trop souvent à des simplifications outrancières, dans la mesure où elle conduit à sous-estimer les relations entre éléments du système. Ainsi, Jakobson |35| a réuni 11 consonnes du palestinien druze sous le même label flat/plain. Il réduit à un seul trait distinctif un phénomène complexe. Or, il est clair que ces phonèmes correspondent à des réalités très différentes, et que l'analyse de Jakobson est à cet égard "mutilante": au plan articulatoire, les vélaires, les uvulaires et les pharyngales ne sont caractérisées que par un seul lieu d'articulation (Dkhissi-Boff |36|) alors que les emphatiques en possèdent deux |27|. D'autre part, les emphatiques sont beaucoup plus stables que leurs correspondantes non emphatiques; ce n'est pas le cas pour l'uvulaire |q| vs. |k| (pour nous, |q| n'est pas une emphatique) |27|. De son côté, Dkhissi-Boff |36| indique que les consonnes d'arrière possèdent des caractéristiques acoustiques qui les différencient très nettement des emphatiques. Se plaçant sur un terrain proprement phonologique, Mc Cawley |37| infirme lui aussi l'analyse de Jakobson: la batterie de traits est présentée comme étant non seulement universelle, mais également minimale, et les oppositions arrondi/non arrondi et pharyngalisé/non pharyngalisé sont subsumées sous le trait flat/plain. Or, le système à trois voyelles de l'arabe standard possède un |u| arrondi. La composante phonologique devrait comporter une règle du type:

$$|+\text{syllabic}| \rightarrow |+\text{flat}| \,/ \left\{ \begin{array}{l} \_\_ \; \begin{array}{l} |+\text{flat}| \\ |-\text{syll}| \end{array} \\ \\ \begin{array}{l} |+\text{flat}| \\ |-\text{syll}| \end{array} \; \_\_ \end{array} \right\}$$

"Flat" devrait être interprété comme "pharyngalisé" lorsqu'il se rapporte à la voyelle basse ou à celle d'avant, comme "arrondi" pour la voyelle postérieure au contact d'une consonne non pharyngalisée, enfin comme "arrondi+pharyngalisé", s'agissant de la voyelle au contact d'une consonne pharyngalisée. La généralisation "une voyelle est pharyngalisée lorsqu'elle est adjacente à une consonne pharyngalisée" doit donc figurer aussi bien dans la composante phonologique que dans la composante interprétative des traits: Mc Cawley souligne que cela oblige à traiter un processus phonologique unique comme s'il s'agissait de deux opérations indépendantes, localisées dans deux composantes séparées de la grammaire: "even if there are no languages in which rounding and pharyngealization function as independant oppositions, a theory which treats them as separate must still be held superior to a theory which subsumes them under a single feature"(p.524).

2.4. Les tendances naturelles sont encore limitées par les spécificités phonotactiques des langues. Fowler et coll.|38| soutiennent que la parole comporte des régularités universelles, s'agissant de la respiration, du mode phonatoire, de l'alternance des consonnes et des voyelles ou des syllabes accentuées et inaccentuées. Il va de soi qu'il y a nécessairement une succession de voyelles et de consonnes dans la chaîne parlée, mais les syllabes ne sont pas des "clones structuraux", comme l'observe justement Nolan |39|: même en anglais, on peut très bien produire un énoncé ayant la structure VC.'CCV:.CV.VCC.'CV:C.'V: . D'autre part, il faut opérer un choix au plan paradigmatique, et ce choix conditionne largement l'encodage séquentiel. Certaines langues possèdent un inventaire consonantique très riche. Il en va ainsi du sui, langue extrême-orientale, qui compte 70 unités: on y dénombre 6 séries d'occlusives (aspirées dévoisées, aspirées murmurées, voisées, implosives dévoisées et implosive voisée (1 unité))(Hagège et Haudricourt |40| p.58-60). A l'inverse, d'autres systèmes ont un nombre de voyelles fort élevé: c'est le cas des langues mon-khmer et des dialectes kuy de Thaïlande et du Cambodge (jusqu'à 48 voyelles)(ibidem p.60). Si l'on considère de plus qu'à cette extrême variété segmentale s'ajoutent un grand nombre de contraintes portant sur la forme syllabique, sur le rapport morphème/mot, sur la sélection des unités susceptibles d'être porteuses d'un degré long de quantité, ou encore sur le statut de l'accent, on comprendra qu'il n'est pas possible qu'une théorie comme celle de Fowler et coll. ait une portée universelle. Pour reprendre l'exemple du sui, l'abondance de consonnes s'accompagne d'une contrainte de taille: il y a coïncidence entre mot et morphème (langue monosyllabique) et l'on a donc affaire à un type parfaitement isolant. Au contraire, l'esquimau est extrêmement synthétique, puisque le rapport morphème/mot atteint 3.72 (anglais: 1.68)(Lyons |41|p.144-5).

3. FORMALISATION ET REGLES DE COARTICULATION

3.1. Nous considérons donc qu'il existe non seulement une continuité, mais encore une solidarité entre les processus d'encodage moteur et les structures linguistiques. Il est singulièrement difficile, s'agissant de la coarticulation, d'établir de façon irréfutable un criterium de naturel. Cette situation d'intrication complexe nous autorise théoriquement à décrire l'ensemble des phénomènes d'encodage à l'aide de règles. Cette procédure a notamment l'avantage d'être en accord avec une perspective théorique qui fait actuellement de nombreux adeptes et qui consiste, comme le souligne Keating |42|, à attribuer à la phonologie des tâches de décomposition segmentale qui étaient, il y a encore peu, l'apanage exclusif de la phonétique. Bien entendu, le corpus de règles doit être replacé dans un cadre qui, stricto sensu, n'est pas exclusivement génératif, car il doit largement faire place aux rapports entre les éléments du système, ainsi qu'à la variabilité: il est certes possible de cerner de nombreuses causes de variabilité (voir notamment Bonnot et Chevrie-Muller |43|), mais cela ne signifie pas que l'on soit en état d'en prédire à coup sûr toutes les formes. Aussi suggérons-nous de limiter les descriptions formalisées aux plans articulatoire et acoustique. L'interprétation de l'activité neuromusculaire soulève encore de nombreuses interrogations et, si nous sommes d'accord avec Jakobson et Waugh |12| pour soutenir qu'il existe "une infrastructure cérébrale des traits distinctifs", il reste pour le moins hasardeux de réunir une collection d'indices invariants (voir toutefois Bonnot |44|). En particulier, on est loin de maîtriser tous les aspects de l'interface mouvement/activité musculaire.

3.2. Pour terminer, et revenant à notre problématique de départ, nous pouvons souligner que la production de la parole n'est pas isolée des autres comportements moteurs. Ceci ressort avec une particulière netteté de l'étude de certaines pathologies (Chevrie-Muller et coll.|45|). Benjamin |46| se montrait donc particulièrement perspicace lorsqu'il écrivait, en 1935, à propos des travaux de Richard Paget: "l'articulation comme geste de l'appareil linguistique se rattache à l'ensemble de la mimique du corps. Son élément phonétique est le porteur d'une communication dont le substrat originaire était une gesticulation expressive"(p.112). La gesticulation ordonnée, spatialement et temporellement ciblée, permet la communication, que celle-ci soit orale ou non: elle est expressive, et par conséquent productrice de sens. Nous ne nions pas qu'il soit possible de mettre en évidence certaines "synergies fonctionnelles", et Chomsky nous paraît

être dans l'erreur lorsqu'il soutient qu'il est abusif de procéder à des rapprochements avec des études portant sur les animaux et même avec des activités humaines non linguistiques (marche, saut, mouvement en général). Nous croyons toutefois indispensable d'insister sur les propriétés "volitionnelles" de l'encodage moteur. C'est finalement durant l'apprentissage d'un système linguistique particulier - l'arabe, le français ou le japonais ... - que se mettent en place, puis se stabilisent et s'automatisent, des règles de coarticulation spécifiques.

REFERENCES

|1| Hagège C. L'homme de paroles, Fayard, Paris,1985.
|2| Longacre R.E. An Anatomy of Speech Notions, The Peter de Ridder Press, Lisse, 1976.
|3| Pos H. "Phénoménologie et linguistique" Keur uit de verspreide geschriften, 1, Arnhem, 1957.
|4| Chomsky N. Réflexions sur le langage, Collection "Champs", Flammarion, Paris, 1981.(éd.américe.:1975).
|5| Mehler J. "Review of The Biology and Evolution of Language, by Ph.Lieberman" JASA, 1986, 80, 1558-1560.
|6| Loucel H. "L'origine du langage d'après les grammairiens arabes" Arabica, 1963, 10, 255-267.
|7| Laitman J.T. "L'origine du langage articulé" La Recherche, 1986, 17/181, 1164-1173.
|8| Lieberman Ph."A Reply to Jacques Mehler's Review of The Biology and Evolution of Language", JASA, 1986, 80, 1521-1522.
|9| Wind J. "Phylogeny of the Human Vocal tract" , in Harnad et coll.(eds) Origins and Evolution of Language and Speech, Annals of the New York Academy of Sciences, 1976, 612-630.
|10| Lieberman Ph. "Interactive Models for Evolution: Neural Mechanisms, Anatomy and Behavior", Origins and Evolution of Language etc., 1976, 660-672.
|11| Chomsky N. et Halle M. "Some Controversial Questions in Phonological Theory" J.of Linguistics, 1965, 1, 97-138; repris dans Becker-Makkai V.(ed), Phonological Theory. Evolution and Current Practice, Holt, Rinehart and Winston, New York, 1972, 457-485.
|12| Jakobson R. et Waugh L. La charpente phonique du langage, Editions de Minuit, Paris, 1980.
|13| Hochberg J. "Toward a Speech-Plan Eye-Movement Model of Reading" in Monty R.A. et Senders J.W.(eds), Eye Movements and Psychological Processes, Lawrence Erlbaum Ass., Hillsdale NJ., 1976.
|14| Haber L.R. et Haber R.N. "Does Silent Reading Involve Articulation? Evidence from Tongue-Twisters", American Journal of Psychology, 1982, 95, 409-419.
|15| Locke J.L. "Ease of Articulation" J.of Speech and Hearing Research, 1972, 15, 194-200.
|16| Smith B.L., Hillenbrand J., Wasowicz J. et Preston J. "Durational Characteristics of Vocal and Subvocal Speech: Implications Concerning Phonological Organization and Articulatory Difficulty" J.of Phonetics, 1986, 14, 265-281.
|17| Bonnot J-F.P. Contribution à l'étude phonétique et phonologique de l'organisation temporelle de l'activité électromyographique labiale et vélaire. Coarticulation et processus d'encodage moteur, Thèse Etat Strasbourg, 1986, 704 pages.
|18| Bonnot J-F.P., Chevrie-Muller C., Arabia-Guidet C., Maton B. et Greiner G.F."Coarticulation and Motor Encoding of Labiality and Nasality in CVCVCV Nonsense Words" Speech Communication, 1986, 5, 83-95.
|19| Garrity L.I. "Electromyography: a Review of the Current Status of Subvocal Speech Research" Memory and Cognition, 1977, 5/6, 615-622.
|20| Mc Guigan F.J. et Winstead C.L. "Discriminative Relationship Between Covert Oral Behavior and the Phonemic System in Internal Information Processing" J.of Experimental Psychology, 1974, 103, 885-890.
|21| Keating P. "The Phonology-Phonetics Interface", UCLA Working Papers in Phonetics, 1985, 62, 14-33.
|22| Hammarberg R. "The Metaphysics of Coarticulation" J.of Phonetics, 1976, 4, 353-363.
|23| Fowler C. "Realism and Unrealism: a Reply" J.of Phonetics, 1983, 11, 303-322.
|24| Fowler C. "Converging Sources of Evidence on Spoken and Perceived Rhythms of Speech: Cyclic Production of Vowels in Monosyllabic Stress Feet" J.of Experimental Psychology: General, 1983, 112, 386-412.
|25| Keating P. A Phonetic Study of a Voicing Contrat in Polish, PhD., Brown University, 1979.
|26| Port R.F. "On the Structure of the Phonetic Space With Special Reference to Speech Timing" Lingua, 1981, 55, 181-219.
|27| Bonnot J-F.P. Contribution à l'étude des consonnes emphatiques de l'arabe à partir de méthodes expérimentales, Thèse de doctorat de 3e cycle, Strasbourg, 1976, 582 pages.
|28| Martinet A. La description phonologique: application au parler franco-provençal d'Hauteville, Droz, Genève, 1956.
|29| Bothorel A. "La cavité pharyngale: configuration et variations dans la chaîne parlée" Mélanges à la Mémoire de Louis Michel, Université Paul Valéry, Montpellier, 1979, 103-118.
|30| Lindau M. "|Features| for Vowels" UCLA Working Papers in Phonetics, 1975, 30, 1-155.
|31| Purcell E.T."Formant Frequency Patterns in Russian" JASA, 1979, 66, 1691-1702.
|32| Ohman S."Coarticulation in VCV Utterances" JASA, 1966, 39, 151-168.
|33| Harris K.S. "Coarticulation as a Component in Articulatory Description" in Daniloff R.G.(ed) Articulatory Assessment and Treatment Issues, College Hill Press, 1983, 147-167.
|34| Cantineau J. "Le parler d'El-Hâmma de Gabès" Etudes de linguistique arabe - Mémorial J.Cantineau , Klincksieck, Paris, 1960, 205-240 et BSL, 1951, 47, 64-105.
|35| Jakobson R. "Mufaxxama, the Emphatic Phonemes in Arabic" Selected Writings, 1, 1962, Mouton, La Haye, 510-522.
|36| Dkhissi-Boff M-C. Contribution à l'étude expérimentale des consonnes d'arrière de l'arabe classique (locuteurs marocains), Thèse de Doct.de 3e cycle, publiée dans Travaux de l'Institut de Phonétique de Strasbourg, 15, 1983.
|37| Mc Cawley J.D. "The Role of a Phonological Feature System in a Theory of Language" Langages, 1967, 8, 112-123 (en français); repris dans Phonological Theory etc., 1972, 522-528.
|38| Fowler C., Rubin P., Remez R. et Turvey M."Implications for Speech Production of a General Theory of Action" in Butterworth B.(ed) Language Production, 1, Speech and Talk, Academic Press, New York, 1980, 373-420.
|39| Nolan F.J. "The Role of Action Theory in the Description of Speech Production" Linguistics, 1982, 20, 287-308.
|40| Hagège C. et Haudricourt A. La phonologie panchronique, PUF, Coll."Le linguiste", Paris,1978.
|41| Lyons J. Linguistique générale: introduction à la linguistique théorique, Larousse, Paris, 1970.
|42| Keating P. "CV Phonology, Experimental Phonetics and Coarticulation" UCLA Working Papers in Phonetics, 1985, 62, 1-13.
|43| Bonnot J-F.P. et Chevrie-Muller C. "Analyse phonétique et phonologique et segmentation du signal électromyographique", 14e JEP, Paris, 1985, 168-171.
|44| Bonnot J-F.P."Timing extrinsèque et timing intrinsèque: le temps est-il une variable contrôlée?" 16e JEP, Hammamet, 1987.
|45| Chevrie-Muller C., Séguier N., Spira A. et Dordain M."Recognition of Psychiatric Disorders from Voice Quality" Language and Speech, 1978, 21, 87-111.
|46| Benjamin W. L'homme, le langage et la culture, Denoël-Gonthier, Paris, 1971.

ELEMENTS D'UN MODELE INTONATIF DE LA PHRASE AFFIRMATIVE EN ARABE

L. ES-SKALLI, A. RAJOUANI, M. NAJIM, M. ZYOUTE, D. CHIADMI

LEESA, Faculté des Sciences B.P 1014 Rabat - MAROC

ABSTRACT

This paper describes a procedure to compute the fundamental frequency contour of an affirmative sentence in Arabic language. The computation rules are based on three results pointed out from the analysis of a wide corpus :

- the place of lexical stress remains available in word sentence,
- the FO contour may be approximated by elementary segments corresponding to the syntactic units,
- each elementary segment is characterized by the stress level of the words embedded in the syntactic unit,

The rules are implemented as a module on the Arabic text-to-speech sythesis system by diphones.

INTRODUCTION

Ce travail rentre dans le cadre d'une étude sur la prosodie de l'Arabe dans le but d'améliorer le naturel de la parole des systèmes de synthèse developpés dans notre laboratoire [1,2]. Dans cette étude, nous nous interessons particulièrement à la formulation d'un modèle intonatif de la phrase affirmative et à l'implémentation de ce modèle sur un système de synthèse par diphones.

ELABORATION DU MODELE INTONATIF

1-Analyse des faits intonatifs.

Le corpus analysé est constitué de 120 phrases affirmatives, enregistrées par trois locuteurs, deux masculins et un feminin. Deux types de structures syntaxiques sont étudiés :

La phrase nominale composée du groupe syntaxique sujet et du groupe syntaxique attribut (GN1 + GN2).

La phrase verbale composée du verbe, du groupe syntaxique sujet du verbe et du groupe syntaxique complément (GV + GN1 + GN2).
Pour chaque structure nous avons considéré des groupes sytaxiques de différents niveaux de complexité.

Exemples : /ʔalʕilmu nūrun/ .
/ʔarrajulu ššujāʕu maḡbūbun min ṭarafi ljamīʕi/ .

L'analyse automatique du corpus effectuée aux laboratoires de l'institut de phonetique d'Aix a consisté à une représentation oscillographique de l'onde acoustique et au tracé superposé de la fréquence fondamentale et de l'intensité. Le découpage de chaque phrase en phonèmes et le relevé des parametres durée et fondamental sont effectués manuellement.
Les variations de FO étant plus significatives au niveau des voyelles, nous avons relevé les valeurs de ce paramètre conjointement avec la durée des voyelles.

Pour chaque voyelle nous avons relevé :

- Les valeurs de début et de fin de FO dans le cas où la variation sur la voyelle est supérieure au seuil de perception [3].

- La moyenne des valeurs de début et de fin de FO dans le cas où la variation sur la voyelle et inférieure au seuil de perception.

- La valeur située au 2/3 de la pente de FO de la voyelle dans le cas où la variation sur celle-ci est inférieure au seuil de perception et superieure a 10 Hz.

Pour pouvoir comparer les valeurs relatives aux différents locuteurs pour une même phrase, nous avons procédé en deux étapes :

- La détermination de la dynamique de base de chaque locuteur.

- le tracé de la courbe d'intonation en unités de perception par rapport à la dynamique de base est effectué en considérant les valeurs prises par les voyelles. Si la variation Δ entre deux valeurs successives est supérieure à un certain seuil fixe [4], elle sera percue et la différence entre les deux valeurs est exprimée en unités de perceptions UP(nombre de fois que le seuil fixe est inclu dans Δ), sinon Δ n'est pas perçue et la deuxième valeur est alignée sur la première. La moyenne en unités de perception des trois locuteurs est obtenue comme le montre le tableau suivant pour la phrase :/fataha lwaladu ṭbāba lkabīra/

syllabe	locuteur 1	locuteur 2	locuteur 3	moyenne
fa	+6	+2	+2	+3
ta	+5	+2	+2	+3
ḥa	+1	+2	+0	+1
lwa	+5	+3	+3	+4
la	+3	+3	+2	+2
du	+1	+2	+2	+2
lbā	+3	+2	+3	+3
ba	+3	+2	+2	+2
lka	+3	+3	+2	+3
bī	+5	+5	+5	+5
ra	-3	-5	-4	-4

2-Résultats de l'analyse.

Nous avons dégagé à partir de cette analyse les effets intonatifs suivants :

a- L'attaque intonative d'une phrase se situe sur la dynamique de base.

b- Chaque mot de la phrase conserve son accent de l'état isolé.

c- Globalement, deux classes de schémas intonatifs peuvent être définis pour les structures sytaxiques étudiées :

- Une première classe de schémas qui se réali-

sent dans le medium avec des maximums relatifs sur les syllabes accentuées des différents mots de la phrase. Cette classe est caractéristique des groupes sytaxiques GV, GN1, et GN2 d'une phrase verbale dans le cas où ces groupes sont tous simples.

Exemples : /dahaba lwaladu/
/fataḥa lwaladu lbāba/

- Une deuxième classe de schémas qui se réalisent dans le medium et l'infra-aigu avec des maximums relatifs sur les syllabes accentuées des différents mots de la phrase. Cette classe est caractéristique des groupes syntaxiques GN1 et GN2 d'une phrase verbale lorsqu'ils ne sont pas simples et des mêmes groupes de la phrase nominale.

Exemples : /fataha lwaladu ʒʒaʃʃru lbāba lkabīra/
/ʔalʕilmu nūrun/
/ʔarrajulu ššujāʕu maḥbūbun/

d- Toutes les phrases ont un schéma intonatif final descendant dont le minimum se réalise dans le grave.

Nous precisons que les zones fréquentielles du grave, du medium et de l'infra-aigu sont obtenues à partir de la dynamique de base par des coefficients multiplicatifs définis par Rossi [5].

3-Formalisation des règles :

Après avoir défini grossièrement en forme et en grandeur les contours intonatifs de chaque structure syntaxique, nous avons essayé de définir ces contours d'une manière plus precise par des tests de perception. Pour cela nous considérons un ensemble de phrases de différentes structures syntaxiques, les groupes syntaxiques étant simples ou composés. A chaque phrase nous avons affecté plusieurs contours différents en faisant varier les niveaux d'accentuation des différents mots dans les limites détérminées par analyse au paragraphe precédent. Les phrases synthétisées sont soumises à des tests de préférence réalisés avec la participation de huit sujets marocains. Les contours préférés par la majorité des participants et que nous avons adopté dans nos calculs peuvent être définis de la manière suivante :

- L'attaque et la fin intonative de chaque mot sont situées sur la dynamique de base fixée à 120Hz.

- Le fondamental final de la phrase se situe à -20Hz.

- Pour les phrases à unités syntaxiques simples, les maximums intonatifs sur le verbe et le sujet se situent autour de +18 Hz. Le maximum intonatif sur le complément se situe autour de +25 Hz. Le maximum intonatif sur le sujet et l'attribut d'une phrase nominale se situe autour de +35 Hz.

- Pour les phrases à unités sytaxiques composées, les mots du même groupe sytaxique ont le même niveau d'accentuation. Le maximum des groupes sujet du verbe et complément se situe autour de +35 Hz. Pour les groupes sujet et attribut d'une phrase nominale le maximum intonatif se situe autour de +25 Hz.

IMPLEMENTATION DE L'ALGORITHME DE TRAITEMENT DE L'INTONATION

Cet algorithme est implémenté comme module dans le système de synthèse par diphones à partir du texte. La génération du contour intonatif est effectuée après une ségmentation de la phrase en groupes syntaxiques et la syllabification des mots de la chaine phonétique.

1-Segmentation de la phrase en groupes syntaxiques :

Cette ségmentation est réalisée au cours de la saisie du texte à synthétiser. Des marqueurs intonatifs sont positionnés manuellement dans la chaine phonétique délimitant les groupes syntaxiques. Ils sont définis comme suit :

(#) : marqueur placé après le verbe d'une phrase verbale. Il signifie une montée intonative jusqu'à +18 Hz.

(*) : marqueur placé après le groupe nominal sujet (GN1) d'une phrase verbale ou nominale. Lorsque GN1 est simple il signifie une montée intonative sur le sujet du verbe jusqu'à +18 Hz, alors qu'il signifie une montée jusqu'à +35 Hz sur le sujet d'une phrase nominale. Lorsque GN1 est composé ce marqueur signifie une montée intonative jusqu'à +35 Hz sur tous les mots du groupe dans une phrase verbale et une montée jusqu'à +25 Hz dans une phrase nomnale.

(ʒ) : marqueur placé après le groupe complément GN2 d'une phrase verbale. Lorsque GN2 est simple, ce marqueur signifie une montée intonative jusqu'à +25 Hz sur le complément. lorsque GN2 est composé, il signifie une montée intonative jusqu'à +35 Hz sur chaque mot du groupe.

($) : marqueur placé après le groupe nominal atttribut GN2 d'une phrase nominale. Lorsque GN2 est simple, ce marqueur signifie une montée intonative jusqu'à +35 Hz et lorsque GN2 est composé, il signifie une montée intonative jusqu'à + 25 Hz sur chaque mot du groupe.

(%) : marqueur de fin de phrase. Il signifie une descente intonative finale jusqu'à 100 HZ environ.

Exemples d'utilisation des marqueurs :

- /daxala (#) lwaladu (*) %
- /fataha (#) lwaladu ʒʒapīru (*) lbāba lkabīra(ʒ) %/

2-Décomposition des mots en syllabes :

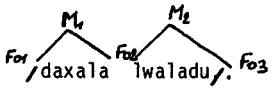
Exepté le positionnement des marqueurs intonatifs effectué par l'opérateur, toutes les autres étapes du traitement sont automatiques. Tous les marqueurs sont repérés, les groupes syntaxiques sont délimités, les limites et le nombre de mots de chaque groupe syntaxique sont memorisées afin de réaliser la syllabification de chaque mot et déterminer la place de l'accent [6]. Nous précisons que dans le cas d'un texte, il apparait d'autres types de syllabes non prévues dans la morphologie du mot isolé tel que les syllabes CCV, CCV, CCVC, CCVC et CCVCC comme dans les exemples suivants :

/daxala lwaladu/
/ʔaššala nnāra/

3- Calcul des contours de FO :

Le contour d'une phrase est obtenu par la juxtaposition des schémas intonatifs des différents groupes syntaxiques constituant la phrase, chaque schéma intonatif étant obtenu par une juxtaposition des schémas élémentaires correspondant aux mots constituant le groupe. Le contour élémentaire de FO d'un mot est formé de deux pentes, une pente montante suivie d'une pente descendante adjacente. La pente ascendante commence au début du mot à 120 Hz (dynamique de base) et se termine à la fin de la syllabe accentuée à une valeur de FO déterminée par le type du marqueur syntaxique. La pente descendante est adjacente à la première et se termine à la fin du mot à 120 Hz, pour le dernier mot de la phrase, cette pente descend jusqu'à 100 Hz.

Exemples :



FO1 et FO2 sont situées surla dynamique de base (120 Hz), FO3 se situe autour de 100 Hz. M1 et M2 sont déterminées à partir des marqueurs.

4-Résultats et discussions :

Pour évaluer les résultats obtenus, nous avons procédé, dans une première étape à des tests de perception de phrases synthétisées avec et sans intonation. Tous les auditeurs ont préféré les phrases avec intonation. Cependant le rythme est jugé très lent.
Dans une deuxième étape, nous avons comparé pour un certain nombre de phrases prononcées par deux locuteurs, les contours naturels de FO et les contours calculés par l'algorithme implémenté. La figure 1 donne les trois contours de l'une des phrases considérées. Nous signalons que la dynamique de base du premier locuteur se situe à 160 Hz environ, celle du deuxième locuteur a 125 Hz et celle adoptée dans les calculs a 120 Hz Nous constatons que pour la majorité des phrases considérées, l'allure générale du contour naturel et en particulier les lieux des maximums intonatifs sont reproduits par le contour calculé. Cependant, des variations d'ordre microprosodique apparaissent dans les contours naturels sans être reproduites dans les contours calculés. De même la durée d'une phrase synthétisée est plus longue que celle de la même phrase naturelle. Nous suggérons un traitement de rythme qui améliorait davantage le naturel d'une phrase.

Nous remercions le professeur M. ROSSI, qui a bien voulu nous recevoir dans le laboratoire de l'institut de phonétique d'Aix en Provence pour effectuer les analyses nécessaires en temp réel et qui nous a recommandé la métbode utilisée pour le relevé et l'exploitation des données d'analyse.

REFERENCES

[1] A. MOURADI, M. NAJIM. A. RAJOUANI "Unlimitted vocabularly synthesis system for Arabic language " Proc. of the 4th inter. conf. on digital processing of signal in communications. Conghbrough 22-26 Apr. 1985.

[2] A. RAJOUANI, M. NAJIM, D. CHIADMI, M. ZYOUTE "Synthesis by rule of Arabic language " to be publisched in Proc. of European Conférence on Speech Technology, Edinburgh, 1-3 Sep. 87.

[3] M. ROSSI " Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole "Phonetica, 23, pp:1-33 - 1971.

[4] M. ROSSI et A. Di CRISTO " Modele de perception et de production de l'intonation ", Edition klinkcink, Paris 1981.

[5] M. ROSSI " Les niveaux intonatifs " Traveaux de l'Institut de phonetique d'Aix en Provence, 1, page 167-176. 1972.

[6] SH. AL-ANI " Arabic phonology ". Mouton 1970.

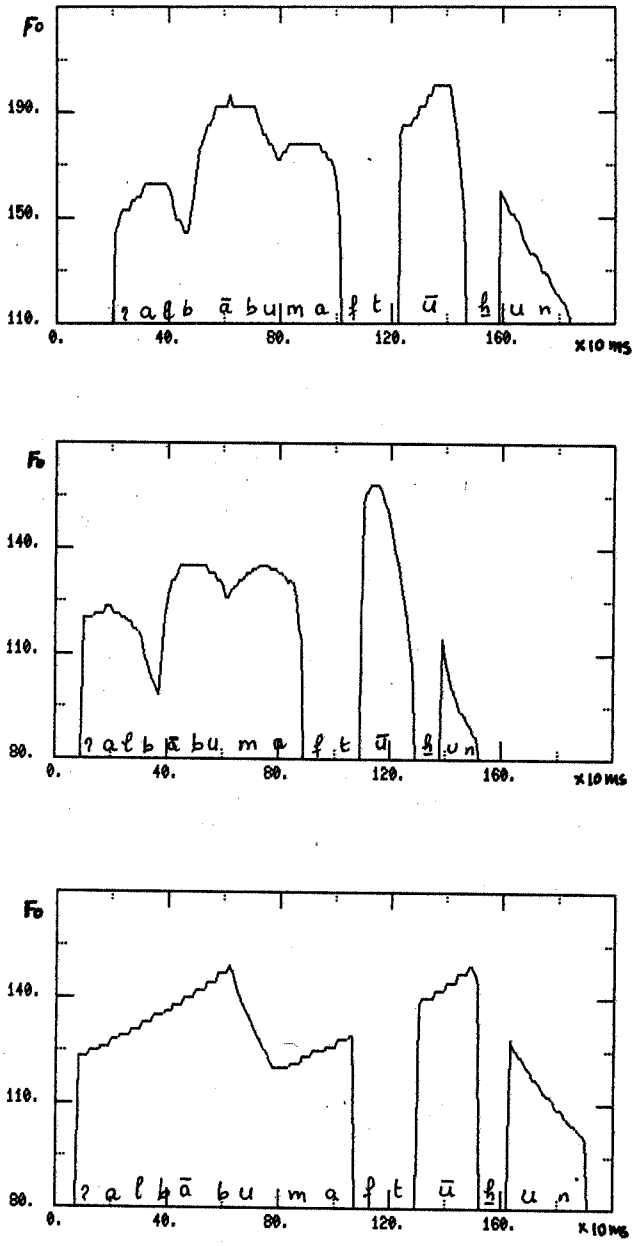


Fig. 1 Differents contours de Fo de la phrase : /ʔalbābu maftūḥun/ .

1. a : Contour du premier locuteur.
1. b : Contour du deuxième locuteur.
1. c : Contour calculé.

ETUDE EMG PRELIMINAIRE SUR LES CONSONNES ARRIERES DE L'ARABE

SALEM GHAZALI

IBLV-IRSIT, TUNIS

ABSTRACT

This EMG investigation is an attempt to determine the role of the styloglossus, the Genioglossus (3 placements), the inferior longitudinal, the middle and the inferior pharyngeal constrictors during the production of Arabic pharyngeal ; uvular and pharyngealized consonants. In this paper, only results pertaining to the activity of the Genioglossus muscle are reported. These results show that the anterior (GGA) and the middle (GGM) -but not the posterior (GGP)- portions of the genioglossus are highly active during the production of pharyngeal consonants. Results regarding the role of this muscle during the production of uvulars are inconclusive.

INTRODUCTION

La langue arabe distingue entre trois classes de consonnes articulées dans la cavité postérieure du conduit vocal. Il s'agit de :

a- Deux pharyngales fricatives [ħ] et [ʕ] produites par une constriction du laryngo-pharynx, au niveau et au dessous du niveau de l'épiglotte.

b- Une occlusive uvulaire [q] et deux fricatives révulaires -ou post-velaires- [x] et [ɣ] produites dans la zone velo-pharyngienne.

c- Un nombre de consonnes coronales pharyngalisées ([t̪] , [s̪] , [ ]...) qui, en plus de leur articulation antérieure primaire, manifestent une articulation secondaire résultant de la rétraction du dos de la langue vers la paroi postérieure de l'oro-pharynx au niveau de la deuxiéme vertebre cervicale.

Les caractéristiques acoustiques et articulatoires de ces consonnes ont fait l'objet de plusieurs études expérimentale. On trouve un compte rendu de ces études ainsi que les résultats d'une investigation acoustique et cinéfluorographique comparant ces trois classes de consonnes et leur effet coarticulatoire dans [1] et [2].

Si les informations disponibles sur les correlats acoustiques et articulatoires des consonnes arrières de l'arabe sont adéquates, rien n'est encore connu sur les structures des commandes motrices derriére les mouvements observés. Nous savons en outre qu'une consonne pharyngalisée est l'origine d'un mouvement de posteriorisation (Emphase) qui dépasse le segment adjacent pour atteindre d'autres segments plus éloignés dans le même mot [1] [2]. Si les muscles responsables de ces mouvements étaient identifiés, l'examen de leurs patrons de contractions pourrait nous informer sur les stratégies motrices derrière ce phénoméne de coarticulation.

PROCEDURE EXPERIMENTALE

1. Mise en place des électrodes . Aprés l'examen des données cinéfluorographiques relatives à la production des consonnes arriéres de l'arabe, il a été décidé de recueillir l'activité EMG des 5 muscles suivant : le génioglosse, le styloglosse, le longitudinal inférieur et deux constricteurs du pharynx (moyen et inférieur).

Au niveau du génioglosse, les électrodes étaient mises en place par voie transcutanée à trois endroits différents : la partie antérieure (GGA), la partie médiane (GGM) et la partie postérieure (GGP).

Les activités du styloglosse et du longitudinal inférieur étaient recueillies à partir d'un seul placement par muscle. La mise en place des électrodes était éffectuée par voie perorale dans la partie gauche de chaque muscle.

Au niveau des constricteurs, les électrodes étaient introduites peroralement pour le constricteur moyen et par voie transcutanée pour le constricteur inférieur. Une description détaillée des techniques d'insertion des électrodes et leur vérification se trouve dans [3].

2. Enregistrement et Analyse . Les signaux EMG étaient amplifiés et filtrés puis enregistrés simultanément avec les signaux acoustiques sur un magnétophone d'instrumentation FM (Honeywell 2600). L'analyse qui a porté sur le signal EMG rectifié et intégré représenté simultanément avec le signal EMG brut et la parole était éffectuée à l'aide d'un ordinateur PDP12. La durée et l'intensité des activités EMG aussi bien que leur début et fin par rapport aux événements acoustiques étaient identifiés et mesurés en se servant d'un programme conçu pour ces tâches. Les signaux EMG et la parole étaient aussi concervés sur papier à l'aide d'un enregistrement (siemens, oscillomink).

3. Sujet. Un homme tunisien parlant un dialecte du sud tunisien servait de sujet dans cette expérience. Le même sujet avait servi avant dans les expériences cinefluorographiques et acoustiques.

4. Corpus. Le matériau phonétique comportait 88 mots et deux phrases (arabe tunisien), choisis pour tester les consonnes pharyngales, uvulaires, pharyngalisées et non-pharyngalisées dans différents entourages vocaliques et consanantiques. Le nombre important des données phonétiques avait permis de comparer plusieurs classes de consonnes dans les mêmes conditions experimentales. Il faut cependant admettre que celà était réalisé aux dépens d'une étude plus approfondie de chacune des classes. Les mots n'étaient pas inclus dans une phrase porteuse, mais prononcés dans un ordre aléatoire avec une pause de 500 à 600 msec.

La première série d'enregistrements (12 répétitions) a permis de recueillir simultanément l'activité des trois parties du génioglosse (GGA, GGM, GGP), du longitudinal inférieur et du styloglosse. Dans une deuxiéme étape, ces électrodes étaient retirées et d'autres étaient mises en place dans les constricteurs du pharynx (12 répétitions). Au début de chaque série et réguliérement après quelques répétitions, l'emplacement des électrodes était contrôlé en prononçant les voyelles [i, e, a, o, u] bréves et longues, des syllabes du type [ki] [ka] [ku] [kit] [kæt] [kat] [pip] [pap] [pup] etc ... et en pratiquant d'autres tâches.

RESULTATS

Il ne sera pas possible de présenter tous les résultats de l'expérience dans la présente communication, la qualité d'informations recueillies étant énorme et hétérogène. Nous nous limiterons ici à l'étude des activités du génioglosse.

Avant d'examiner l'activité du GGA, GGM et GGP pendant la production des consonnes, il convient de présenter l'activité de ces trois parties pendant la production des voyelles. Ceci permettra une interprétation plus rigoureuse des bouffées EMG des consonnes dans les différens environnements vocaliques. Les résultats qui suivent (tableau 1) sont basés sur la production des voyelles [i e a o u] breves et longues, isolées, dans des logatomes, et dans le corpus expérimental. Il s'agit de moyennes d'activités évaluées sur la base de 15 répétitions. Les valeurs 0-4 représentent 4 degrés d'activité déterminés arbitrairement.

Les résultats indiquent que :

- Toutes les parties du génioglosse sont actives durant la production des voyelles antérieures fermées [i, e].

- Le GGP ne s'active que pour les voyelles fermées antérieures et postérieures [i, e, u]. Il semble servir à soulever la masse de la langue vers les parois superieures du conduit vocal.

	GGA	GGM	GGP
I	4	2	4
e	3	2	2
a	3	2	0
o	1	1	1
u	1	1	3

Tableau 1. Activité relative de 5 voyelles de l'arabe tunisien obtenue de 3 parties différentes du génioglosse.

- Le GGA trés actif durant les voyelles antérieures [i, e] manifeste très peu d'activité pour les voyelles postérieures [u, o].
- En plus des voyelles fermées, le GGA est aussi actif durant la voyelle ouverte [a].
- Le GGM montre les mêmes patrons d'activité que le GGA mais avec moins d'intensité.

Ces résultats rejoignement globalement ceux de Miyawaki [4] pour les voyelles du japonais, sauf pour la voyelle ouverte [a] où les données du japonais montrent très peu d'activité quel que soit l'emplacement choisi (5 placements).

Concernant les consonnes arrières, les signaux EMG enregistrés indiquent que :

- Le GGP n'est pas associé à la production des consonnes pharyngales, uvulaires ou pharyngalisées.
- Le GGA et le GGM sont très actifs pendant la production des pharyngales [ħ] et [ʕ], l'activité du GGM étant toujours moins intense que celle du GGA. Les figures 2 3 4 et 5 illustrent la moyenne des amplitudes relatives des activités EMG du GGP, GGA et du GGM pendant la production des pharyngales [ħ] ou [ʕ] avant les voyelles longues [æ], [i], [a] et [u].

L'activité du GGP observée dans la figure 3 commence relativement tard par rapport aux activités du GGA et GGM et est associée à la production de la voyelle [i] qui suit la pharyngale [ʕ].

Le déclenchement des activités EMG est observé jusqu'à 230 msec avant la consonne pharyngale quand celle-ci est en position initiale (fig. 2).

Tout le matériau phonétique étudié montre que quel que soit son environnement, une consonne pharyngale est systématiquement associée à une activité EMG du GGA et du GGM. Ces activités sont plus marquées pour la consonne sourde [ħ] que pour la consonne voisée [ʕ].

En se contractant pour une consonne pharyngale, les fibres du GGA et du GGM retire la partie antérieure de la langue et la poussent vers le bas produisant la configuration pharyngo-buccale illustrée (fig.6) par un cinéfluorogramme produit par le même sujet qui a servi dans la présente expérience.

Si les résultats montrent clairement que le GGA et le GGM s'activent systématiquement durant la production des pharyngales, ils ne sont

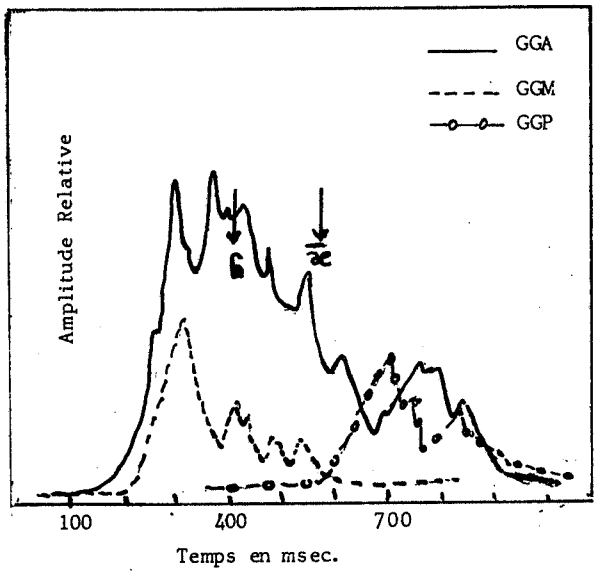


Figure 2. Activite des trois parties du genioglosse pendant le mot /ħāli /

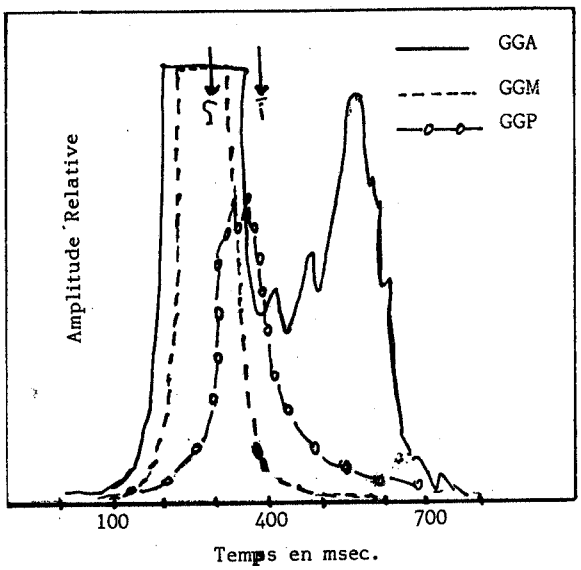


Figure 3. Activite des trois parties du genioglosse pendant le mot /bʕirɤ/

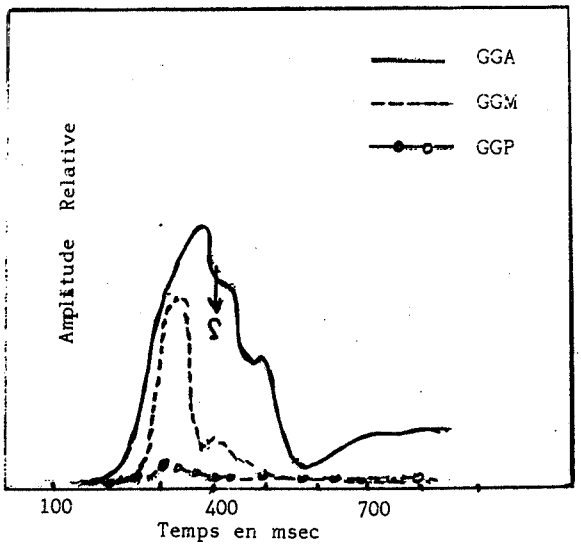


Figure 4. Activite des trois parties du genioglosse pendant le mot /tʕām/

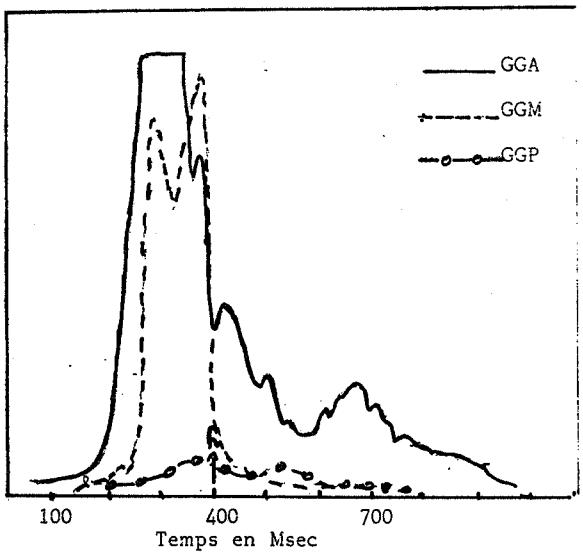


Figure 5. Activite des trois parties du genioglosse pendant le mot /ṣħūr/

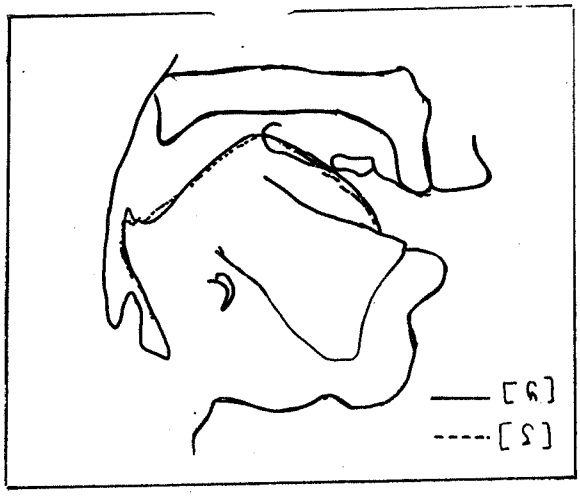


Figure 6. Forme du conduit vocal pendant la production des consonnes pharyngales /ħ/ et /ʕ/ dans les mots /ħāli/ et /ʕāli/

pas concluants pour les uvulaires et les pharyngalisées. Des conditions experimentales plus contrôlées sont nécessaires. Nous avons constaté que les bouffées EMG qui semblent accompagner la consonne uvulaire ou pharyngalisée correspondaient plutôt aux voyelles [i] et [a] qui suivent ou precedent la consonne. Ceci est indiqué par le moment du declenchement des activités EMG par rapport au segment concerné, et par le fait que cette activité est négligeable pour les uvulaires et absente pour les pharyngalisées au cas où la voyelle adjacente est [u]...

Il faut toutefois noter que dans une sequence cv où v=[a] et c=uvulaire [q] [x] [ɣ] ou pharyngalisée [ṭ] [ṣ] [ð],

- l'activité EMG associée à [a] est plus intense et commence plutôt quand la consonne est une uvulaire que quand elle est pharyngalisée.

Parmi les consonnes uvulaires, l'activité EMG est la plus intense pour [q] et la moins intense pour [ʁ]

CONCLUSION

D'un point de vue fonctionnel, le genioglosse ne semble pas se comporter comme un seul muscle [4]. Le GGA et le GGM, et non le GGP, s'activent systématiquement durant la production des consonnes pharyngales. Leurs activités EMG ne sont cependant pas forcément correlées à la production des consonnes uvulaires et pharyngalisées.

REMERCIEMENTS

Cette expérience a été rélaisée au laboratoire de phonétique de l'université du Texas grâce à une assistance du NSF (BNS77-07686). Les électrodes étaient mises en place par Seiji Niimi et Hirohide Yoshioka. Les conseils de Peter MacNeilage dans l'identification des muscles à étudier étaient précieuses.

REFERENCES

[1] GHAZALI, SALEM. Back consonants and Backing coarticulation in Arabic. These de PhD. Université du Texas, 1977.

[2] GHAZALI, SALEM. La Coarticulation de l'Emphase en Arabe. ARABICA, Tome XXVIII, Fascicule 2-3, 1983.

[3] HIROSE, HAJIMI. Electromyography of the Articulatory muscles : current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, pp73-86, 1971.

[4] MIYAWAKI, K. HIROSE, H. USHIJIMA, T. & SAWASHIMA, M. Prealiminary Report on the Electromyographic Study of the Activity of Lingual Muscles. Annual Bulletin, RILP, N° 9, pp91-106, 1975.

CONTRIBUTION A LA SYNTHESE DE LA PAROLE EN ARABE STANDARD

Mhania GUERTI

Ecole Nationale Polytechnique, Hassen Badi, El-Harrach, ALGER - ALGERIE -

ABSTRACT :

The object of this study was to synthesize in real time any standard arabic message on the basis of a diphone dictionary.

The analysis of the vocal signal has been made by Linear Predictive Coding (LPC).
The manual segmentation of diphones led to an intelligible synthetic speech.

1. INTRODUCTION.

Cette étude représente à notre connaissance, la première tentative de synthèse de parole en Arabe standard *.

Pour parvenir à cette fin, nous avons constitué un dictionnaire de diphones, en utilisant la technique de prédiction linéaire. L'ensemble de ces diphones permet de synthétiser tout le vocabulaire de l'Arabe.

Nous avons illustré la segmentation par diphones, par le choix des principales particularités qui existent en Arabe : géminée; emphatique; longue.

2 - LES SONS DE L'ARABE STANDARD

Le système consonantique de l'Arabe se compose de 28 consonnes, chacune d'elles correspond à un phonème particulier (fig. 1).

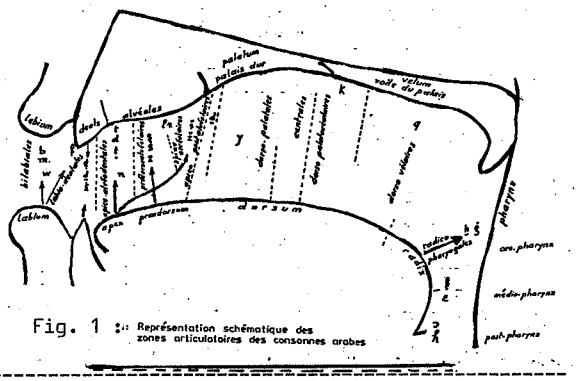


Fig. 1 : Représentation schématique des zones articulatoires des consonnes arabes

(*) *Ce travail a été effectué au CNET (Centre National d'Etudes des Télécommunications) de LANNION, dans le Département RCP (Recherches en Communication parlée). Il s'insère dans le cadre d'une coopération franco-algérienne.*

L'originalité de la phonétique arabe se fonde, pour une part importante sur les consonnes pharyngales, laryngales et emphatiques (fig.2).

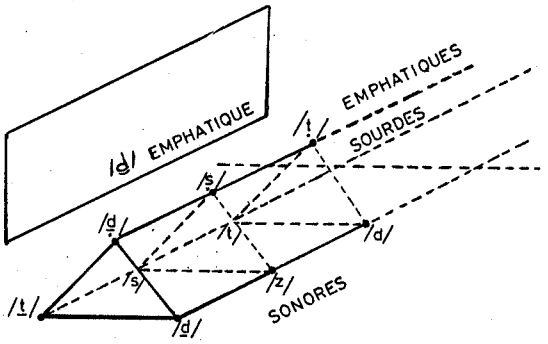


Fig. 2 : Les quatre emphatiques de l'Arabe.

Le système vocalique comprend 3 voyelles brèves, s'opposant aux 3 voyelles longues (fig.3).
Remarque : Il n'existe pas de voyelles nasales en tant que phonèmes, en Arabe.

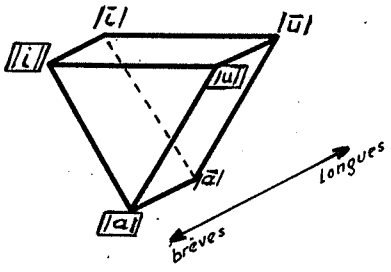


Fig. 3 : Système vocalique de l'Arabe du point de vue phonologique.

On a noté que dans la réalisation acoustique, les phonèmes fondamentaux /a,u,i/ comportent diverses variantes :

-/l'ɔimāla / est une prononciation antérieure du /a/, son point d'articulation se rapprochant de /e/ voire de /i/;

-le /tafẖīm/ : la présence de ces consonnes/r,l, y,q,ẖ,ġ,ṣ,ḍ,ṭ,ẓ/ reportent en arrière le point d'articulation des voyelles /a,u,i/, de sorte qu'elles deviennent /a,o,e/ (fig.4).

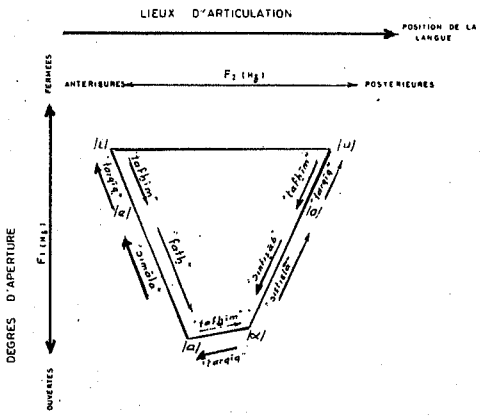


Fig. 4 : Système vocalique de l'Arabe standard (documents élaborés par A.HADJ SALAH.)

3 - LA CONSTITUTION DU DICTIONNAIRE DE DIPHONES

Pour constituer un dictionnaire de diphones, il faut disposer de toutes les combinaisons réalisables. Si la langue concernée comporte "n phonèmes", le nombre de diphones nécessaires est de l'ordre de "n^2".

Pour l'Arabe nous avons "1681" diphones dont "325" d'entre eux n'apparaissent pas (tableau 1).

Tableau 1 : Calcul du nombre de diphones de l'Arabe standard.

$C_{\bar{E}}$: Consonne non emphatique

C_E: consonne emphatique

V_E: Voyelle emphatisée

$V_{\bar{E}}$: Voyelle non emphatisée.

Diphones	Nb.de C.R*	Combinaisons irréalisables
silence-consonne :		
$\#\, C_{\bar{E}}$	24	
$\#\, C_E$	4	
consonne-silence :		
$C_{\bar{E}}\, \#$	24	
$C_E\, \#$	4	
consonne-voyelle :		
$C_{\bar{E}}\, V_{\bar{E}}$	24x6=144	
$C_E\, V_E$	4x6=24	
$C_{\bar{E}}\, V_E$	24x6=144	- Une consonne non emphatique ne peut pas être suivie d'une voyelle emphatisée.
$C_E\, V_{\bar{E}}$	4x6=24	- Une consonne emphatique ne peut pas être suivie d'une voyelle non emphatisée.
voyelle-consonne :		
$V_E\, C_E$	6x4 = 24	
$V_E\, C_{\bar{E}}$	6x24 = 144	
$V_{\bar{E}}\, C_E$	6x 4 = 24	
$V_E\, C_{\bar{E}}$	6x24 = 144	
consonne-consonne :		
$C_E\, C_{\bar{E}}$	4x24 = 96	
$C_{\bar{E}}\, C_E$	24x4 = 96	
$C_E\, C_E$	4x4 = 16	
$C_{\bar{E}}\, C_{\bar{E}}$	24x24 = 576	
silence-voyelle :		
$\#\, V_E$	6	On ne peut pas commencer un mot par une voyelle, car celle-ci est intimement liée à la consonne précédente.
$\#\, V_{\bar{E}}$	6	
voyelle-silence :		
$V_E\, \#$	6	
$V_{\bar{E}}\, \#$	6	
voyelle-voyelle :		
VV	12x12 = 144	On ne peut pas avoir deux voyelles consécutives en arabe
silence-silence :		
$\#\, \#$	1x1 = 1	Cette combinaison ne représente pas un diphone.

La mise en oeuvre d'une telle procédure comporte les étapes suivantes :

3.1. Choix du code phonétique

Pour homogénéiser la lecture des caractères par le calculateur (PDP 11/34), nous avons codé tous les phonèmes à l'aide de 2 caractères latins. Nous avons fait précéder d'un blanc ceux qui n'en comportent qu'un seul (tableau 2).

Tableau 2 : Code phonétique choisi.

---

* Nombre de combinaisons réalisables.

	PHONEMES	CODE DES ARABISANTS	CODE PROPOSE
CONSONNES NON EMPHATIQUES	ب ت ث ج ح خ د ذ ر ز س ش ك ل م ن ه و ي ء ق ر ل	b t c j h x d z r s c k l m n h w y ' q r l	B T TH J HH D DH Z C CH X F K L M N H W Y KH GU Q R L
CONSONNES EMPHATIQUES	ص ض ط ظ	s d t z	S DG DK TJ
VOYELLES AU VOISINAGE DES NON EMPHATIQUES	ـَ ـِ ـُ ـَا ـِي ـُو	a i u aa ii uu	A I U AA II UU
VOYELLES AU VOISINAGE DES EMPHATIQUES	ـَ ـِ ـُ ـَا ـِي ـُو	a e i o ie ou	AE IE UE AL IL UL

3.2. Choix du corpus de mots.

Le corpus que nous avons choisi pour extraire les diphones comporte des mots trisyllabiques artificiels.

3.3. Choix du locuteur

En ce qui concerne ce choix, le substrat dialectal doit être le plus faible possible, afin de respecter les principes de l'Arabe standard.

3.4.Définition des phonèmes

La liste des phonèmes a été mise dans un fichier Chaque phonème est écrit en code phonétique avec des indications binaires (tableau 3).

Tab.3:Les traits distinctifs des phonèmes arabes.
- Le guillemet indique la présence d'un son codé à l'aide de deux caractères latins.

NUMERO DU PHONEME	PHONEMES	CONSONNES	VOISEES	PLOSIVES	FRICATIVES
1	A'	2	0	0	0
2	B	1	0	1	0
3	A	1	0	0	0
4	T	1	0	1	0
5	TH"	1	0	0	1
6	J	1	1	0	1
7	HH"	1	0	0	1
8	D	1	1	1	0
9	DH"	1	1	0	1
10	Z	1	1	0	1
11	C	1	0	1	0
12	CH"	1	0	0	1
13	X	1	1	0	1
14	F	1	0	1	0
15	K	1	0	1	0
16	M	1	1	1	0
17	N	1	1	1	0
18	H	1	0	0	1
19	W	1	1	1	0
20	Y	1	1	1	0
21	KH"	1	0	0	1
22	GU"	1	1	0	1
23	Q	1	0	1	0
24	R	1	1	1	0
25	L	1	1	1	0
26	S	1	0	1	0
27	DG"	1	1	1	0
28	DK"	1	0	1	0
29	TJ"	1	1	1	0
30	A	0	1	0	0
31	I	0	1	0	0
32	U	0	1	0	0
33	AA"	0	1	0	0
34	II"	0	1	0	0
35	UU"	0	1	0	0
36	AE"	0	0	0	0
37	IE"	0	0	0	0
38	UE"	0	0	0	0
39	AL"	0	0	0	0
40	IL"	0	0	0	0
41	UL"	0	1	0	0

3.5. Numérisation et analyse du corpus

Après numérisation le corpus est analysé en prédiction linéaire. L'analyse fournit une valeur du gain de la fréquence fondamentale et 12 coefficients de réflexion.

3.6. Segmentation en diphones

Deux méthodes de segmentation manuelle ont été possibles :- à partir des courbres résultant de l'analyse LPC et
- à partir de la représentation en format vocodeur à canaux.

Pour extraire un diphone, on doit chercher les zones stationnaires de la fonction de stabilité spectrale. Le diphone est extrait entre les minima les plus marqués de cette fonction pour les deux phonèmes qui le constituent. Son milieu est pris au maximum de la fonction de stabilité (fg. 5,6 et 7).

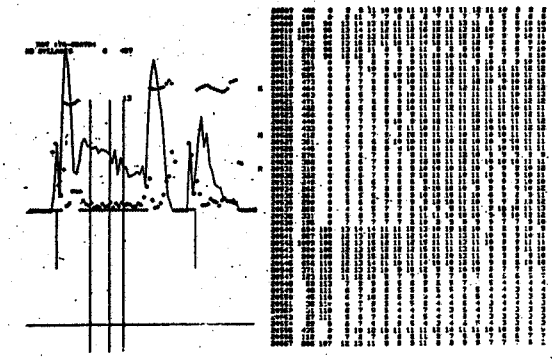


Fig.5:Résultat de l'analyse LPC du mot /TA-CCATA/
- : Gain oo : Stabilité spectrale
xx: Période fondamentale ++ : Dérivée du gain

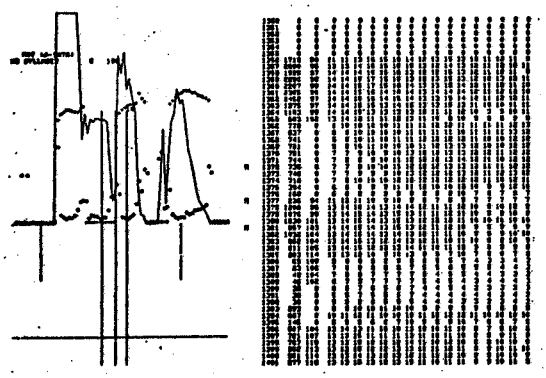


Fig.6: Résultat de l'analyse LPC du mot / A-SATA /

Les géminées et les emphatiques (cf.fig.5 et 6) n'ont pas présenté de difficulté particulière. Pour ce qui est des géminées, nous les avons considérées comme des consonneslongues.

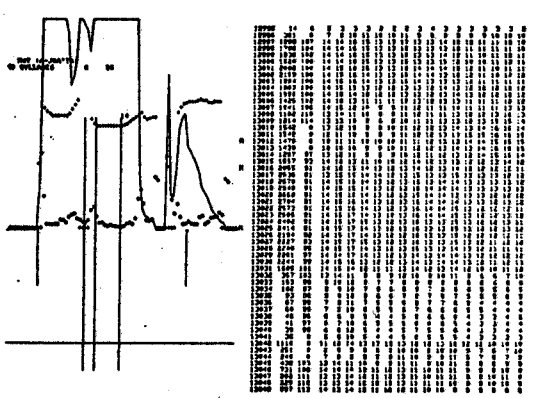


Fig.7: Résultat de l'analyse LPC du mot /A-JAA'TA/

4 - CONCLUSION

L'intelligibilité est généralement le critère le plus important de qualité de la parole synthétique, mais pour apprécier celle-ci un autre paramètre: "le naturel" est à considérer également.

La parole synthétique que nous avons obtenue était intelligible. On arrivait à identifier le caractère masculin ou féminin du locuteur. Mais le naturel reste à améliorer par des paramètres prosodiques.

L'Arabe standard trop souvent réputé difficile, souffre cependant de l'insuffisance des travaux sur le traitement automatique de la parole.

C'est pourquoi une collaboration entre les spécialistes en acoustique, électronique, informatique, linguistique, phonétique, traitement du signal... nous semble particulièrement indispensable dans ce domaine.

BIBLIOGRAPHIE

BONNOT,J.F.(1976),"Contribution à l'étude articulatoire et acoustique de certaines consonnes emphatiques de l'Arabe à partir de méthodes expérimentales". Thèse de Doct. 3e cycle, STRABOURG II

CANTINEAU,J.(1960),"Cours de phonétique arabe".Paris, Librairie C. KLINCKSTECK.

EMERARD,F.(1977),"Synthèse par diphones et traitement de la prosodie".Thèse de Doctorat de 3e cycle, GRENOBLE.

GUERTI,M.(1983),"Contribution à la synthèse de la parole en Arabe standard (synthèse par diphones et technique de prédiction linéaire).Thèse de Magister-Université d'Alger.

STELLA,M.(1983),"Fabrication semi-automatique de dictionnaires de diphones". Recherches Acoustiques Vol.III,PP.51-63 CNET-LANNION.

DE QUELQUES ASPECTS RYTHMIQUES DE L'ARABE DIALECTAL TUNISIEN

DRISS KORCHANE et FRANCOIS WIOLAND

INSTITUT DE PHONETIQUE - STRASBOURG

L'analyse et la comparaison de corpus d'arabe dialectal tunisien et de francais parlé à la radio révèlent des différences significatives dans la distribution des structures syllabiques et dans les rapports respectifs entre syllabes ouvertes / syllabes fermées, durées vocaliques / durées consonantiques dans le cadre d'une même structure syllabique et durée / intensité vocaliques.

Dans le cadre de nos recherches sur l'accent et le rythme de l'arabe dialectal tunisien et du français parlé nous avons enregistré d'une part un conte radiophonique dit par Abdelaziz LAROUI et d'autre part un extrait de discours radiophonique sur Europe I, bénéficiant ainsi des avantages reconnus de ce type de document à savoir la qualité sonore et la relative spontanéité des locuteurs enregistrés.

Distribution syllabique

- Structures syllabiques observées

La distribution respective des structures syllabiques observées est la suivante :

I - En arabe dialectal tunisien

1 - CVC environ 40%
2 - CV " 35%

puis dans des proportions comparables mais avec une fréquence inférieure à 10%

3 - CCV " 7%
CCVC " "
VC " "

et enfin

6 - CVCC " 2%
CCVCC " "

la structure V étant plus rare encore.

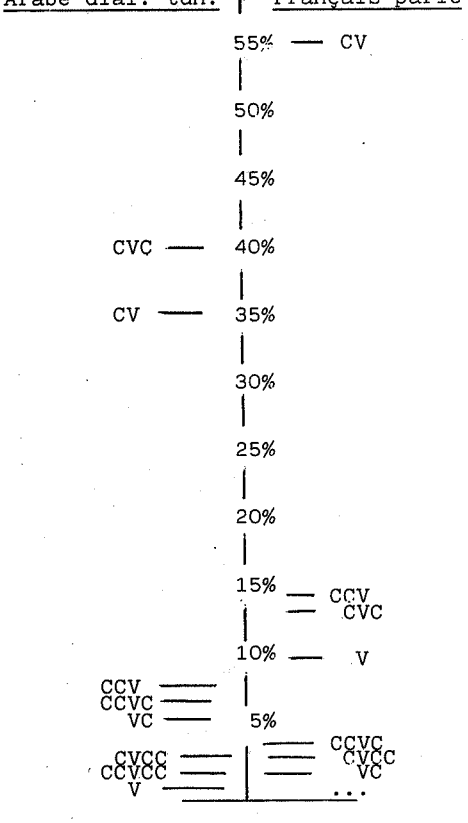
L'approximation des pourcentages en arabe dialectal tunisien s'explique par le corpus moins important qu'en français.

II - En français parlé

1 - CV 55,61%
2 - CCV 13,90%
3 - CVC 13,55%
4 - V 9,80% soit 92,86% du total
5 - CCVC 2,65%
6 - CVCC 1,50%
7 - VC 1,32% soit 98,33% du total

les autres structures ne représentant que 1,67%.

En comparant les résultats nous observons que la distribution des structures syllabiques diffère de façon importante :



Si les deux structures CVC et CV représentent 75% des occurrences en arabe dialectal tunisien et 70% en français parlé, l'ordre est inversé d'une langue à l'autre et

les pourcentages respectifs très différents.

En arabe dialectal tunisien aucune structure n'atteint la moitié des occurrences alors qu'en français parlé la structure CV la dépasse.

La structure CCV qui représente un pourcentage légèrement supérieur à CVC en français est deux fois moins fréquente en arabe dialectal tunisien alors que pour CCVC le rapport est inverse.

La structure V qui en français atteint 10% environ des occurrences et occupe le 4° rang n'apparaît en arabe dialectal que lorsque l'occlusive glottale disparait.

- Rapport syllabes ouvertes / fermées

En arabe dialectal tunisien les structures syllabiques fermées l'emportent sur les structures syllabiques ouvertes dans un rapport d'environ 57% contre 43%.

En français du fait de l'enchainement dans le discours les structures syllabiques ouvertes atteignent 80% contre 20% pour les structures syllabiques fermées.

	Structures syllabiques fermées	ouvertes
arabe dia. tun.	57%	43%
français parlé	20%	80%

Rapports sur le plan de la durée entre voyelles et consonnes

Pour des raisons évidentes de nombre d'occurrences statistiquement représentatives nous nous sommes limité aux deux structures syllabiques CV et CVC.

- Structure syllabique CV

En français parlé toutes positions accentuées, accentuables et inaccentuables confondues nous observons que la durée vocalique est en moyenne inférieure à celle de la consonne dans un rapport de 1,15 au bénéfice de la consonne et de 0,85 pour la voyelle.

Dans 46% des cas la durée vocalique est inférieure à celle de la consonne, dans 32% des cas elle est égale et supérieure dans seulement 21% des cas. Le rapport de 1;8 au bénéfice de la voyelle n'est atteint que pour certaines voyelles accentuées à durée très marquée mais relativement peu nombreuses dans ce type de discours. Mais pour certaines voyelles inaccentuées il peut descendre à 0,30.

En arabe dialectal tunisien toutes positions confondues la durée vocalique est en moyenne supérieure à celle de la consonne dans un rapport de 1,97 contre 0,51 pour la consonne.

- Structure syllabique CVC

Si dans les deux langues le rapport est en faveur des consonnes il est plus élevé en français qu'en arabe dialectal tunisien :

		Arabe dia.tun.	Français parlé
C	C	1,12	1,68
	V	0,89	0,53

- Structure syllabique CCV

En français parlé nous avons noté une différence très marquée entre les syllabes accentuées et les syllabes inaccentuées qui affichent respectivement des rapports de 1,01 et de 2,03 au bénéfice des consonnes.

Rapports durée / intensité vocalique

Sur le tracé oscillographique nous avons observé au cours de la durée vocalique en arabe dialectal tunisien, en moyenne plus importante qu'en français, une augmentation de l'énergie acoustique et non pas comme souvent en français une diminution voire même une chute notoire de la courbe de l'intensité. Ceci peut expliquer en partie l'impression que l'intensité perçue en arabe dialectal tunisien serait plus forte qu'en français parlé alors que les valeurs absolues relevées sont du même ordre dans les deux langues.

Bien que l'on puisse remarquer une augmentation de la durée syllabique en fin de syntagme aussi bien en arabe dialectal tunisien qu'en français les quelques différences relevées sont des constituants de schémas rythmiques qui n'ont rien de commun.

VALIDITE ET LIMITES DU DIPHONE EN TANT QU'UNITE DE SYNTHESE POUR LA LANGUE ARABE STANDARD

A. MOURADI

FACULTE DES SCIENCES
B.P. 1014 RABAT-MAROC

ABSTRACT

In a previous work we presented a synthesis system based on diphone concatenation [1]. The purpose of this paper is to show the validity and limits of the diphone as a minimal unit for Arabic language synthesis assuming that each diphone has only one version in the dictionary. The following aspects will be discussed:

- pharyngealized consonants,
- gemination of plosives,
- phonemes strongly altered by the neighbourhood,
- consonants that the voiced/unvoiced attribute is context dependent.

I. INTRODUCTION

La synthèse à partir du texte peut être réalisée en utilisant l'une des deux approches suivantes:

- établir les règles régissant les transitions entre phonèmes successifs,
- mémoriser ces transitions dans un dictionnaire.

Cette seconde approche a été utlisée pour la synthèse de plusieurs langues, notamment le français [2,3], l'allemand [4] et l'italien [5]. Elle a été aussi appliquée pour la synthèse de la langue arabe standard [6,7].

Dans un travail précédent [1], nous avons présenté un système de synthèse de la langue arabe basé sur un dictionnaire de diphones codés sous forme de paramètres LPC. Dans ce papier, nous nous proposons de discuter la validité du diphone en tant qu'unité de synthèse pour l'arabe standard et de montrer les limites dans certains cas particuliers.

II. DESCRIPTION SUCCINTE DU SYSTEME DE SYNTHESE PAR DIPHONES DE L'ARABE.

La langue arabe standard comporte 34 phonèmes de base : 6 voyelles et 28 consonnes. Un dictionnaire de 1225 diphones a été élaboré correspondant aux 34 phonèmes de l'arabe et au silence. Chaque diphone n'a qu'une seule version dans le dictionnaire codée sous forme de paramètres LPC. Le système de synthèse de l'arabe à partir du texte que nous avons réalisé est basé sur l'exploitation de ce dictionnaire de diphones.

III. SYNTHESE DE L'ARABE PAR DIPHONES : VALIDITE ET LIMITES.

Les tests effectués à l'aide du système de synthèse nous ont permis de constater que dans certains cas une seule version pour chaque diphone n'est pas suffisante pour avoir une parole acceptable. D'autres tests nous ont même poussé à étendre les phonèmes de base en ajoutant d'autres voyelles. Quatre classes de problèmes ont été étudiées:

III.1 Les consonnes emphatiques

L'arabe contient quatre consonnes emphatiques, à savoir: /t̲/, /s̲/, /d̲/ et /ð̲/. Ces consonnes affectent la qualité des voyelles adjacentes. Le phénomène de l'emphase peut aller au déla de la voyelle adjacente et influencer les voyelles du mot en entier. Ce fait va à l'encontre de la synthèse par diphones qui suppose que le phénomène de coarticulation ne dépasse pas le demi phonème.

Nous avons constaté que les voyelles adjacentes à une consonne emphatique sont mal perçues. Sur l'exemple suivant nous allons voir les limites du diphone pour ce type de phonèmes.

Supposons que nous voulons synthétiser le mot "saqat̲a" (il est tombé), alors nous allons procéder à la concaténation des diphones:

#s sa aq qa at̲ t̲a a#

Il est facile de remarquer que la dernière voyelle /a/ du mot "saqat̲a" est constituée de deux parties. Une première partie a été enregistrée dans un contexte emphatique et se trouve dans le diphone t̲a. Par contre la seconde partie se trouvant dans le diphone a# a été enregistrée dans un contexte non emphatique. Pour résoudre ce problème, nous avons proposé d'élargir le système vocalique arabe en ajoutant 6 autres voyelles, appelées voyelles emphatiques. Cette solution augmente la taille du dictionnaire mais elle permet d'améliorer la qualité de la parole synthétisée. Pour tenir compte de l'influence d'une voyelle emphatique sur les voyelles du mot il faudrait alors établir des règles de propagation de l'emphase.

III.2 Les plosives géminées

En arabe, toute consonne en position médiane peut être géminée. La gémination d'une consonne se traduit, lors de la transcription graphèmes-phonèmes, par un dédoublement de cette consonne.

La gémination des consonnes autres que les plosives peut être réalisée d'une manière simple. Il suffit d'ajouter un diphone correspondant à la consonne géminée comme le montre l'exemple suivant: pour synthétiser le mot "kassara" (ss désigne la consonne /s/ géminée) il suffit de mettre bout à bout les diphones: #k ka as ss sa ar ra a#

Nous avons essayé d'appliquer la même procédure aux plosives. Le diphone correspondant à la gémination étant extrait dans ce cas du silence de fermeture. Les résultats obtenus n'étaient pas satisfaisants et on n'arrivait pas à créer une "bonne" plosive géminée avec le burst d'une plosive non géminée. Alors, nous avons procédé d'une manière inverse en essayant de réaliser une plosive non géminée à partir de la géminée correspondante en diminuant la durée de fermeture. Dans ce cas les résultats obtenus sont satisfaisants.

III.3 Consonnes fortement colorées par les voyelles adjacentes

Certaines consonnes de l'arabe sont fortement colorées par les voyelles voisines et il est difficile de pouvoir effectuer la synthèse avec une seule représentation par diphone dans ce cas.

Prenons l'exemple de la pharyngale /ع/ suivie des voyelles /a/ et /i/ comme dans les mots "عilmun" et "عalamun" (science et drapeau). La décomposition en diphones de ces deux mots est:
"عilmun" ---> #ع عi il lm mu un n#
"عalamun" ---> #ع عa al la am mu un n#

Selon le mot à synthétiser, Le diphone #ع est concaténé au diphone عi ou عa. Or ce diphone est unique dans le dictionnaire et il a été extrait d'un logatome où la consonne /ع/ figure à l'initiale suivie de la voyelle /a/. Si la synthèse du mot "عalamun" ne pose pas de problèmes, celle du mot "عilmun" est mal produite au niveau de la pharyngale /ع/ car le diphone #ع est coloré par la voyelle /a/ par construction alors que le diphone عi est coloré par la voyelle /i/. Pour des raisons analogues, il difficile de synthétiser le mot "taعiba" (il s'est fatigué) pour lequel il faudrait juxtaposer les diphones aع et عi. Pour ce genre de problèmes, il convient d'avoir plusieurs représentations d'un même diphone et choisir celle qui convient en fonction du contexte ou trouver des solutions ad-hoc.

III.4 Consonnes dont le voisement dépend du contexte

Certaines consonnes de l'arabe comme la laryngale /h/ sont voisées à l'intervocalique et peuvent être non voisées dans d'autres contextes comme à l'initiale. Or lors de l'élaboration du dictionnaire et la mémorisation du diphone ha on ne sait pas à priori si la consonne /h/ sera utilisée au début du mot ("hamada") ou entre deux voyelles ("dahaba") et par conséquent est-ce-qu'il faut mémoriser la partie h du diphone ha en tant que voisée ou non. Pour ce type de phonèmes nous avons constaté qu'il est préférable de les mémoriser en tant que consonnes voisées. Toutefois, il est souhaitable de faire un traitement particulier lorsqu'elles sont en position initiale ou dans un contexte sourd et de dévoiser la partie du diphone correspondant à la consonne en question.

IV. TRAITEMENT DE LA PROSODIE

La parole arabe produite par le système de synthèse réalisé est intelligible mais elle manque de naturel. Pour améliorer la qualité de cette parole il est nécessaire d'intégrer les variations prosodiques. Un traitement préliminaire est effectué au niveau du mot, ce dernier est décomposé en syllabes et l'accent primaire est placé sur la syllabe appropriée.

V. CONCLUSION

Ce premier système de synthèse nous a été très utile pour expliciter certains problèmes liés à la langue arabe. Il nous a permis de tester la validité du diphone pour la synthèse de l'arabe. D'autres unités, peut être plus appropriées, telles que les demi-syllabes vont être essayées.

Un effort considérable doit être fait pour dégager des règles prosodiques, nécessaires à tout système de synthèse.

BIBLIOGRAPHIE

[1] A. MOURADI, M. NAJIM, A. RAJOUANI
" Synthèse de l'arabe à partir du texte."
13èmes J.E.P Bruxelles Mai, 1984.

[2] J.S. LIENARD, D. TEIL, C.CHOPPY, G.RENARD, J.SAPLAY
" Diphone Synthesis of French : Vocal Response Unit and Automatic Prosody from Text. "
IEEE ICASSP 1977.

[3] J.L. COURBON, F. EMERARD
" SPARTE : A Text-to-Speech machine using synthesis by diphones. "
IEEE ICASSP, Paris 1982.

[4] E. VIVALDA, S. SANDRI, C. MIOTTI
" Real-time Text Processing for Italian Speech Synthesis. "
Proc. ICASSP, 1979

[5] H.W. STURBE, R. WILHELMS
" Synthesis of Unrestricted German Speech from Interpolated LOG-AREA-RATIO Coded Transitions. "
Speech Communication, 1982

[6] M. GUERTI
" Contribution à la synthèse de la parole en arabe standard. "
Diplôme de Magister des Sciences Univ Alger, 1983.

[7] A. MOURADI
"Synthèse de la parole arabe à partir du texte par la méthode des diphones."
Thèse Doctorat d' Etat Jul. 1985, Rabat Maroc.

LA PHARYNGALISATION DES CONSONNES LABIALES

G. PUECH, N. LOUALI ET R. HAMDI

UNIVERSITE LUMIERE-LYON 2, C.R.L.S., F - 69500 BRON

X - ray films were taken of one Berber native speaker from El Aderj (Morocco) and of one Moroccan Arabic native speaker from Casablanca. The purpose of the present study was to determine to what extent labial consonnants, which do not involve primary lingual articulation, may be pharyngealized by retraction of the tongue root. From the measurements carried out on the film frames, it appears that, in the appropriate contexts, labials are indeed pharyngealized. Evidence is presented that, in order to realize a word terminal pharyngealized labial, constriction of the pharynx affects the adjacent segment across the word boundary; this segment, however, perceptually retains its non pharyngealized allophonic identity.

L'arabe et le berbère, qui appartiennent à deux branches distinctes de la famille chamito-sémitique [1] font un usage distinctif de la prosodie d'emphase. Une des principales caractéristiques de l'emphase est de surimposer une pharyngalisation sur une articulation primaire. D'après Catford [2] "Pharyngealized sounds involve some degree of contraction of the pharynx either by a retraction of the root of the tongue, or by lateral compression of the faucal pillars and some raising of the larynx, or a combination of these" (p. 193).

Dans la koinè [3] et certaines des langues arabes contemporaines, le contraste d'emphase n'affecte pas les consonnes labiales, contrairement au berbère et à de nombreuses autres langues arabes. Pour Jakobson [4] la rareté des labiales pharyngalisées s'explique par le fait que "the narrowing of a wide orifice separating a pharyngealized dental from its plain counterpart is much more contrastive than the narrow orifice with an additional narrowing which distinguishes the pharyngealized from the non pharyngealized labials" (p. 108). Au contraire, Ghazeli [5] constatant l'existence de consonnes labiales pharyngalisées en tunisien écrit "The reason that [b] can exhibit a high degree of backing may be attributed to the nature of its articulation wich does not actively involve the tongue. In other words, the entire mass of the tongue can move posteriorly, without having the rearward movement of the back of the tongue being cancelled in part by an anterior movement of its blade or tip..." (p. 93).

Nous avons étudié ce qu'il en est de la pharyngalisation des labiales dans le berbère d'El-Aderj (groupe Tamazight) et l'arabe de Casablanca. Comme c'est le cas pour l'ensemble du domaine berbère et du domaine arabe, seules les consonnes coronales induisent une prosodie de pharyngalisation. Le sujet propre de cette communication est de savoir comment elle se manifeste pour les labiales. Des films cinéradiographiques ont été réalisés en collaboration avec l'Institut de Phonétique de Strasbourg au centre médico-chirurgical de Schitilgheim, avec deux locuteurs natifs l'un d'El-Aderj et l'autre de Casablanca.

Les films (50 images/s.) ont été exploités selon les méthodes décrites dans Bothorel [6] notamment. Après détermination de l'origine d'un système d'axes (voir schéma du système de mesure), nous avons procédé à un certain nombre de mesures pour mettre en évidence la pharyngalisation des labiales dans les deux parlers étudiés.

Le corpus comprenait notamment les occurrences suivantes (les segments inclus dans la prosodie d'emphase sont soulignés):

Berbère (El-Aderj) :

[imtlit]	~	[emtlet]
[imda]	~	[emdɑ]
[ismda]	~	[esmdɑ]
[iʃmt]	~	[eʃmt]
[ini bsl da]	~	[ini bsl da]
[iblatn]	~	[eblɑtn]
[ini tfrit]	~	[ini tfret]

Arabe (Casablanca) :

[ʕli sæ:b bəzzæ:f]	~	[ʕli sɑ:b bəzzæ:f]
[ʕli sæ:m hna]	~	[ʕli sɑ:m hna]
	~	[wæ:ʃ sɑ:mu hna]

Chaque occurrence a été répétée dix fois au moins. Les mesures ont été effectuées sur l'image centrale correspondant à la réalisation du segment concerné, en se repérant d'après l'oscillogramme. Deux occurrences pour chacun des deux parlers analysés ont fait l'objet d'un suivi dynamique pour toutes les images de la séquence. Sur les graphes 1 à 4 (illustrant les données berbères) a été portée en ordonnée la valeur moyenne de la distance entre:

a. le dos de la langue et le palais dur (axes 45°, 60°, 75°)
b. le dos de la langue et le palais mou (axes 90°, 105°)
c. le dos de la langue et le voile du palais (axes 120°, 135°)
d. la racine de la langue et la paroi pharyngale (axes 150°, 165°, 180°).

Pour la paire coronale [t / ṭ] - prise comme référence de contraste d'emphase - et les paires de consonnes labiales [m/ṃ, b/ḅ , f/ḟ], on constate que l'axe 165° correspond à la constriction pharyngale maximale (partie médiane du pharynx). Il

s'agit donc bien d'une pharyngalisation, concomitante d'un élargissement de la cavité buccale.

Les croquis (2, 3, 4, 5) ont été dessinés d'après les mêmes moyennes, auxquelles sont adjointes les mesures concernant les lèvres supérieures (axes 0° et 15°), l'épiglotte (axe 195°) et l'os hyoïde (projection sur l'axe 90° et 180°): on constate avec la rétraction de la racine de la langue un léger recul de l'épiglotte et de l'os hyoïde vers la paroi pharyngale.

La frontière de mot:

Pour l'arabe tunisien, Ghazeli [3] écrit: "The fact that a word boundary prevents the spread of pharyngealization is a good reason to believe that this coarticulation phenomenon is neurally programmed in this dialect and not caused by mechanico-inertial factors. If it were caused by mechanical constraints on the articulations, the coarticulation of backing would have affected an adjacent segment in the following or in the preceding word".

Dans les parlers que nous étudions la prosodie de l'emphase peut aussi s'étendre à tout le mot, mais nous avons constaté que la pharyngalisation commence ou se prolonge sur le segment du mot adjacent, ce qui semble impliquer une plus grande importance du facteur d'inertie que ne l'a observé Ghazeli pour l'arabe tunisien. Les graphes 5 et 6 illustrent cette tendance. Pour l'énoncé berbère [i n i b s l d a] le i précédant la frontière de mot amorce le mouvement de pharyngalisation qui se prolonge également sur le d, pourtant les deux voyelles de [i n i] sont perçues comme identiques, ce qui confirme que la frontière de mot bloque la propagation de l'emphase même si la pharyngalisation déborde son domaine. De même pour l'énoncé en arabe de Casablanca [ʕ l i s ɑ:b b ə z z æ:f], le i amorce le mouvement de pharyngalisation (voir graphe 6).

La frontière syllabique:

Dans le disyllabe berbère [e m t l e t] contrastant avec [i m t l i t] (graphe 7), la qualité des voyelles correspond à une articulation emphatique de la voyelle i. La prosodie d'emphase est déclenchée par la coronale radicale (consonne médiane) ṭ. La pharyngalisation affecte la consonne labiale et s'étend à la voyelle qui la précède. Cette fois il ne s'agit plus d'une contamination imputable à l'inertie puisque la pharyngalisation est concominante d'une articulation principale rétractée; il faut donc distinguer coarticulation et propagation de pharyngalisation. En arabe de Casablanca au contraire, la frontière syllabique bloque la propagation progressive de l'emphase. Dans [s ɑ m u] (graphe 8) la première voyelle est emphatisée contrairement à la seconde. On constate sur le graphe 8 que le mouvement de constriction pharyngale est inversé pendant la production de la première voyelle [ɑ] ce qui suggère que le faible degré de pharyngalisation de la labiale est dû à un phénomène de coarticulation et non d'extension de l'emphase à ce segment. Au contraire pour [i n i b s l d a] nous pouvions conclure que la prosodie d'emphase inclut la labiale dans la mesure où le mouvement de pharyngalisation ne s'amorce pas sur ce segment mais sur la voyelle qui le précède.

Conclusion

Il ressort de cette étude que les consonnes labiales peuvent être pharyngalisées. Nous avons constaté par ailleurs que la coarticulation de la constriction pharyngale déborde le domaine de la propagation de l'emphase: il existe donc des phénomènes de compensation qui maintiennent la qualité de référence (non bémolisée) des voyelles, lors même que le mouvement de pharyngalisation les affecte.

REFERENCES

[1] Cohen D., "Les langues chamito-sémitiques", *Le langage*, volume publié sous la direction d'A. Martinet, Paris NRF, pp. 1288-1330, 1968.

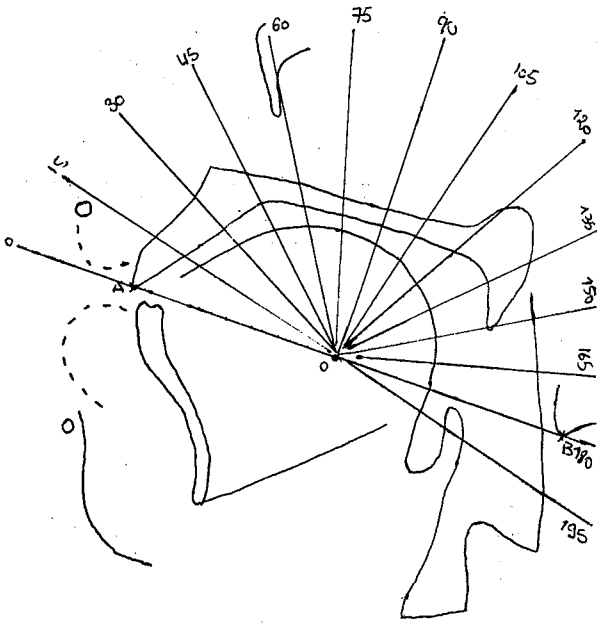
[2] Catford, J.C., *Fundamental problems in phonetics*, Indiana University Press, Bloomington, 1977.

[3] Roman A., *Etude de la phonologie et de la morphologie de la koinè arabe*, Tome 1, Université de Provence, 1983.

[4] Jakobson R., "Mufaxxama, The 'emphatic' phonemes in Arabic", in *Studies presented to Joshua Whatmough*, Pulgram ed., Mouton, The Hague, pp. 105-115, 1957.

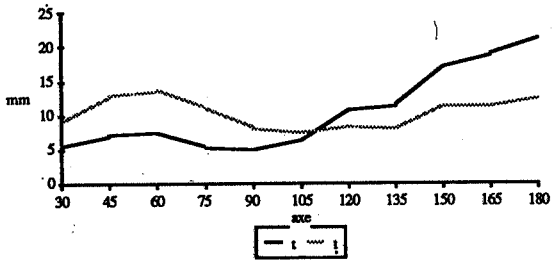
[5] Ghazeli S., *Back consonants and backing Coarticulation in Arabic*, Ph.D. Thesis, University of Texas, Austin, 1977.

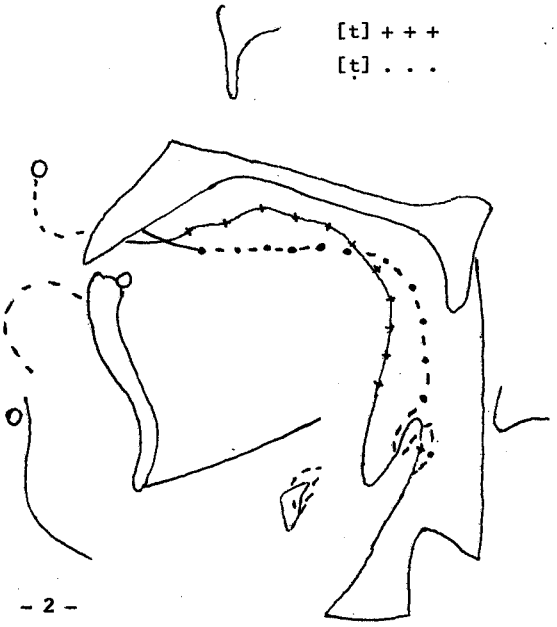
[6] Bothorel A., *Etude phonétique et phonologique du Breton parlé à Argol (Finistère Sud)*, Thèse 1978, Ed. Breizh, 1982.



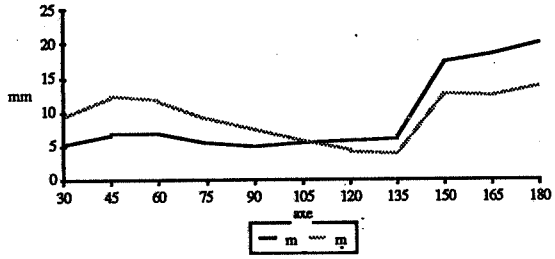
1 - Croquis du système de mesure

Graphe 1

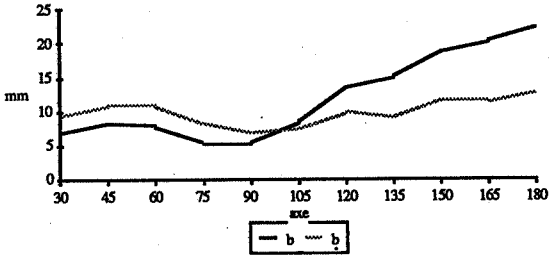


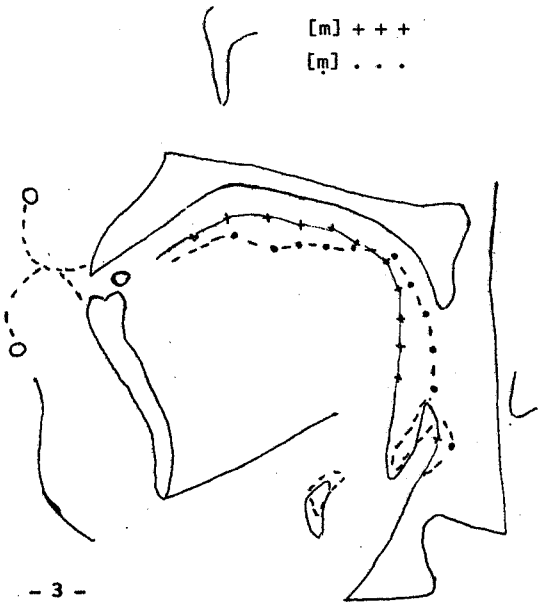


Graphe 2

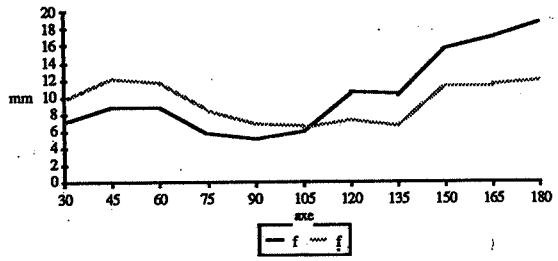


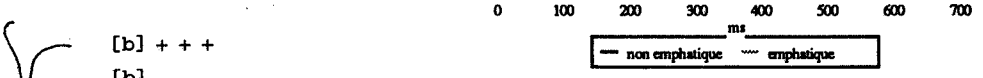
Graphe 3





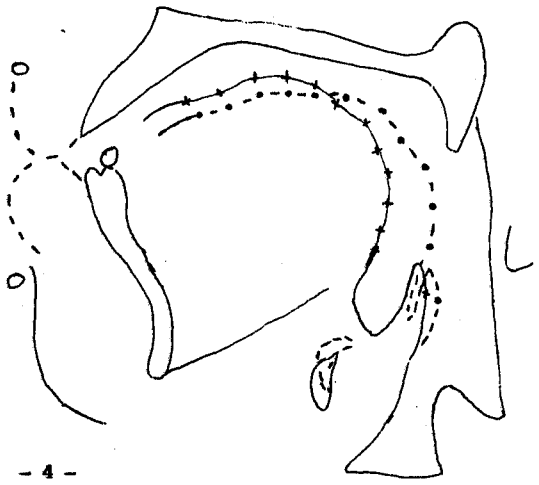
Graphe 4



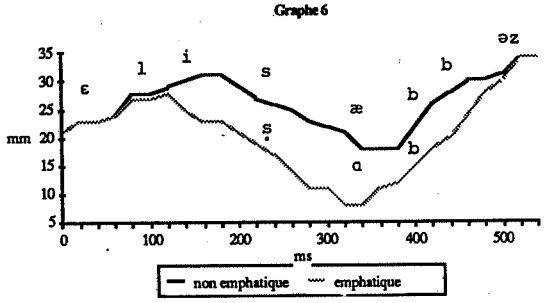


— [ini bsl da] ... [ini bsl da]

[b] + + +
[b] . . .

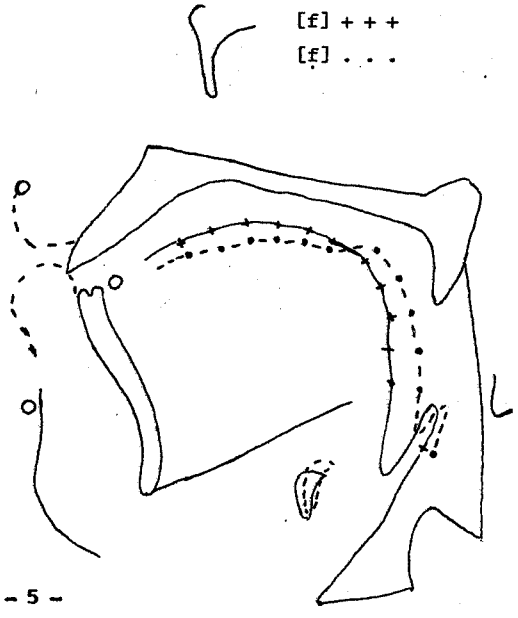


- 4 -

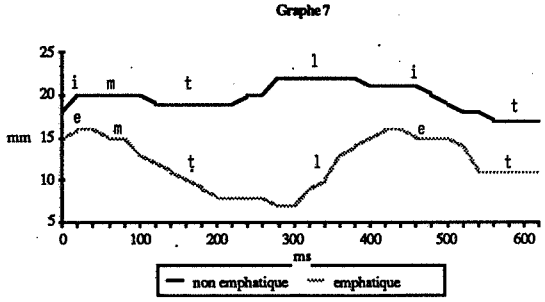


— [ɛli sæ:b bəzzæ:f] ...[ɛli sɑ:b bəzzæ:f]

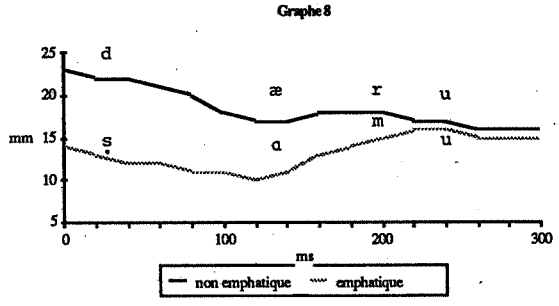
[f] + + +
[f] . . .



- 5 -



— [imtlit] ... [emtlet]



— [dæru] ... [sɑmu]

SYNTHESE ET PERCEPTION DE L'ACCENT LEXICAL EN ARABE

A. Rajouani, D.Chiadmi, M.Najim et M. Ouadou

LEESA -Faculté des sciences -B.P. 1014 -Rabat-Maroc

ABSTRACT

The experiments reported in this paper are an attempt to synthesize a natural lexical stress in Arabic. In the first experiment we investigate the effects of the fundamental frequency, the intensity and the duration on the perception of the stress. Test responses nalyzed in relation to the various parametric values present in each stimulus, consistently pointed to duration as the less pertinent cue. The 2nd experiment is devoted to find the optimal combination of the acoustic parameters values to synthesize a natural stress depending on the quantity of the stressed vowel and the type of the embedding syllable. The 3rd expriment deals with the determination of the place of the lexical stress in the CVCCVCVCV sequences. The results obtained basing on a preference test showed that the stress is onto the first syllabae.

SUR L'EXISTENCE DE L'ACCENT LEXIAL ET LA SYLLABIFICATION

1/Existence de l'accent lexical.

Existe-t-il un accent lexical, ou accent de mot, en Arabe ? La question a longtemps constitué un sujet d'intérêt et de controverses avant que la réponse par l'affirmative ne soit admise. Les arguments avancés par les chercheurs niant l'existence de l'accent lexical en Arabe sont parfaitement resumés par la citation de Fleish /1/ 'Comment est ce possible de parler de l'accent alors qu'ils ne joue aucun rôle distinctif dans la langue; les grammairiens arabes l'ont complètement ignoré...' .La position de Ghalib /2/ est plus subtile, l'accent existe en Arabe mais sa pertinence est beaucoup moindre que dans l'Anglais et l'Allemand, ainsi selon l'auteur le deplacement de l'accent d'une syllabe à une autre ne change ni le sens du mot ni sa fonction grammaticale même si un tel deplacement peut deformer la prononciation correcte du mot. Abdou /3/, Brame/4/ et Bohas /5/ réfutent tous les arguments cités ci-dessus et demontrent l'existence de l'accent en Arabe. Nous retenons notamment que ces auteurs rapportent des exemples, certes peu nombreux, où le rôle distinctif de l'accent au niveau semantique est irréfutable.

Parmi les études instrumentales attestant l'existence de l'accent lexical en Arabe, nous citerons Al Ani /6/, et Belkaid /7/.

2/ Structure syllabique et syllabification.

L'existence de 6 types de syllabes peut être postulée en Arabe /8/: CV, CVV, CVC, CVCC, CVVC. De très nombreux travaux ne considérent cependant que 5 types en négligeant la structure CVVCC. Nous sommes plutôt d'accord avec ce point de vue au niveau phonétique. En effet la structure CVVCC ne pourrait apparaitre qu'en position finale en forme pausale et à condition que les deux consonnes finales soient identiques. Les analyses spectrographiques que nous avons realisées montrent qu'une consonne geminée en position pausale est realisée acoustiquement comme une seule et unique consonne /9/.

Les syllabes sont decrites en terme courte/longue, ouverte/fermée :

	ouverte	fermée
Courte	CV	
Longue	CVV	CVC, CVVC, CVCC

Les régles admises pour la syllabification d'un mot sont :

- Le nombre de syllabes est égal au nombre de voyelles, la voyelle étant le noyau de la syllabe,
- Le premier élèment de la première syllabe coincide avec le debut du mot. Cet élèment ne peut être qu'une consonne et une seule.
- Le coda de la dernière syllabe coincide avec la fin du mot. Le coda est soit une voyelle brève ou longue, soit une ou deux consonnes.

CORRELATS ACOUSTIQUES DE L'ACCENT

1.Position du problème.

La description de l'accent lexical en Arabe reste dans la majorité des travaux intuitive et partielle. Pour Ghalib /2/, les syllabes accentuées tendent à être plus hautes ou plus intenses que les syllabes non accentuées. Concernant la durée l'auteur repporte que l'accent est en relation avec la durée de la syllabe. Bohas et Kouloughli /10/ suggèrent que les syllabes bréves accentuées sont plus fortes, ce qui leur donne des propriétés phonétiques et phonologiques que les brèves non accentuées n'ont pas (p 34). Le travail de Belcaid /7/, rapporte à partir d'une étude instrumentale des voyelles de l'Arabe, une description qualitative relativement plus precise 'toutes les voyelles accentuées de notre corpus, comparées aux inaccentuées, sont intenses quels que soit leur timbre et leur durée. Un pic d'intensité manifeste leur presence'(p230). L'auteur ajoute aussi que les voyelles accentuées n'ont pas une durée plus importante que les voyelles non accentuées (p 231). Aucune indication sur le comportement de la fréquence fondamentale des voyelles accentuées n'est signalée.

2. Etude instrumentale de l'influence de FO, de l'intensité et de la durée des voyelles sur la perception de l'accent lexical.

a/ Objectif et considérations expèrimentales.

Notre objectif consiste à dégager à partir des tests de perception l'influence des trois paramètres acoustiques précités sur la perception de l'accent lexical en Arabe. Les tests de perception sont realisés sur un ensemble de stimuli synthétisés par règles utilisant un synthétiseur à formants /11/, les voyelles de chaque stimuli étant réalisées avec une combinaison différente de valeurs

pre-définies des paramètres étudiés. La determination des plages de variation adequates pour chaque paramètre demeure encore un problème très délicat /12/,/13/, /14/, /15/. Les intervalles que nous avons adoptés pour chaque paramètre ont été définis à partir de tests de perception preliminaires lors de chaque expérience. Les pas d'incrementation à partir des valeurs references (120 Hz,55 DB,80 ms) sont arbitraires.

Rappelons qu'en Anglais, par exemple, la position de l'accent peut differencier deux mots constituant une paire minimale au sens strict (contrat/contract, torment/torment.....). Sans s'étendre sur la définition du mot en Arabe, nous précisons que deux élèments d'une paire minimale, au sens de la place de l'accent ne peuvent être que de la forme suivante : A/B+A, où A est un mot sans prefixe et B est un prefixe tel ka, fa, wa, ... Parmi les paires minimales du lexique nous avons choisi /fa'alaa/ et / fa'alaa/, où fa de la deuxième séquence est un prefixe. Deux raisons ont justifié ce choix : la qualité acoustique des stimuli synthétisés et la fréquence relativement grande des deux séquences dans le vocabulaire usuel. Nous précisons que les valeurs des formants sont identiques pour tous les stimuli, que la variation de l'intensité est realisée par l'intermédiaire des gains des sources d'excitation et que la variation de FO est réalisée de la manière suivante :

FO référence A ——— C ——— B. FO référence
FO. max

où AB est la durée de la partie stable de la voyelle et AC est égal au 2/3 de AB.

b/ Expérience 1.

Le but de cette expérience est d'étudier l'influence des variations combinées des paramètres fréquence fondamentale FO et intensité I des voyelles sur la perception de l'accent lexical. Les stimuli sont réalisés selon deux modalités.

Pour 50 % des stimuli, les variations des paramètres sont appliquées à la première voyelle de la séquence /fa alaa/, alors que pour l'autre moitié des stimuli les variations sont relatives à la deuxième voyelle. La plage de variation est fixée pour FO dans l'intervalle 120-170 Hz , alors que l'intensité peut être augmentée jusqu'à +8 dB par rapport à la valeur référence. Les durées de tous les segments de la séquence restent constantes. Les pas d'incrementation utilisés pour chacun des paramètres au cours de cette expérience sont :
- +0,+15,+20,+25,+30,+35 Hz pour FO,
- +0,+2,+4,+6,+8 dB pour E.

Les 59 stimuli synthétique sont enregistrés dans un ordre aléatoire sur bande magnetique avec une durée de silence interstimuli égale a 2 secondes. L'ensemble des stimuli a été écouté separément, dans une salle relativement calme, par chacun des 8 sujets marocains participant a l'expèrience. Chaque sujet était muni d'une feuille de réponse avec une en-tête à quatre colonnes : numéro de stimuli/ils ont fait / et il a pris de l'altitude/ hesitation (en français sur la feuille). Les indications données aux sujets sont ainsi formulées "Pour chaque stimuli, cochez une croix dans la colonne 2 (resp.3) si vous comprenez 'il ont fait' (resp. 'il a pris de l'altitude')". En cas d'hésitation ajouter une deuxième croix dans la 4ème colonne
L'exploitation des résultats est basée sur une analyse de variance realisée sur les réponses codées sur 4 niveaux. Les ésultats de l'experience 1 sont presentés dans la table 1.

	ddl	SC	CM	F	Prob
I	4	16.36	4.09	3.65	.05
FO	5	28.15	5.63	5.02	.001
INTER	20	44.41	2.22	1.97	.01
RESIDU	450	506	1.21		
TOTAL	479	594.92			

Table 1. Résultats de l'analyse de variance pour l'expérience 1.

c/ Expérience 2.

Cette expérience est realisée de la même manière que l'experfence 1, excepté que le paramètre intensité est remplacé par le paramètre durée D. Les pas d'incrementation pour la durée sont en ms : 0, 10, 20, 30, 40, 50. Les variations de l'intensité par rapport à la séquence inaccentuée sont nulles.
Les résultats sont présentés dans la table 2.

	ddl	SC	SM	F	Prob
FO	5	42.55	8.51	7.12	.001
D	5	15.70	3.14	2.62	.01
INTER	20	40.84	1.63	1.36	N.S
RESIDU	540	644.88	1.21		
TOTAL	575	743.97			

Table 2. Résultats de l'analyse de variance pour l'expérience 2.
(N.S. = non significatif).

d/ Expérience 3.

Nous étudions dans cette experience l'influence de la combinaison des paramètres intensité et durée. Les pas d'incrépentation sont :
- +0, +10, +20,+30,+40,+50 ms pour D.
- +0, +2, +4, +6, +8 dB pour I.

Les résultats sont présentés dans la table 3.

	ddl	SC	CM	F	Prob
I	4	41.27	10.31	8.52	.001
D	5	18.66	3.73	3.08	.05
INTER	20	39.57	1.97	1.62	.01
RESIDU	450	544.56	1.21		
TOTAL	479	644.08			

Table 3. Résultats de l'analyse de variance de l'expérience 3.

e/ Conclusion et commentaires.

Les résultats de l'expérience 1 montrent que l'intensité est moins significative que FO, alors qu'après les résultats de l'expérience 3 l'intensité est plus significative que la durée. Nous suggérons que la hiérarchie des 3 paramètres pour la perception de l'accent est dans l'ordre suivant : FO,I,D. Cette hiérarchie n'est pas conforme à la hiérarchie 'universelle' FO,D,I, suggérée par Hymann /16/. Cette non conformité a été déjà montrée pour certains dialectes /17/ qui ont la particularité, à l'image de l'Arabe, d'avoir un système vocalique basé sur une opposition de quantité (breve/longue). McCarthy /18/ suggère aussi que 'Particular languages have perceptual or articulatory reasons for keeping some class of vowels short, so these vowels resist stressing, since it would inevitably cause them to length. Probably the most common of this is a language with

phonemic length. Since lenghtening of short vowels under stress would tend to neutralize this lenght contratst, the, speaker has a perceptutal motivation to draw stress away from short vowel and maintain the phonemic contrast (p 444)".

3/ Syntèse de l'accent lexical.

a/ Matériau linguistique.

Notre objectif est de determiner de manière quantitative la combinaison des paramètres FO,E,D, suceptible de créer, au niveau perceptif, le meilleur naturel de l'accent. La matériau linguistique est constitué de trois mots réels /wa'ada/ (promettre), / saa'ada/ (aider), / ma'bada/ (temple). Parmi les raisons de ce choix, nous signalons que :

- La place de l'accent est connue avec certitude pour ces types de séquences.
Nous avons ainsi éliminé les prefixes et les suffixes et toutes les autres structures syllabiques objets de controverses (voir paragraphe suivant),
- Les syllabes CV, CVV, CVC, accentuées dans chacune des séquences sont d'une part les plus fréquentes dans le lexique, et d'autre part permettraient par les oppositions qu'elles présentent (courte/longue, ouverte/fermee) d'envisager la généralisation des résultats obtenus,

- Les trois séquences sont strissyllabiques, et la consonne suivant la voyelle accentuée est toujours /'/. Ces précautions sont prises afin de neutraliser les varitions temporelles dûes au nombre de syllabes et à la nature de la voyelle post-vocalique.

b. Procédure expérimentale et résultats

Des tests preliminaires ont permis de dégager, pour chaque type de syllabe, un premier ensemble de combinaisons suceptibles de créer un accent acceptable. Ces combinaisons sont rapportées dans les tableaux 4,5 et 6. Pour la séquence / wa'ada/, par exemple, 15 stimuli sont synthétisés correspondant aux 15 combinaisons (P,E,D) pre-établies. Le test de préférence consiste à comparer 2 a 2 l'ensemble des stimuli {a,b,...} . Le format de chaque combinaison est (a) pause (ba) (b pouvant être similaire à a), a et b étant ainsi l'étalon une fois sur deux. Chacun des 3 sujets informateurs devait indiquer lequel des deux stimuli il préfère.

Le score de préférence Pref (i) est obtenue par la formule suivante :

$$Pref\ (i) = (20/\pi)\ \sin\left[f\ (i)/\ (n-1)\right]$$

où f (i) est le nombre de fois que la condition a été préferée par rapport aux autres (n-1) conditions. La constante $20/\pi$ assure que le score de préférence est compris entre 0 et 10 (voir pour détails /19/ et /20/).

Les résultats des tests de préférence sont rapportés dans les tables 4,5 et 6. Nous n'avons représenté dans chaque table, pour plus de clarté, que les résultats relatifs aux 5 meilleurs scores moyens et le score moyen le plus faible.

D	10	20	20	20	20	20
I	6	2	4	6	6	6
P	20	15	15	15	20	25
Sujet1	7.5	10	5.9	5.9	4.5	1.7
2	3.6	5.5	10	4.5	6.4	4.5
3	4.5	6.4	5.5	8.3	4.5	3
Moy.	5.2	7.3	7.13	6.23	5.13	3.06

Table 4. Résultats du test de préférence pour / wa'ada/.

D	10	10	10	10	20	20	20
I	2	2	4	6	2	4	6
P	15	20	15	15	15	15	20
Sujet1	2.2	5.9	8.1	5.9	4.7	7.2	1.9
2	8.I	5.3	8.1	5.9	5.3	8.1	1.9
3	7.2.	4.7	7.2	5.9	4.7	7.2	1.9
Moy	7.5	5.3	7.8	5.9	4.9	7.5	1.9

Table 5. Résultats du test de préférence pour / saa'ada/.

D	10	10	10	10	10	10	20
I	2	2	4	4	6	6	6
P	15	20	15	20	15	20	20
Sujet1	5.9	5.3	4.7	7.2	5.9	6.5	1.9
2	8.1	7.2	7.2	7.2	6.5	3.5	0
3	8.1	7.2	6.5	7.2	6.5	5.9	1.9
Moy	7.36	6.56	6.13	7.2	6.3	5.3	1.26

Table 6. Résultats du test de préférence pour / ma'bada/.

Parmi les remarques qui se dégagent de la lecture de ces tableaux, nous signalons que :

- pour les 3 séquences, le score de préférence le plus bas est ralisé par la combinaison correspondant aux valeurs maximales des 3 paramètres, les effets des 3 paramètres seraient ainsi cumulatifs.

- dans les 3 cas, le meilleur score est realisé pour la même valeur de Fo.Ce paramètre serait ainsi le plus pertinent pour la perception de la qualité de l'accent.
- la voyelle brève accentuée est plus courte en syllabe fermée. Céla confirme nos résultats preliminaires sur la durée des voyelles de l'Arabe.

- la voyelle longue accentuée est plus intense que la voyelle brève accentuée.

PLACE DE L'ACCENT

1/ Position du problème.

Différentes règles pour l'assignation de l'accent lexical sont presentées dans la littérature /3/, /4/, /6/, /8/. Quatre faits pertinents se dégagent de l'examen de ces régles :

- le nombre de types d'accent diffère selon les auteurs,

- aucune suggestion n'est validée (à notre connaissance) par synthèse ou toute autre méthode reproductible.

- l'automatisation de la plupart des règles implique de développement d'un analyseur morphologique, et d'un traitement ad-hoc pour les exceptions.

- Les schémas accentuelle obtenus différent fréquemment selon l'ensemble de règles appliquées. L'exemple suivant est parfaitement revélateur :

Séquence	règles appliquées	Place de l'accent principal
CVCCVCVCVCV (ex:maktabatuhu)	Al Ani/6/, Mc Carthy /18/ Abdo /3/	CVCCVCVCVCV
	Selim et Anbar /18/	CVCCVCVCVCV

Nous nous proposons dans ce travail d'étudier un aspect qui semble être le sujet majeur des controverses : quelles sont les limites de la remontée de l'accent primaire vers le début d'une séquence de type CVCCVCVCV ? En d'autres termes, l'accent peut-il être assigné à la première syllabe CVC dans ce type de séquence ? McCarty répond par l'affirmative 'Classical Arabic allows retraction of stress a potentially infinite distance from the right boundary' (p 461), alors que Abdo /3/, Brame /4/, Salem et Anbar /18/ Dohas et Koulougli /10/ suggèrent que l'accent ne remonte jamais au delà de l'antépénultième et par conséquent' dans les mots polysyllabiques dont la penultième n'est pas une lourde (et

dont les deux dernières ne sont pas surlourdes, bien sûr), c'est toujours l'antepénultième qui est accentuée' /10/. Notre objectif est d'apporter des éléments de réponse basés sur la perception de l'accent par des auditeurs marocains.

2/ Expérience et résultats.

Le corpus soumis au test de préférence est constitué de 8 stimuli synthétiques, 4 correspondant à /máf'alata/ et 4 à / maf'álata/. Les combinaisons (P,E,D) utilisées pour la synthèse de /máf'alata/ sont les 4 combinaisons ayant realisé les 4 meilleurs scores pour la synthèse de /ma'bada/, la syllabe accentuée dans les deux cas étant de type CVC. Le même raisonnement est reproduit pour / maf'álata/ avec les paramètres de /wa'ada/, la syllabe accentuée étant CV. Le test de préférence est réalisé comme dans les trois expériences précédentes avec 3 sujets. Les résultats sont présentés dans le tableau 4 :

	stimuli / máf'alata/				stimuli /maf'álata/			
D	10	10	10	10	20	20	20	20
E	2	4	2	6	2	6	4	4
P	20	20	15	15	15	15	15	20
Sujet 1	6.42	5.45	5.45	5.45	3.59	2.46	0	0
2	6.41	6.41	5.45	5.45	2.46	0	0	0
3	5.45	5.45	5.45	5.45	0	0	0	0
4	5.45	6.41	6.41	5.45	2.46	0	0	0
5	7.53	6.41	5.45	6.41	0	0	0	0
Moy	6.24	6.02	5.64	5.64	1.70	0.49	0	0

Table 4. Résultats du test de préférence.

Les résultats montrent clairement que les sujets préferent la séquence /máf'alata/ . L'accent principal serait ainsi assigné à la 1ère syllabe de la séquence. Nous suggérons par conséquent que l'hypothèse sur les limites de la remontée de l'accent principal est infirmée dans le cas des sujets à substrat dialectal marocain.

CONCLUSION ET PERSPECTIVES

Les deux conclusions degagées à partir de ce travail sont :
- la fréquence fondamentale (rep. la durée) est le paramètre le plus (rep. le moins) pertinent pour la perception de l'accent lexical en Arabe.

- l'accent lexical principal peut remonter au delà de l'antepénultième. Des études preliminaires sur l'intonation en Arabe /22/ ayant montré que le calcul du contour intonatif peut être deduit à partir de la place de l'accent lexical principal, nous suggérons un examen critique basé que des méthodes instrumentales de l'ensemble des règles d'accentuation présentées dans la littérature.

Références.

1/ H. Fleish, Traité de philologie arabe. Imprimerie Catholique, Beyrouth, 1961.

2/ M. Ghalib, 'Etude de quelques aspects de l'intonation en Arabe', Revue de l'Université de Dassorh, 10,pp :198-228 (en arabe).

3/ D. Abdo, On stress and Arabic phonology, a generative approach.Khayats, Beyrouth,1969.

4/ M.Brame, 'Stress in Arabic and generative phonology'. Foundations of Language,7,pp:556-591, 1971.

5/ G.Bohas,'Peut on parler de l'accent en Arabe classique'.Actes du colloque Recherches Linguistiques et Semiotiques, Rabat 1981, pp:165-173. (en arabe).

6/ S. Al Ani, Arabic Phonology. Mouton, The Hague, 1970.

7/ Y.Belkaid,' Les voyelles de l'Arabe littéraire moderne. Analyse spectrographique'.T.I.P. de Strasbourg, 16,pp: 217-240,1984.

8/ S. Al Ani et D.R. May, 'The phonological structure of the syllabe in Arabic'. Américain Journal of Arabic studies,1,pp: 37-49,1973.

9/ A. Rajouani, M. Najim, D. Ciadmi et M. Ouadou, 'Etude de la geminations des occlusives en Arabe'. Actes GALF, 15ème J.E.P.,pp: 16-18, 1985.

10/ G. Bohas et Kouloughli. 'Processus accentuels en Arabe'. Analyse/Théorie,1, pp: 1-59, 1981.

11/ H.Klatt, 'A software for a cascade/parallel formant sunthesizer'. J.A.S.A. 67,pp:971-995,1980.

12/ D.B.Fry, 'Duration and intensity as physical correlates of linguistic stress'.J.A.S.A.,27,pp: 765-768, 1955.

13/ D.B. Fry, 'Experiments in the perception of stress'. Language and Speech,1,p: 126-152,1958.

14/ J.Morton et W. Jassem, 'Acoustic correlates of stress'. Language and Speech, 8,pp: 159-180,1965.

15/P.M. Bertinetto,' The perception of stress by Italian speakers'. Jour. of Phonetics, 8,pp: 385-395,1980.

16/ L.M.Hymann, 'On the nature of linguistic stress' in Hyman ed. Studies in Stress and Accent,Sou. Calif. Occasional Papers in Linguistics,4,pp: 37-82.

17/ A.E.Berinstein, 'A cross-linguistic study on the perception and production of stress'. UCLA Working Papers in Phonetics,47,pp:1-59,1979.

18/ J.J. McCarthy, 'On stress and syllabification'. Linguistic Enquiry,10,pp: 443-465,1979.

19/ H.A. David,The method of paired comparisons, Hafner.Pub.Cie.,New York, 1963.

20/L.R. Rabiner, M. Levitt et A.E. Rosenberg, 'Investigation of stress patterns for speech synthesis by rule'.J.A.S.A, 45,pp: 92-101,1969.

21/ H. Selim et T.Anbar, 'A phonetic transcription system for Arabic text'. Proc IEEE-ICASSP, Dallas, pp: 1446-1449, 1987.

22/ L. Esskali, A. Rajouani, M. Najim et D. Chiadmi, 'Eléments d'un modèle intonatif pour la phrase affirmative en Arabe'. Actes GALF, 16J.E.P., Hammamet, 1987 (à paraître).

JOURNEE EVALUATION BASES DE DONNEES

16e Journée d'Étude sur la Parole - Société Française d'Acoustique - Hammamet, 5-9 oct 1987.

EVALUATION D'UN SYSTEME DE RECONNAISSANCE VOCALE DANS DES TACHES DE CONTROLE AERIEN

CHRISTINE BAILLEUL (CENA)

Centre d'Etudes de la Navigation Aérienne
B.P. 205 Orly 94542 Orly-Aérogare

In cooperation with the LIMSI, the CENA has been working on a vocal post prototype, capable of speech synthesis and recognition, which allows the direct voice piloting of an air trafic simulator. To asses the performances of its recognition system, the CENA has undertaken a series of tests, at first independantly from any operationel context, and then in a dynamic work context. The evaluation required the definition of methods and criteria of assessment. The uncorrected recognition rates of error can be compared with results coming from different studies; others rates are more relevant for the CENA but they are so dependant from the context of the task that they don't have any signification whithout being refered to the application.

INTRODUCTION

Lors d'une séance de simulation de contrôle du trafic aérien, l'élève contrôleur se trouve dans une situation proche de la réalité, et communique ses instructions à un pseudo-pilote qui assure un dialogue oral et l'entrée dans le simulateur des données correspondant aux instructions.

Le CENA étudie depuis 1983, en collaboration avec le LIMSI, une maquette de poste vocal (1) comprenant reconnaissance et synthèse vocale, dont le but est d'assurer la fonction pilote d'une simulation. La reconnaissance permettrait de tirer directement de la parole du contrôleur les données pour la machine, et une voix synthétique remplacerait celle du pilote.

Toute étude d'un système de reconnaissance vocale passe par de phases d'évaluation de performances, qui fournissent divers taux de reconnaissance. La comparaison de résultats provenant d'études de même nature n'est pas aisée, si l'on ne précise pas les critères utilisés, ni les conditions du déroulement de séances d'évaluation.Dans cette optique, mais aussi afin de suivre la progression de l'étude, le CENA a défini des méthodes et des cadres d'évaluation des performances du système de reconnaissance vocale. Les performances du terminal vocal sont mesurées lors de séances de tests spécifiques, puis en contexte dynamique de travail. Enfin, le CENA effectue des mesures en cours de simulation.

I-SEANCES DE TESTS EN MODE AUTONOME

Une séance de test a pour but de mesurer les performances du terminal vocal en mode autonome. Elle consiste à faire prononcer, en dehors de tout contexte opérationnel, des séries de phrases par plusieurs locuteurs naïfs, trois ou quatre hommes et femmes non familiarisés avec le système de reconnaissance. Chaque séance est liée à une syntaxe et un vocabulaire donnés, faisant en général partie du langage opératif utilisé pour les échanges contrôleur-pilote.

1-a Déroulement d'une séance de test

Le terminal étudié est monolocuteur; un apprentissage préalable du vocabulaire doit être effectué par chaque locuteur. Une séance se déroule en deux phases, et son déroulement est suivi par un opérateur dont le but est d'entrer à certains moments des données pour le fichier d'archivage élaboré pendant la séance.

Durant la première phase d'environ une heure par locuteur, chaque locuteur effectue un premier apprentissage de tout le vocabulaire, puis prononce quelque phrases afin de se familiariser avec le système. Le locuteur alors plus décontracté effectue immédiatement un second apprentissage qui sera retenu pour la seconde phase.

Au cours de la deuxième phase , d'une durée d'une demi-heure par locuteur, chaque locuteur lit des phrases affichées sur un écran. Il peut éventuellement être informé du résultat de la reconnaissance. Les phrases présentées sont choisies aléatoirement parmi le corpus des phrases autorisées par la syntaxe. L'opérateur qui suit le déroulement de la séance archive si besoin est le type d'erreur: ajout, confusion, élision, ou rejet de mot.

1-b Taux de reconnaissance

On distingue les performances brutes de la reconnaissance des performances obtenues après la mise en application de processus de compréhension.

Les performances brutes de reconnaissance proviennent de la comparaison de la phrase prononcée à la phrase fournie en sortie du terminal, sans aucun traitement. Les couches supplémentaires de traitement qui interviennent alors -détection des indicatifs, mécanismes d'instanciation de schémas- mettent en jeu des processus de "compréhension" propres à l'application.

Leur apport est mesuré en se référant aux performances brutes du terminal vocal.

Les taux bruts de reconnaissance:

Les taux définis ci-aprés s'appliquent à des mots enchaînés, pour des phrases prononcées séparément les unes des autres.

Taux d'erreur par mots:

On estime que lorsqu'une erreur intervient dans une phrase, les mots fournis après erreur ne sont pas nécessairement en relation avec les mots prononcés. C'est pourquoi ne sont pris en compte que les mots reconnus avant erreur, y compris celui sur lequel a porté la première erreur. Le taux d'erreur par mots est le rapport du nombre de mots incorrects au nombre de mots prononcés avant erreur. On calcule des taux partiels d'erreur par mots, relatifs à un mot du lexique, à un type d'erreur, à un locuteur, ou à une combinaison de ces critères.

Taux bruts d'erreur de reconnaissance par phrases:

On considère qu'une phrase est bien reconnue si tous les mots qui la compose sont correctement reconnus. La phrases affichée sur l'écran n'est prononcée qu'une seule fois. La première erreur qui intervient dans la phrase donne le type d'erreur (ajout, confusion, élision ou rejet). Le taux brut d'erreur par phare est le rapport du nombre de phrases non reconnues au nombre de phrases prononcées. On calcule des taux partiels par locuteur et par type d'erreur.

Evaluation de la qualité de la compréhension:

La qualité de la compréhension est évaluée par le taux d'échec après compréhension: c'est le rapport du nombre de phrases pour lesquelles le processus de compréhension a échoué au nombre de phrases prononcées.

La notion d'échec de la compréhension est spécifique de l'application envisagée. Pour la définir, le contexte de la tâche est indispensable. Ainsi, pour la détection des indicatifs d'appel radio (2) on a défini plusieurs catégories d'échec. Dans un message émis par un contrôleur l'indicatif désigne l'aéronef destinataire de ce message. La compréhension des indicatifs repose sur le fait que l'on dispose à un moment donné d'un nombre limité d'indicatifs susceptibles d'être appelés par le contrôleur: ce sont les vols en compte. La reconnaissance fournit des éléments d'indicatifs, à savoir le nom de compagnie, le numéro de vol, et des lettres. A chaque indicatif de la liste des vols en compte est associé un score, fonction du nombre d'éléments que contient cet indicatif. Le résultat de la compréhension est un indicatif isolé, ou une liste des indicatifs, ayant obtenu le meilleur score. On obtient quatre types d'échec ,selon que le résultat du processus de compréhension existe et est un indicatif isolé mais incorrect, une liste d'indicatifs où figure l'indicatif prononcé, ou une liste ne contenant pas cet indicatif. A ces quatre types d'échec correspondent naturellement quatre taux d'échec. Dans l'hypothèse où l'on sélectionne un indicatif d'une liste, les taux d'échec relatifs aux listes prennent leur importance. Suivant que la liste contient ou non le bon indicatif, on peut espérer "corriger" les erreurs de la reconnaissance, ou au contraire on va accentuer le taux de fausse information. Ce taux de fausse information est un renseignement essentiel, puisqu'il donne la proportion de données cohérentes mais erronées introduites dans le simulateur, sans que rien dans l'état actuel de l'application ne permette de les détecter.

1-c Utilisation des taux de reconnaissance:

Trois taux caractérisent une séance:
-taux brut d'erreur par mots
-taux bruts d'erreur par phrases
-taux d'échec de la compréhension.

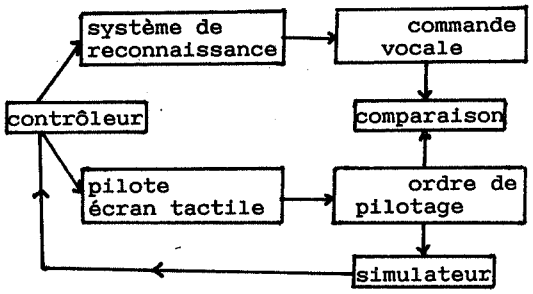
Les deux premiers taux mesurent les performances de la reconnaissance pure. Le taux de reconnaissance par phrases est un indicateur "pessimiste" puisqu'il suffit d'un seul mot non reconnu dans une phrase pour qu'elle soit déclarée erronée. Ces taux sont des taux de référence, permettant de comparer différentes syntaxes, dont on connaît par ailleurs le degré de complexité, par la taille du vocabulaire et les facteurs de branchement. Les taux partiels par locuteur, dont il est intéressant d'étudier les variations -ou la stabilité- selon les différentes syntaxes, font apparaître que certains types d'erreurs sont caractéristiques d'un locuteur. On peut envisager d'adapter certains paramètres du système de reconnaissance au locuteur.

Le dernier taux est le plus pertinent pour caractériser les performances du système de reconnaissance, ou plus exactement l'apport des processus de compréhension à la reconnaissance pure. Ses divers composants,dont les taux partiels de non information ou de fausse information, sont des renseignements essentiels du point de vue de l'utilisateur. Mais contrairement aux deux taux précédents, le taux d'échec de la compréhension n'a de signification que relativement à l'application considérée. Indiscociable du contexte de la tâche, il ne peut avoir de portée en dehors d'elle.

II-EVALUATION EN CONTEXTE DYNAMIQUE DE TRAVAIL

On fait fonctionner de manière autonome le système de reconnaissance vocale sur des phrases prononcées par les élèves contrôleurs en cours de simulation de formation au contrôle aérien. On compare les phrases reconnues aux ordres de pilotage transmis au simulateur par les pseudo-pilotes au moyen d'une interface de type écran tactile. Ces ordres de pilotage sont une traduction directe des instructions des contrôleurs.

Schéma de l'expérimentation :



On obtient deux fichiers d'archivage, correspondant tous deux à une interprétation des instructions prononcées par le contrôleur. De leur comparaison proviennent des taux pratiques de reconnaissance, globaux ou spécifique d'une commande vocale.

On peut effectuer des séances d'évaluation sur un échantillon quelconque de trafic aérien, du moment que la syntaxe utilisée pendant l'exercice cadre avec les possibilités du système de reconnaissance.

III-EVALUATION DES PERFORMANCES DU SYSTEME DE RECONNAISSANCE VOCALE EN COURS DE SIMULATION

Une fois terminées les campagnes d'évaluation du terminal vocal, en mode autonome puis en contexte dynamique de travail, le CENA envisage d'évaluer l'influence de l'utilisation de techniques de synthèse et reconnaissance de la parole sur le déroulement d'une simulation du contrôle aérien. IL sera possible de se référer aux performances du terminal qui auront été préalablement déterminées.

Les sources d'erreur sont multiples, et une classification s'impose. On distingue notamment les erreurs propres à la reconnaissance de la parole et que le processus de compréhension n'a pas écartées des erreurs propres au locuteur. Les cas de fausse information ou de non communication d'information, peuvent avoir pour origine une défaillance de la reconnaissance, du type ajout, élision, confusion, ou rejet de mots. Mais ces erreurs peuvent aussi être provoquées par le locuteur qui "trébuche" dans la prononciation du message, hésite, ou ne respecte pas la syntaxe . On doit aussi tenir compte des erreurs causées par les conditions d'enregistrement du signal sonore: bruits brefs, mauvaise position du microphone ou encore mauvaise manipulation de l'interrupteur que le contrôleur doit positionner lorsqu'il s'adresse au pilote (alternat).

Si un certain nombre de ces erreurs peuvent être déduites des données enregistrées, d'autres doivent être codées et introduites par un opérateur disposant de tous les éléments permettant de décider rapidement du cas d'erreur.

On peut craindre -à juste titre- que les erreurs soient difficiles à isoler et identifier. D'autre part, la notion d'échec de la compréhension est complexe. De même que l'on n'évalue pas le "degré de réussite" d'une simulation classique où intervient un opérateur pseudo-pilote et une interface homme-machine, comment s'appuyer sur une notion claire "d'échec de la compréhension", pour définir un taux de compréhension ? On peut en revanche définir quelques indicateurs facilement mesurables, tels que l'indicateur de fausse information. Cet indicateur rend compte du nombre d'instructions "mal comprises", c'est-à dire ayant donné lieu à des ordres de pilotage corrects mais non conformes aux instructions prononcées, dans la mesure où ce qu'a voulu exprimer le contrôleur est valide. On écarte ainsi pour le calcul de ce taux les phrases où la syntaxe ou le vocabulaire n'ont pas été respectés, et où la manipulation de l'alternat ou la position du microphone n'ont manifestement pas été normales. Cet indicateur correspond aux cas où le simulateur reçoit une fausse information correspondant à un message "correctement" exprimé par le contrôleur.

CONCLUSION

Le CENA a développé une méthodologie ayant pour but d'évaluer les performances du terminal de reconnaissance vocale en mode autonome. Les taux bruts de reconnaissance , par mots et par phrases, sont des taux de référence mesurés lors de séances de tests se déroulant dans des conditions bien précises. C'est par eux que l'on peut apprécier l'influence de la taille du vocabulaire et de la complexité de la syntaxe, au fil des différentes séances de tests. Outre que les taux bruts sont un moyen de contrôler l'influence du niveau de langage sur les performances du système de reconnaissance, ils permettent la confrontation de résultats provenant d'études différentes. Ils donnent une signification claire à de chiffres qu'ils serait sinon peut-être difficile d'interpréter.

L'évaluation de la qualité du processus de compréhension se définit par rapport aux taux bruts de reconnaissance. Les indicateurs que l'on utilise ne rendent pas compte de la qualité globale du système mais renseignent sur certains aspects, tel que le niveau de fausse information d'une simulation. Ils reposent sur une classification des types d'erreurs provoquées de façon évidente ou indirecte par le système de reconnaissance. En cas d'hésitation ou de non respect des contraintes liées à la syntaxe-phraséologie, que mettre en cause, le système de reconnaissance ou le niveau de l'élève? Du point de vue de l'utilisateur, l'intérêt de ces indicateurs réside dans le fait qu'il est toujours possible de les obtenir, d'une façon reproductible, et qu'ils ont une signification bien précise dans le cadre de l'application. S'ils ne permettent la confrontation avec d'autres applications, ils constituent une base sur laquelle l'utilisateur peut s'appuyer pour apprécier les performances de son système de reconnaissance de la parole, mesurées au cours de différentes séances d'évaluation.

BIBLIOGRAPHIE

(1) J. MARIANI A. AOUATI Utilisation des techniques vocales dans des tâches de contrôle aérien Rapport final de la convention LIMSI/CENA 1983-1985

(2) C. BAILLEUL Détection des indicatifs tests d'évaluation Rapport CENA R8708 1987

COMPARAISON DE QUALITE SUBJECTIVE DE TROIS SYNTHETISEURS DE PAROLE

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

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

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

I INTRODUCTION

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

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

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

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

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

II LE MATERIEL DE TEST

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

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

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

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

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

Conditions du test en A/S et en SàpT

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

III DEROULEMENT DU TEST

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

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

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

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

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

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

IV ANALYSE DES RESULTATS

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

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

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

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

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

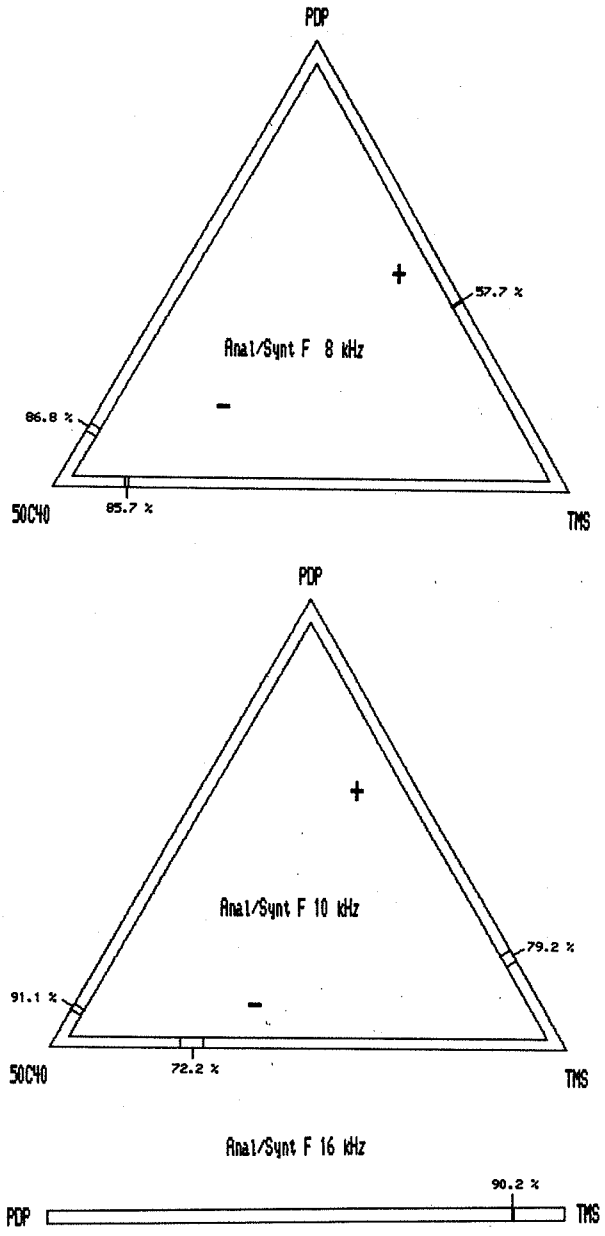


Figure 1-a

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

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

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

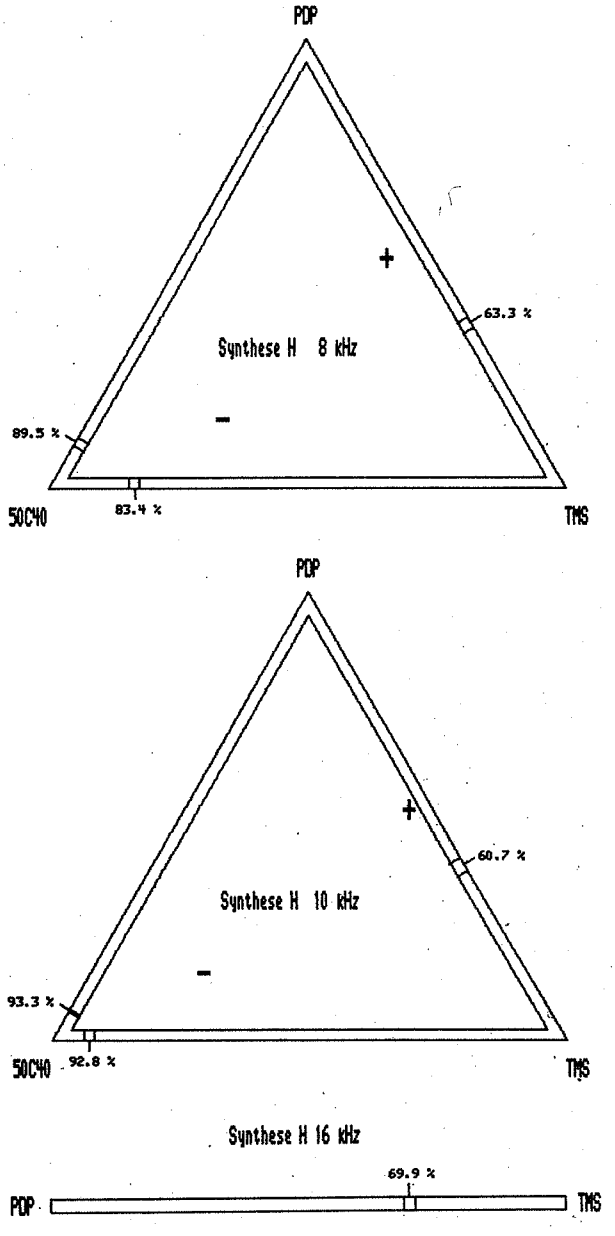


Figure 1-b

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

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

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

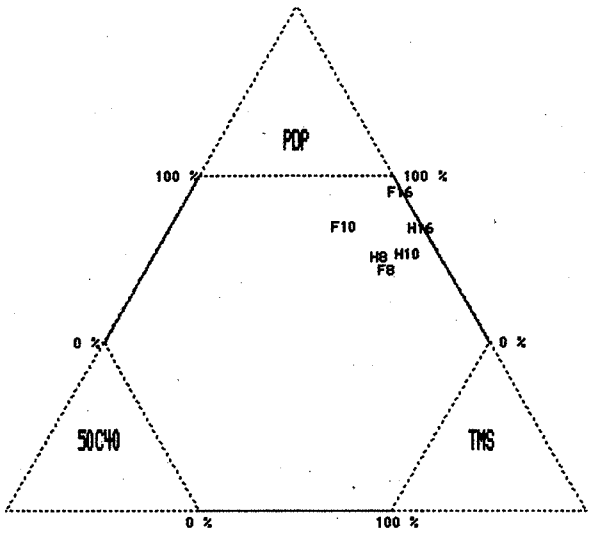


Figure 2

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

V ANALYSE DES REPONSES

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

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

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

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

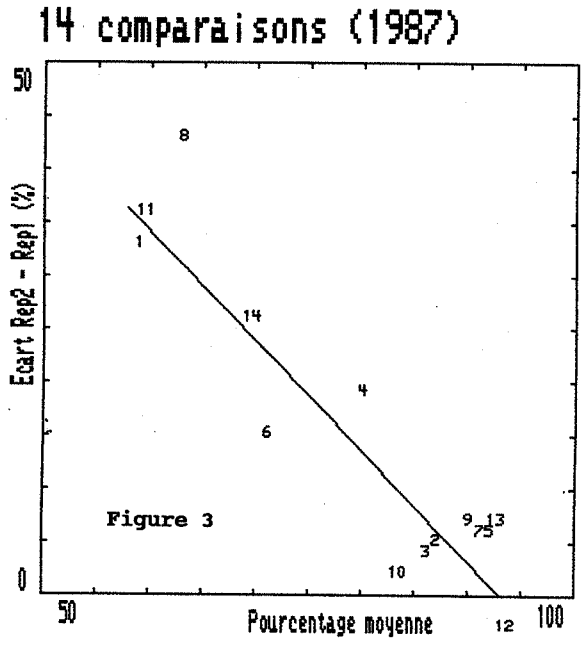


Figure 3

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

$$N = a\ N1 + b$$

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

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

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

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

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

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

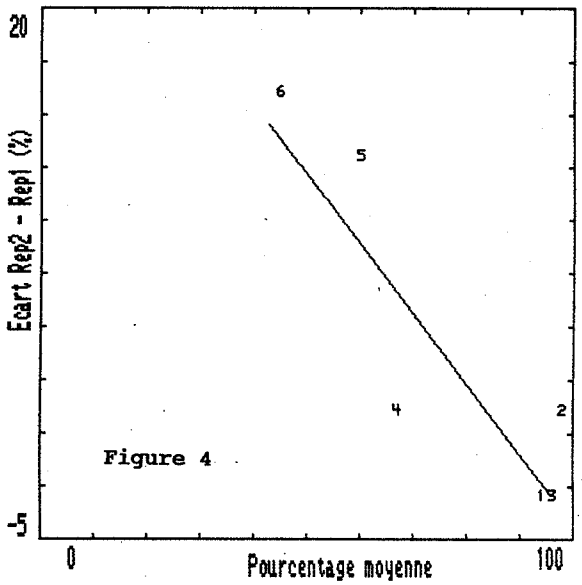


Figure 4

VI CONCLUSION

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

VII BIBLIOGRAPHIE

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

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

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

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

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

LE PROJET EUROPEEN "SAM" : EVALUATION MULTI-LINGUE DES DISPOSITIFS D'ENTREE-SORTIE VOCALE

JM DOLMAZON (I.C.P. GRENOBLE); C. BENOIT (CNET LANNION); JL. GAUVAIN (LIMSI PARIS);

G. PERENNOU (CERFIA TOULOUSE)

C.N.R.S. GRECO "COMMUNICATION PARLEE", C.R.I.N. Université de Nancy I
B.P. 239 54506 VANDOEUVRE-LES-NANCY (FRANCE)

RESUME : ce papier présente le contenu et l'état d'avancement (JUIN 87) du projet européen "SAM" (SPEECH ASSESSMENT METHODS) auquel participent les laboratoires du GRECO "COMMUNICATION PARLEE" et le CNET Lannion. L'objectif principal de ce projet concerne l'évaluation multi-lingue des dispositifs d'entrée-sortie vocale. Pour conduire ces évaluations de synthèse et de reconnaissance, les partenaires ont décidé de mettre en place une base de données parole pour différents langages. Cette base de donnée sera gérée sur une station de travail standardisée au niveau européen. Ce poste de travail servira aussi aux tâches d'évaluation et d'analyse de performances. Les partenaires souhaitent arriver ainsi à une standardisation des méthodes qui permette un libre échange d'informations dans la communauté et une progression des connaissances dans le domaine du traitement de la parole.

I-INTRODUCTION : L'utilisation au sein de la communauté économique européenne (C.E.E.) d'une grande variété de langues n'est pas seulement un obstacle à la communication entre les individus, c'est aussi un handicap pour les travaux en synthèse et reconnaissance de parole car elle limite l'utilisation de ces dispositifs dans les différents pays. L'état d'avancement des recherches dans ce domaine, laisse prévoir que de plus en plus de systèmes commerciaux seront disponibles dans les prochaines années. Pour pouvoir faire un choix objectif du système le plus approprié à une application donnée, il est nécessaire de disposer de critères objectifs pour l'évaluation de ces systèmes. Cette évaluation devant être faite au sein de la communauté elle doit être fondamentalement multilingue. De nombreuses méthodes d'évaluation existent déjà pour la plupart des systèmes mais il est rare que ces méthodes puissent être appliquées à de multiples dispositifs et encore plus rare qu'elles soient objectivement indépendantes du langage.

Un des objectifs généraux du projet est donc de conduire ce travail sur une base multilingue pour obtenir des méthodes d'évaluation standardisées indépendantes du langage aussi bien pour ce qui concerne la synthèse que la reconnaissance. Pour appliquer ces méthodes, il sera nécessaire de définir le "matériau" de parole utilisé. C'est pourquoi une des premières tâches va consister à mettre en place une base de données de signaux de parole. Pour être véritablement utile, cette base de données de grande ampleur devra être manipulée par un dispositif de gestion de base de données. De plus pour un examen approfondi des performances des systèmes d'entrée-sortie vocale testés, une connaissance approfondie des caractéristiques acoustiques et phonétiques des différents sons sera nécessaire. Pour cela une analyse des signaux enregistrés devra être conduite sous une forme normalisée et les procédures d'analyse ainsi définies, mises à disposition des utilisateurs de la base de données. Un poste de travail type est donc en cours d'élaboration. Il a pour objectif de définir une station de travail standard utilisable dans tous les centres de parole de la communauté pour conduire les taches essentielles de gestion de la base, d'analyse du signal acoustique (étiquetage, analyse FFT, analyse LPC, détection du fondamental, etc.) et enfin d'analyse de performances.

II-LES PARTENAIRES et L'ORGANISATION MATERIELLE DU PROJET :

Le projet SAM est actuellement dans une phase préliminaire dite "de définition". Cette première phase devrait être prolongée par une phase intermédiaire de 12 à 18 mois pour aboutir en 1989 à une phase définitive de 3 années portant la durée globale du projet à 5 années (fin du projet 1992).

Cinq pays participent à la première phase du projet. Pour chacun de ces pays, le tableau suivant donne le nom des laboratoires, firmes ou organisations qui assurent la mise en place des actions ainsi que le pourcentage total de leur participation :

ROYAUME UNI (32 %) :
- University College of LONDON ("PRIME CONTRATOR")
- National Physical Lab.
- British Telecom.
- Royal Signals & Radar Establishment
- SMITHS Industries
- LOGICA

DANEMARK (15 %) :
- Jutland Telco. (JTAS)
- Inst. of Electr. System (Univ. d'Aalborg)

ITALIE (12 %) :
- CSELT Turin

HOLLANDE (7 %) :
- TNO Inst. for Perception

FRANCE (34 %) :
- CNET Lannion
- GRECO "Comm. Parlée" avec :
  - I.C.P. - Grenoble
  - LIMSI - Orsay (Paris)
  - ENST - Paris
  - CRIN - Nancy
  - CERFIA - Toulouse
  - I.P. Aix en Provence

Pour la France, la gestion scientifique de l'ensemble du projet est assurée par le GRECO Communication Parlée.

III-ORGANISATION SCIENTIFIQUE :

III-a) PHASE DE DEFINITION :

Pendant la phase de définition, l'objectif principal est de définir les bases sur lesquelles sera entrepris le travail futur du projet. Pour cela, cinq tâches ont été simultanément entreprises :

1. Revue des bases de données existantes et des dispositifs de stockage utilisés (sous la responsabilité des Danois et en relation étroite avec les Français).
2. Revue des stations de travail utilisées en parole (responsabilité France). Cette revue donnera les recommandations techniques de la station de travail à utiliser par les différents partenaires.
3. Revue des méthodes d'évaluation existant en reconnaissance (responsabilité Italie). Cette revue s'intéresse à tout type de système (mots isolés ou connectés, parole continue, dépendant ou non du locuteur, au niveau acoustique ou au niveau phonétique).
4. Revue des méthodes d'évaluation en synthèse (et pour les systèmes de synthèse à partir de texte).
5. Revue des méthodes d'étiquetage, de contrôle phonétique et des références normatives.

L'objectif commun à ces différentes revues et de définir de façon précise les outils (tant matériels que méthodologiques) et les normes du travail commun à entreprendre.

III-b) PHASE SUIVANTE :

Pour l'ensemble du projet, le travail a été découpé en quatre "WORKING PACKAGES" :

WP1 : BASE DE DONNEES - Mise en place d'une base de données commune.
WP2 : RECONNAISSANCE - Evaluation en mots isolés ou connectés, grands vocabulaires.
WP3 : SYNTHESE - Evaluation de la qualité de synthèse (naturel, acceptabilité, intelligibilité, etc.)
WP4 : RECONNAISSANCE - Evaluation de parole continue, contrôle phonétique et analyse des langages.

Pour ce qui concerne la participation française au projet, le GRECO a désigné les responsabilités suivantes :

WP1 : I.C.P. Grenoble (JM DOLMAZON)
WP2 : LIMSI Paris (JL GAUVAIN)
WP3 : CNET Lannion (R DESCOUT)
WP4 : CERFIA Toulouse (G PERENNOU)

JM DOLMAZON est chargé de la coordination d'ensemble de cette participation.

IV-DESCRIPTION DES ACTIONS :

IV-1) WP1 : BASE DE DONNEES :

Pendant la phase de définition les travaux consacrés à cette action se sont focalisés sur l'étude de la future station de travail et sur les systèmes de gestions de bases de données.

a) La station de travail : deux niveaux techniques de station standard ont été définis. Une station de bas niveau (bon marché) basée sur le standard IBM-PC a été définie. Elle inclut principalement un système de stockage de masse de grande capacité (500 Moctets) pour la base de données, un convertisseur D/A pour les sorties analogiques de signaux de parole et une liaison série pour les entrées de résultats en provenance des systèmes sous test. La station de "haut niveau" est pour des raisons de compatibilité basée aussi sur le standard IBM-PC, mais la station est ici de haut de gamme (PC-AT3) avec de gros moyens de stockage (disque dur et CD ROM), des moyens de calculs additionnels (carte de traitement parole spécialisée avec accés direct mémoire et convertisseurs D/A et A/D) et un certain nombre de logiciels de gestion (SGBD pour la base de données, procédures automatisées pour l'enregistrement de corpus et leur contrôle), de logiciels de travail (éditeur de signal ou de sonagramme, étiquetage, ..), de logiciels de traitement (analyse du signal de parole : FFT, LPC, détection de formant, ...) ainsi que des logiciels d'analyse statistique. Cette station de travail sera utilisée pour la mise au point des procédures automatiques d'évaluation qui, une fois qualifiées, seront transportées pour exploitation sur la station de bas niveau. Cette station de travail "haut niveau" sera munie de possibilités standard de communication pour que son utilisation soit facilitée dans le cadre de centres équipés de moyens lourds (VAX, MASSCOMP, APOLLO, etc.).

Ces deux types de stations de travail doivent recevoir la base de données qui sera enregistrée. C'est pourquoi une grande attention a été portée sur les dispositifs de stockage de masse. Devant la difficulté de choix, les partenaires européens ont décidé de faire réaliser un disque compact de test (CD-ROM) avec 100 Moctets de parole pour chacune des 5 langues : FRANCAIS, ANGLAIS, HOLLANDAIS, ITALIEN et DANOIS. Le corpus choisi est un corpus de digits (isolés, en triplets ou en parole continue pour 4 locuteurs dans chaque langue). Un lecteur de CD-ROM sera couplé aux premiers prototypes de la station "SAM" pour une évaluation objective de performances. Ce premier essai devrait être poursuivi par une étude approfondie des solutions à base de disque optique numérique (WORM) ou d'enregistreur magnétique numérique (DAT).

b) Base de données : avant de déterminer avec précision les fonctionalités que devra avoir le système de gestion de la base de données, il est nécessaire de définir le contenu (et donc les corpus et les "objets") de la future base de données européenne. C'est pourquoi, sur la base des travaux entrepris depuis plusieurs années par le GRECO, une étude comparative de plusieurs systèmes de gestion de base de données a été entreprise. Ses conclusions serviront de base à la définition du logiciel SGBD à utiliser. L'expérience acquise au sein du GRECO servira également à la définition d'un schéma conceptuel pour la gestion des données de la base européenne.

IV-2) WP2 : RECONNAISSANCE :

La difficulté actuelle d'obtenir une évaluation objective des performances d'un système de reconnaissance est bien connue. En effet, bon nombre de constructeurs annoncent des taux de reconnaissance supérieurs à 99% sans préciser toutes les conditions de tests.

Le but de WP2 est d'établir des protocoles standard d'évaluation et d'étudier une station de test pour mesurer les performances des systèmes.
En fait, pour évaluer les performances d'un système de reconnaissance, il n'y a pas de meilleure solution que de le tester dans les conditions réelles d'utilisation. Ces

conditions dépendent de la tâche à accomplir, du vocabulaire, des locuteurs, de l'environnement acoustique, de la prise de son, et du type de liaison entre le microphone et le système de reconnaissance. A moins d'enregistrer un corpus pour chaque application particulière, il faut donc disposer d'une base de données très importante, quelques centaines de milliers de mots prononcés par quelques centaines de locuteurs, intégrant de manière contrôlée les principales sources de variabilité. La station de test doit d'une part, être capable de présenter au système des signaux calibrés et controlés, et d'autre part, dépouiller les réponses qui lui sont retournées en fonction des informations acoustiques et non-acoustiques contenues dans la base de données. Elle doit en outre intégrer un certain nombre d'outils, tel qu'un logiciel d'aide à la définition des vocabulaires, ou des logiciels d'évaluation des bases de données (par exemple, un logiciel de reconnaissance servant de système de référence).

Durant la première phase de huit mois, les cinq points suivants sont traités:

- L'état de l'art en France et à l'étranger sur les procédures d'évaluation, et la réalisation d'une base de données bibliographique sur l'évaluation.
- L'étude d'un outil pour évaluer la difficulté à priori d'un vocabulaire donné, afin d'un part, de pouvoir comparer les performances des systèmes tests avec des vocabulaires différents, et d'autre part, pour aider l'utilisateur à définir un vocabulaire optimal pour son application.
- L'étude de faisabilité d'un logiciel d'évaluation de bases de données (système de référence pour la reconnaissance de mots enchaînés).
- La définition d'un projet pour l'évaluation des systèmes de reconnaissance de grands vocabulaires (plus de 1000 mots).
- L'initialisation d'un projet pilote consistant à tester quelques systèmes français avec la base de données du GRECO (BDSONS).

Pour les trois années du projet, le programme de travail comprend trois axes:

- La définition d'une base de données en collaboration avec WP1: cette base devra comprendre un grand nombre de locuteurs (quelques centaines) et devra permettre d'évaluer dans diverses conditions, les performances des systèmes de reconnaissance de mots isolés et enchaînés, mono- et multi-locuteurs, pour de petits vocabulaires. Elle devrait également permettre de tester dans des conditions plus restreintes, les systèmes de reconnaissance de grands vocabulaires (plus de 1000 mots) en mode monolocuteur. Les énoncés contenus dans cette base de données doivent couvrir un domaine de variabilité contrôlé et mesurable pour tirer le maximum d'informations des résultats d'évaluation.

- Le développement d'une station de test: la première fonction de cette station est de présenter les signaux au système de reconnaissance sous forme analogique, avec la possibilité de choisir la bande passante et la réponse en fréquence, et d'additionner au signal vocal un bruit de caractéristiques données. La deuxième fonction est de contrôler le système de reconnaissance auquel il peut être relié au moyen d'une liaison série et/ou parallèle et de dépouiller les résultats de l'expérimentation en tenant compte des différents types d'erreurs. Cette fonction est très délicate car elle dépend du système test (protocole, mode de reconnaissance, mode de corrections des erreurs, etc...). Durant le projet, quelques systèmes de reconnaissance de mots isolés et enchaînés, mono- et multi-locuteurs, seront évalués au moyen de cette station de tests avec des bases de données correspondant aux différentes langues.
- Le développement de protocoles et d'outils d'évaluation : les protocoles de tests seront définis en fonction du type de système test (mots isolés ou enchaînés, mono- ou multi-locuteurs), ainsi que les procédures de dépouillement. Un logiciel d'évaluation de la difficulté à priori d'un vocabulaire, ainsi que des systèmes de référence (mots isolés et mots connectés) seront développés.

IV-3) WP3 : SYNTHESE :

Ces dernières années nous ont permis de voir apparaître "sur le marché" bon nombre de "synthétiseurs de parole", qu'il s'agisse de synthèse à partir du texte, de concaténation de segments préenregistrés, ou simplement de stockage-restitution pour différentes langues.

En premier lieu, les constructeurs sont systématiquement confrontés à l'épineux problème du DIAGNOSTIC de qualité de ces systèmes. Celui-ci ne peut, en l'état de l'art, être résolu que par l'intervention d'un "expert", le plus souvent secondé par les résultats de tests subjectifs traditionnels hérités, pour la plupart, de l'expérience de la téléphonométrie.

En second lieu, les Administrations des Télécommunications des différents pays se heurtent régulièrement à l'absence de critères de référence pour leurs décisions de choix, de recette, d'agrément ou d'homologation de systèmes de diffusion de parole. Il n'existe actuellement aucun standard national ou international permettant de définir A PRIORI la "qualité minimale" requise pour un système parlant.

C'est dans ce contexte que les administrations européennes sont vivement intéressées par l'élaboration de tests standard.

Le CNET propose une méthode basée sur la résistivité de l'intelligibilité de la parole synthétique à la dégradation.

La première partie de cette investigation (8 mois) porte sur la définition d'une méthodologie de test indépendante du code linguistique, souple, automatisable et permettant des comparaisons à un ou des systèmes de référence.

La deuxième partie de la contribution au projet ESPRIT sera consacrée à l'implantation de ce test sur un poste léger et transportable, et à l'évaluation de ce test par rapport à d'autres méthodes.

INTELLIGIBILITE DE LA PAROLE SYNTHETIQUE NOYEE DANS LE BRUIT :

Dans une optique de sélection de systèmes, le critère de choix est nécessairement monoparamétrique. Ce paramètre unique peut être un facteur résultant de la pondération de plusieurs indices de qualité. Mais ce problème de la pondération optimale n'étant pas résolu à ce jour, il nous parait urgent de définir un critère simple qui devra ensuite être confronté à des analyses plus détaillées.

Le postulat de départ est que le "meilleur système" sera celui qui RESISTERA le mieux,

du point de vue de l'INTELLIGIBILITE, à la dégradation la plus simple. Au signal sera donc superposé un bruit.

Dans un tel test, l'auditeur entend une phrase noyée dans un bruit dont l'intensité, totalement masquante au début, diminue pas-à-pas tant que le sujet estime n'avoir pas compris cette phrase. Trois mesures objectives sont alors fournies par ce test : le rapport signal sur bruit S/B atteint lorsque le sujet pense avoir compris, le rapport S/B lorsque le sujet a correctement compris la phrase (seuil d'intelligibilité), ainsi que les temps de réaction du sujet à répondre.

La décision de "phrase correctement comprise" sera, autant que possible, et en fonction des applications prévues pour le système testé, la réalisation correcte d'une tâche accomplie par le sujet : composition d'un numéro de téléphone sur un clavier, par exemple.

Cette méthode a déja été utilisée par BACRI et GRAILLOT pour situer un codeur et deux synthétiseurs à partir du texte par rapport à de la voix naturelle. Les réponses fournies par plusieurs auditeurs sur des listes de phrases phonétiquement équilibrées étaient jugées par un opérateur. Cet ensemble de résultats à permis d'établir des courbes comparatives donnant pour les différents types de parole (naturelle, LPC ou synthèses à partir du texte) le pourcentage d'identification par rapport au rapport signal sur bruit.

IMPLANTATION ET EVALUATION DE LA METHODE

Le travail restant à accomplir au cours de la deuxième période d'ESPRIT sera l'implantation de ce test sur la station de travail définie par le GRECO conformément aux recommandations du projet SAM (PC AT + carte AU20), afin de rendre la procédure facilement transportable.

La synthèse à partir du texte du CNET sera ensuite évaluée à travers ses différentes possibilités : voix "masculine" et "féminine", prosodie "naturelle" et "publicitaire", diphones LPC et diphones a formants. La comparaison pourra être étendue ultérieurement à d'autres langues.

En outre, d'autres tests sur ces mêmes synthétiseurs permettront de situer l'apport de cette méthode "monoparamétrique" par rapport à d'autres. Une base de données "stimuli synthétiques", déja grosse de l'expérience antérieure du CNET sera ainsi enrichie par ces tests et pourrait permettre de définir une "référence" de qualité pour les évaluations ultérieures.

IV-4)WP4 : RECONNAISSANCE :

L'objectif de WP4 est la définition d'une méthodologie d'évaluation des constituants d'un système de reconnaissance automatique de parole comportant un niveau phonétique explicite. Les implications concernant un poste de travail pour l'évaluation de ses systèmes relèvent également des objectifs de ce groupe.

Après concertation entre les laboratoires concernés, les thèmes suivants se sont dégagés :

T1 : alignement phonétique automatique, s'appliquant en particulier à la segmentation et à l'étiquetage semi-automatique de corpus ;

T2 : segmentation et étiquetage en classes phonétiques majeures, utilisés pour l'accès au lexique ;

T3 : segmentation et étiquetage automatique en phonèmes, avec comme application certaines classes de systèmes de reconnaissance automatique de parole ;

T4 : tests d'algorithmes de reconnaissance s'appliquant à des entrées représentées au niveau phonétique ;

T5 : performances de l'auditeur humain en matière de reconnaissance phonétique ;

T6 : bases de données et corpus de tests phonétiques.

Au cours d'une réunion de travail, organisée au CSELT de TURIN rassemblant les partenaires du projet SAM, certaines directions se sont précisées.

La première a trait au caractère multi-lingue de la station d'évaluation envisagée, ce qui oriente vers deux types de corpus de test : les premiers seraient destinés à des évaluations de systèmes de reconnaissance indépendamment de la langue ; les seconds contiendraient des corpus rassemblant les difficultés spécifiques des diverses langues. Il appartiendra aux partenaires de WP4 de préciser, dans cette double perspective, leurs besoins propres en corpus et en étiquetage. Les conclusions du groupe d'étiquetage large du GRECO animé par L. MICLET et de la réunion de travail du mois de JUIN 87 à LONDRES, devront être pris en compte; ceci justifie au sein de WP4, le thème T6.

La seconde direction concerne le test des algorithmes. Il semble maintenant bien admis qu'à coté de l'évaluation globale des cartes de reconnaissance, il soit nécessaire d'évaluer d'une manière plus générale les algorithmes de reconnaissance dans les systèmes analytiques, ce qui justifie le thème T4.

Enfin, une troisième direction de recherche est celle qui consiste à rechercher des modèles pour prédire la difficulté des corpus ou pour en mesurer les difficultés relatives. R. MOORE a suggéré une méthode utilisant l'auditeur humain pour déterminer à quels niveaux de bruits relatifs les performances de reconnaissance sont identiques, ce qui permet d'envisager une mesure perceptive de la difficulté des corpus. D'une manière plus générale, la connaissance des performances de l'auditeur humain est une donnée de comparaison qui sera précieuse dans le programme SAM ce qui justifie le thème T5.

IV-CONCLUSIONS :

On le voit, les objectifs de ce projet sont ambitieux et, pour bien des raisons, le projet SAM semble essentiel aux travaux entrepris depuis de nombreuses années sur la parole.

Le premier enjeu essentiel de ce projet est de fournir des méthodes d'évaluations objectives de la qualité des dispositifs d'entrée-sortie vocale. Cette mesure réalisée dans un contexte multi-lingue permettra une véritable évaluation des systèmes au delà des barrières linguistiques naturelles.

Un second enjeu capital réside dans l'important travail de normalisation entrepris par les partenaires du projet. Cette normalisation s'étend du matériel utilisé pour les recherches jusqu'aux méthodologies et aux procédures. Cet effort, sans précédent dans le domaine de la parole, facilitera l'échange des informations dans la communauté scientifique internationale et sera, sans nul doute, un facteur de progrès scientifiques.

INTELLIGIBILITE DE LA PAROLE DANS LES VEHICULES AUTOMOBILES MISE AU POINT EXPERIMENTALE

A. FOTI - G. PACHIAUDI - M. VERNET *
A. MARCHAL **

* INRETS-LESCO 109, avenue S. Allende - 69500 BRON (France)
** Institut de Phonétique d'Aix-en-Provence (France)

ABSTRACT

This research deals with disturbances of speech communication within cars.

Its objectives are to provide car manufacturers with means of assessment of car comfort, and especially concerning the effect of noise on speech communication.

The assessment means will be determined by adjusting indicators measured, calculated and evaluated by panels.

Phonetic tests by panels are used in laboratory ; to this end, a noise recording method in situ together with a laboratory simulation of both noisy environment and phonetic tests have been worked out.

The second step of this study consists in establishing comfort scales whose aim is to compare various noisy situations.

Cette recherche concerne les perturbations de la communication orale à l'intérieur des véhicules légers.

1. LES OBJECTIFS

Les objectifs à moyen terme sont de proposer aux constructeurs des outils d'évaluation de l'ambiance bruitée en fonction des différents modes d'utilisation (types de véhicules, de revêtements, de conduite) et en fonction des différents types d'utilisateurs et de communications vocales (conversation, utilisation du téléphone, conduite assistée et guidage, écoute radio). Les objectifs seront réalisés par une confrontation des performances d'indices mesurés, calculés, évalués par jurys.

Indices mesurés : Au moyen du système RASTI (Rapid Speech Transmission Index) (1), variante du STI, commercialisé par Brüel et Kjaer, a été effectuée une série de mesures dans le véhicule pour différentes places occupées par les composantes émission et réception ainsi que pour des conditions d'enregistrement contrastées selon la vitesse et le type de revêtement.

Indices calculés : Il s'agit des indices classiques d'évaluation des possibilités de communication tels que : A.I. - Articulation Index - (2), L_{SIL} - Speech Interference Level - (3) et leurs dérivés qui prennent en compte le rapport signal à bruit dans une gamme de fréquences plus ou moins étendue, ainsi que des calculs de niveaux équivalents Leq pratiqués sur le bruit et sur la parole.

Indices obtenus à l'aide de jurys : Ceux-ci sont composés de résultats à des tests phonétiques :
a) le test de diagnostic par paires minimales et triplets élaboré par Cartier et Rossi (4) qui repose sur l'analyse qualitative du système de fautes occasionné par les dégradations du signal doit nous permettre de localiser les sources majeures de perturbations, et répond ainsi au souhait de portée opérationnelle d'une grille d'évaluation.

b) un test (chiffres, mots, logatomes) qui, par un indice plus global, propose une différenciation nette des situations.

c) des évaluations subjectives de la qualité d'écoute :
- par réponse à des questionnaires et par notation sur échelles concernant l'ambiance bruitée et le confort global
- par choix d'un niveau de préférence d'écoute du poste. Cette étape du test est réalisée in situ.

2. PROTOCOLE EXPERIMENTAL ET RESULTATS ATTENDUS

Pour comparer l'efficacité des tests d'évaluations par jurys, nous avons mis au point, dans un premier temps, une méthode d'enregistrement in situ, puis de simulation en laboratoire de l'ambiance bruitée.

Les enregistrements sont réalisés dans le véhicule pour des allures de fonctionnement "stables", à l'entrée du pavillon des deux oreilles par l'intermédiaire de microphones montés sur casque. Pour le message, nous utilisons une bande enregistrée selon les normes professionnelles des télécommunications.

La chaîne de reproduction de l'ambiance bruitée et des messages (cf Figure 1) est étalonnée au moyen d'une tête artificielle (KEMAR) (5) qui permet le contrôle rigoureux de la restitution des caractéristiques du signal aux écouteurs (dynamique, bande passante, directivité).

Le déroulement du test a lieu au laboratoire d'évaluation de l'environnement de l'INRETS. Deux options président au recrutement des sujets : faire appel à un jury de professionnels, constituer un échantillon hétérogène représentatif de la population des conducteurs. Le test comprend deux phases :
Pour la première, le mixage du bruit avec le message s'effectue en utilisant la parole enregistrée en chambre sourde, en ayant pour but de décrire les effets du bruit sur la parole.
En un second temps, avant l'étape de mixage, le message est réenregistré dans le véhicule à l'arrêt au moyen d'une chaîne portative qui fait office de locuteur simulé afin de récupérer la parole "colorée" par la fonction de transfert de l'habitacle.

Ainsi seront testés les effets additionnés de l'habitacle et du bruit sur la parole.Le constat d'un écart de résultats entre les deux phases nous permettra de juger de la dégradation d'intelligibilité consécutive aux seules caractéristiques d'habitacles.

La combinaison des places retenues pour l'émetteur et le récepteur mettra en évidence les répercussions possibles au niveau psycho-acoustique des facteurs de provenance (directivité) et d'éloignement des sources de parole et de bruit.

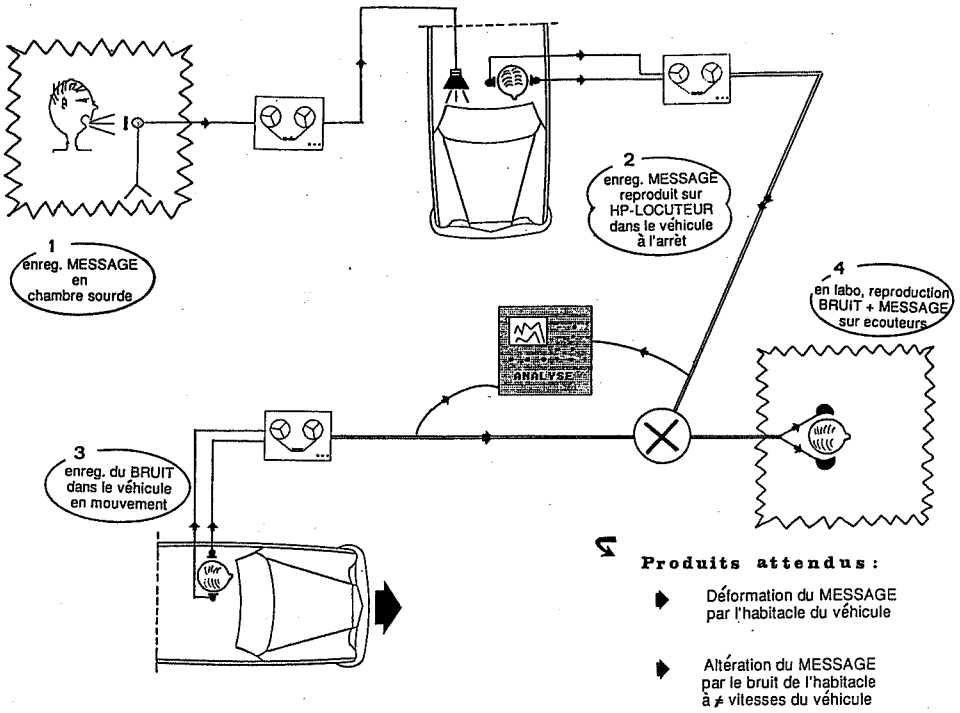


Figure 1 : Etude de l'altération du message parlé dans un habitacle de véhicule

La comparaison des indices énumérés, par matrice de corrélations, est destinée à en établir la redondance, à juger de leurs qualités discriminatives et de leur sensibilité aux caractéristiques d'ambiances bruitées ainsi que d'habitacles.

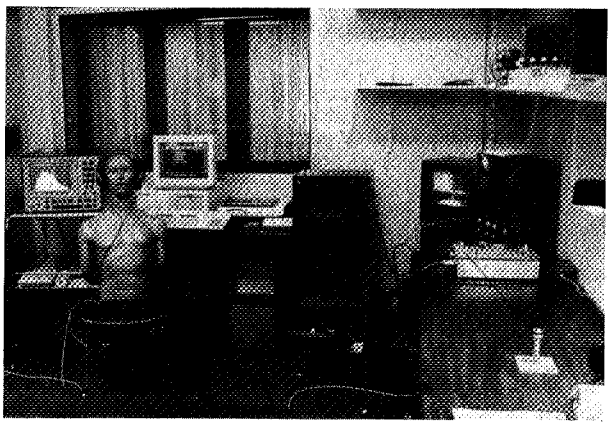
La dernière étape de la recherche consistera à élaborer une grille d'évaluation du confort qui combine l'apport informatif des différents critères.

Cette étude est réalisée dans le cadre d'un accord GIE Renault/PSA - INRETS.

BIBLIOGRAPHIE

(1) RASTI conforme à norme IEC Draft. Pub. 268, Part 16
(2) AI Norme ANSI S3.5-1969
Methods for the calculation of the articulation index
(3) L_{SIL} Document de travail discuté dans le groupe ISO/TC/59/SC5/WG3N235, 1986
(4) CARTIER M. et ROSSI M.
Le test de diagnostic par paires minimales : mise en oeuvre et résultats
Liège - 1973 - Symposium Intelligibilité de la parole, pp 191-208
(5) KEMAR : SHAW E.A.G. and TERANISHI R. (1968)
Sound pressure generated in an external-ear replica and real human ears by a nearby paint source
JASA, 44, pp 240-256
(6) G. PACHIAUDI, M. VERNET, A. FOTI
Rapport INRETS (Convention RENAULT/PSA)
Intelligibilité à l'intérieur des véhicules automobiles - 1987

Système d'enregistrement du bruit dans l'habitacle



La chaîne d'analyse et de reproduction

ALIGNEMENT AUTOMATIQUE OPTIMAL DE DEUX CHAINES PHONETIQUES

J.P. Lefèvre[1] , V. Aubergé[1,2], D. Maret[2,3]

[1]. Société OROS 38240 Meylan ZIRST.
[2]. Institut de la Communication Parlée de Grenoble - IPG.
[3]. CRISS UII Grenoble.

ABSTRACT

This communication gives some solutions to specific problems of phonetic string alignment, met during experiments carried out for an evaluation of the intelligibility of a text-to-speech synthesis system.

The optimal alignment of the "stimuli string" with the "response string" of a listener is performed through a generalization of the algorithm "String to string correction problem" due to Wagner&Fisher.

This alignment strategy involves by a set of rules which is updated during a first phase of semi-automatic learning.

Ultimatly, the parasit errors are corrected in order to produce coherent input for the expected intelligibilty tests.

I. INTRODUCTION

Cette communication entre dans le cadre de l'évaluation de l'intelligibilité d'un système de parole du Français [1] à partir de tests d'écoute. L'évaluation est réalisée par comparaison directe entre stimulis synthétisés par le système et réponses fournies par un groupe d'auditeurs à l'écoute de ces stimulis.

Un telle méthodologie présuppose que les données acquises sont directement interprétables, c'est-a-dire que seules les erreurs de perception apparaissent dans la réponse de l'auditeur. En pratique, les conditions expérimentales introduisent des modifications non négligeables de la chaîne perceptive, excluant ainsi l'exploitation immédiate de la chaîne réponse.

Il est donc nécessaire de traiter les chaînes réponses afin de cerner au mieux, puis de corriger les erreurs parasites.

II. PROTOCOLE EXPERIMENTAL

Afin d'évaluer la qualité acoustique des segments élémentaires mis en oeuvre dans un système de synthèse du Français à partir de diphones, un test d'écoute a été développé pour mesurer l'intelligiblité de chaque segment. Pour que ce test soit complet et efficace, un ensemble de 646 mots sans signification du type VCCV et CVVC décrivant toutes les combinaisons est généré par synthèse. A titre de référence, un sous-ensemble de ces mêmes mots prononcés naturellement puis codés par prédiction linéaire est également utilisé.

Lors de l'expérience, des groupes de quatre auditeurs transcrivent les sons entendus sur un clavier alpha-numérique modifié. Les sons autorisés sont tous accessibles par une seule touche, étiquetée par deux caractères alpha-numériques. Les auditeurs sont naïfs, c'est-à-dire sans expérience antérieure de la synthèse, et sans connaissances phonétiques particulières.

Pour que ce test soit complet et efficace, un ensemble de 646 mots sans signification du type VCCV et CVVC décrivant toutes les combinaisons est généré par synthèse. A titre de référence, un sous-ensemble de ces mêmes mots prononcés naturellement puis codés par prédiction linéaire est également utilisé.

Lors de l'expérience, des groupes de quatre auditeurs transcrivent les sons entendus sur un clavier alpha-numérique modifié. Les sons autorisés sont tous accessibles par une seule touche, étiquetée par deux caractères alpha-numériques. Les auditeurs sont naïfs, c'est-à-dire sans expérience antérieure de la synthèse, et sans connaissances phonétiques particulières. pour faciliter leur apprentissage, un codage des sons par des symboles proches de l'orthographe courant a été adopté:

Voyelles :

/a./ (pAte)	;	/A./ (pAtte)	
/O./ (dOnne)	;	/AU/ (chAUd)	
/EI/ (EtE)	;	/AI/ (mAIs)	
/EU/ (fEU)	;	/OE/ (hEUre)	
/U./ (crU)			
/OU/ (bOUt)			
/I./ (vIte)			
/UI/ (hUIt)		("ué" + i)	
/AN/ (ENfANT)			
/ON/ (mONtagne)			
/IN/ (brIN, brUN)		(archi-phonème)	

Consonnes :

/G./, /P./, /T./, /B./, /D./, /V./
/F./, /M./, /N./, /L./, /R./, /CH/
/GN/ (campaGNe)
/K./ (QUatre)
/Z./ (Zéro)
/S./ (Soif)
/W./ (OUi)
/Y./ (fiLLe)
/J./ (Jour)

Les auditeurs reçoivent typiquement un entraînement d'une demi-journée, leur permettant de se familiariser avec la tâche à accomplir, le clavier, la qualité de la parole synthétique, et la fréquence d'apparition des stimulis. Les mots sont présentés par liste de 50, avec un mot toutes les 4 secondes. Ensuite, l'experimentation proprement dite se déroule sur trois demi-journées consécutives. Normalement, au cours d'une séance, vingt listes sont présentées, incluant à la fois synthèse et codage LPC. Chaque liste dure environ quatre minutes et de fréquentes pauses sont respectées entre les listes. Le strict respect de la présentation aléatoire des différents stimulis a été particulièrement surveillé.

En pratique, dans les réponses, de nombreuses erreurs parasites viennent se superposer aux erreurs perceptives et éliminent toutes possibilités d'exploitation directe des résultats. Les causes en sont diverses, citons :

II.1. Codage:

Le codage proche de l'orthographe et la mauvaise connaissance par les sujets du système phonologique

s'additionnent pour donner des erreurs du type:

- difficultés d'assimilation des semi-voyelles; par exemple :

[UI] -----> [U/I]
[W/A] -----> [U/A] ou [OU/A]
[Y] -----> [I]

- chaînes phonétiquement proches et phonologiquement équivalentes:

[GN/A] -----> [N/I/A]
[GN/A] -----> [GN/I/A]

- réflexe de la dictée:

[W/A] -----> [O/I]

Dans tous ces exemples, le terme de gauche représente le stimuli, celui de droite la réponse, le "/" indiquant le passage à un nouvel élément de la chaîne considérée.

II.2. Clavier :

Les erreurs classiques engendrées par l'utilisation d'un clavier, et plus particulièrement des erreurs d'insertion apparaissent clairement :

[A/B/N/A] -----> [P/A/B/N/A]

II.3. Temps de réponse :

Le temps laissé pour répondre à un stimuli est court. Il n'est donc pas facile pour l'auditeur de respecter la place de chaque son dans sa réponse, lorsque celle-ci est incomplète. Il s'ensuit un décalage horizontal entre la chaîne stimuli et la chaîne réponse :

[AU/P/GN/A] -----> [AU/GN/A/-] (pour [AU/-/GN/A])
[AU/P/GN/A] -----> [A/-/-/-] (pour [-/-/-/A])

Lorsque l'auditeur prend du retard, il est amené à négliger complètement une ou plusieurs réponses. Il en résulte souvent un décalage vertical entre la liste des stimulis et le liste des réponses. Ces décalages, moins nombreux, et plus facilement repérables , ont été traités manuellement.

III CHOIX METHODOLOGIQUE.

Les diverses causes d'erreurs évoquées ci-dessus nous obligent à réaliser un recalage des chaînes observées sur les chaînes de référence avant d'envisager l'exploitation des résultats. Les algorithmes à mettre en oeuvre entrent dans le cadre plus général de la programmation dynamique . Deux approches se distinguent dans la littérature.

La première [2], issue des domaines de la biologie moléculaire [3] et des sciences informatiques [4], consiste à modéliser les différences existant entre les deux chaînes à comparer au moyen de trois transformations : insertion ou effacement d'un caractère, substitution d'un caractère par un autre. Ces transformations sont pondérées par une fonction de coût.

La seconde [5],[6], est issue du domaine de la reconnaissance de la parole. Elle consiste à coupler les deux chaînes de façon optimale, un caractère de l'une des chaînes pouvant être mis en correspondance avec un ou plusieurs caractères de l'autre. La mise en correspondance de deux caractères est également pondérée par une fonction de coût.

Ces deux méthodes n'offrent pas la précision requise pour notre expérimentation. Prenons l'exemple du recalage de [GN/A] avec [N/I/A], en supposant les coûts fixés de façon à favoriser les correspondances voyelle-voyelle et consonne-consonne. La première approche fournira l'alignement optimal :

[GN/-/A]
[N/I/A]

tandis que la seconde proposera la solution suivante :

[GN/A]
[N/IA]

alors qu'un alignement correct doit mettre en correspondance [GN] avec [NI]. Il nous a donc paru indispensable de généraliser la première approche [7], afin de prendre en compte des effacements, insertions et substitutions ne portant pas seulement sur des caractères mais sur des sous-chaînes quelconques des chaînes stimuli et réponse. Il est alors nécessaire de séparer le moteur de comparaison des connaissances qu'il manipule. Dans la base de règles, l'expert peut exprimer ses connaissances au moyen de règles de substitutions généralisées, suivant la syntaxe :

(X) --> (Y) [Coût] X,Y étant deux chaînes de longueur quelconque mais non nulles simultanément.

L'algorithme de comparaison est exprimé par la relation:

DG(REF,OBS) = MIN(DG(U,V) + COUT(X,Y))

où DG :: fonction (algorithme) de Dissimilarité Généralisée,
avec REF = U + X et OBS = V + Y, le minimum étant pris sur toutes les règles (X) --> (Y) dans la base.

La trace de chaque substitution (X,Y) qui a réalisé un minimum local permet de construire l'alignement final, après l'obtention du score global minimal .

IV. DEMARCHE SUIVIE.

Dans notre cas, comme d'ailleurs dans la plupart des applications manipulant des connaissances, il n'apparaît guère envisageable, à priori, de décrire complètement la base de règles. Par contre, une solution élégante consiste à acquérir ces connaissances par le biais d'une phase d'apprentissage semi-automatique. Cette phase d'apprentissage nécessite une connaissance minimale des phénomènes que l'on veut prendre en compte. Un premier ensemble de règles modélisant grossièrement les erreurs parasites qui nuisent à l'exploitation directe des résultats est tout d'abord défini comme suit :

μ étant la chaîne vide,
C1, C2 des consonnes quelconques
V1,V2 des voyelles quelconques,
L1, L2 des phonèmes quelconques:

Coût(V1,μ) = Coût(μ,V1) =100
Coût(C1,μ) = Coût(μ,C2) = 100
Coût(C1,C2) = Coût(C2,C1) = 100
Coût(V1,V2) = Coût(V2,V1) = 100
Coût(V1,C1) = Coût(C1,V1) > 200
Coût(L,LL) = Coût(LL,L) =10
Coût(C1C2,C3) = Coût(C3,C1C2) = 150
Coût(V1V2,V3) = Coût(V3,V1V2) = 150
Coût(L1L2,L2L1) = Coût(L2L1,L1L2) = 100

Ces règles correspondent à un alignement préférentiel voyelle/voyelle et consonne/consonne, à des erreurs typographiques classiques (insertion, effacement, substitution, transposition) et à des extensions/compressions. De plus, les coûts des substitutions C/C et V/V ont été affinés en utilisant leurs descriptions en traits phonétiques.

Ensuite, cet ensemble de règles est appliqué à un échantillon représentatif d'un millier de données, sur les 30000 données qui constituent le corpus global. Les recalages effectués sont présentés sous une forme facilitant l'expertise manuelle, des utilitaires de tri et de

comptage repertoriant les alignements effectués.

L'analyse de ces résultats permet d'affiner et d'enrichir la base par l'addition de règles du type :

[ON] -----> [O/N]) coût faible
[UI] -----> [W/I]) coût faible

Cette phase d'apprentissage est réitérée jusqu'à obtention d'alignements jugés satisfaisants sur l'échantillon.

Les 30000 couples de chaînes (stimuli,réponse) sont alors soumis au système d'alignement. Cette opération se solde par 4000 recalages.

Il est alors possible de décider des corrections à effectuer dans les chaînes réponses. Elles sont formulées par des règles de réécriture contextuelle.

V. CONCLUSION.

Ce travail a été réalisé dans le cadre du projet SPIN sous financement de la Communauté Économique Européenne par l'intermédiaire du programme ESPRIT.
Ainsi, ont été menées, une comparaison avec une correction manuelle des décalages horizontaux, et une comparaison entre les résultats bruts et ceux après alignement automatique. Les résultats [8] de ces comparaisons démontrent d'une part la validité des règles d'alignement, et d'autre part la nécessité d'appliquer un tel traitement avant d'utiliser les données de l'évaluation.

REFERENCES.

[1] J.P. Lefèvre, "A diphone speech synthesis approach applicable to different languages", IEEE ICASSP Proc., Tokyo, 1986.

[2] J.C. Spohrer, P.F. Brown, R.Roth, "Automatic labeling of speech", IEEE ICASSP Proc., Paris, 1982, pp.1641-1644.

[3] P. Sellers, "An algorithm for the distance between two finite sequences", Journal of Combinatorial Theory,1974, vol. 16(2), pp. 253-258.

[4] R.A. Wagner, Fisher, "The string to string correction problem", Journal of the ACM, 1974, vol. 21(1), pp. 168-173.

[5] K.M. Goudie-Marshall, J.W. Picone, W.M. Fischer, "Phonetic string alignment", 109th meeting of the Acoustical Society of America, Austin Texas, april 1985, pp.1-4.

[6] J. Picone, K.M. Goudie-Marshall, G.R. Doddigton, W. Fisher, "Automatic text alignment for speech system evaluation", IEEE ASSSP Proc., August 86, pp.780-784.

[7] D. Maret , "Comparaisons de chaînes ce caractères, accès lexicaux tolérants et applications", Thèse CRISS-IMSS, Université de GrenobleII, Mai 1987.
A paraître dans Les cahiers du CRISS, no 7.

[8] Pols, J.P. Lefèvre, G. Boxelaar, N. van Son, "Word intelligibility of a rule synthesis system for French", European Conference on Speech Technology, Edinburgh, September 1987.

SYSTEMES DE REFERENCE POUR L'EVALUATION D'APPLICATIONS ET LA CARACTERISATION DE BASES DE DONNEES EN RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Claude MONTACIE, Gérard CHOLLET

ENST Dept. SIGNAL, CNRS UA-820
46 rue Barrault, 75634 PARIS cédex 13, FRANCE.

Abstract

Many data bases have been collected by the scientific community in order to compare Automatic speech recognizers and evaluate their usefulness in specific applications. Their availability on CDROM will facilitate their distribution and therefore their use.

A connected word recognizer developped to evaluate these data bases and serve as a state of the art reference to compare with other recognizers is described in this paper.

Introduction

La qualité d'un système de reconnaissance peut et doit être évaluée. Nous devons distinguer les points de vue du chercheur et du fabricant de systèmes. Le chercheur est surtout interessé par des informations relatives (amélioration dues à un changement d'algorithme, performances comparées à d'autres systèmes ...). Le fabricant de systèmes doit d'abord évaluer la technologie par rapport au marché. Il s'intéresse ensuite à la complexité des algorithmes afin de réaliser une implantation temps réel au meilleur coût.

L'évaluation des algorithmes et la diffusion de bases de données de parole est une pratique nécessaire dont il faut favoriser le développement.

Un taux de reconnaissance dépend de nombreux paramètres [1][2] dont certains sont difficilement quantifiables. On peut citer entre autres: le vocabulaire employé (taille, difficulté, syntaxe, ...), l'environnement (nature et niveau du bruit, bande passante, type et position des microphones, micro ouvert/contrôlé, ...), le locuteur (motivé, professionnel/normal/pathologique, entrainé/novice).

Validité statistique

Il est courant de trouver dans les notices publicitaires des taux de reconnaissance supérieurs à 99%, sans autre précision. Ce chiffre seul ne donne qu'une estimation ponctuelle des performances de la machine. En toute rigueur il devrait être accompagné de la méthodologie des tests, de l'intervalle de confiance à x% (intervalle dans lequel il y a une probabilité de x% de trouver le taux exact).

Pour calculer cette intervalle de confiance, il existe plusieurs méthodes statistiques [3][4] elles donnent des résultats comparables. Nous avons considéré pour notre part que les succès suivaient une distribution binomiale, le calcul de l'intervalle de confiance est dans ce cas assez simple [4].

Soit N le nombre de tests, P le pourcentage de succès, l'intervalle de confiance à x% est:

$$p = \frac{P + \frac{z_x^2}{2N} \pm z_x \sqrt{\frac{P(1-P)}{N} + \frac{z_x^2}{4N^2}}}{1 + \frac{z_x^2}{N}}$$

avec $z_{95\%} = 1.96$
$z_{99\%} = 2.48$

Pour les grandes valeurs de N, cette formule se réduit à :

$$p = P \pm z_x \sqrt{P(1-P) / N}$$

On peut observer que l'intervalle de confiance, pour un même nombre de tests, est maximal pour un pourcentage de succés de 50%, diminue quand le taux de reconnaissance augmente.

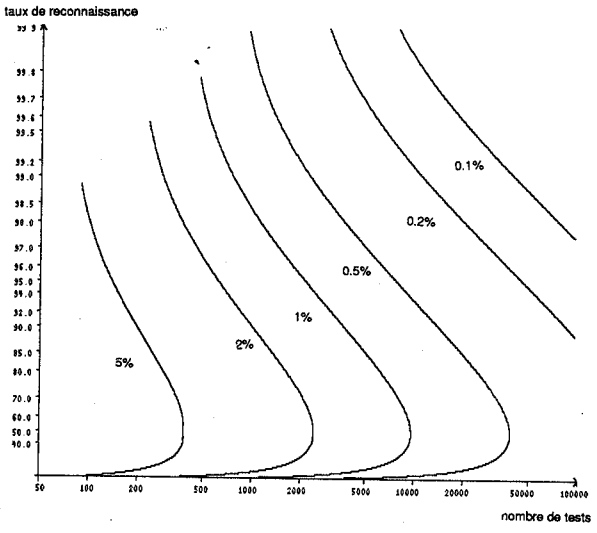


fig1
Précision du taux de reconnaissance en fonction du taux de reconnaissance et du nombre de tests effectués les calculs ont été fait en prenant un intervalle de confiance à 95%.

Exemple: un fabricant fait 200 tests, il obtient 199 succès et une erreur, il annonce un taux de réussite de 99.5%, alors que l'intervalle de confiance à 95% est de (97.2% - 99.9%) soit une précision de 2.3%.

Pour obtenir une précision de 0.2%, un minimum de 8000 tests était indispensable.

Spécification des systèmes d'évaluation et des bases de données de parole

Nous devons pouvoir évaluer toutes sortes de systèmes, nous adapter à leur spécificité :

-mode d'apprentissage (entrée micophone, génération de références, modèles stochastiques ...).

-mode de reconnaissance (mots isolés, enchainés, décodage accoustico-phonétique).

Les bases de données de parole elles mêmes doivent être évaluées sur la complexité du vocabulaire, la qualité des enregistrements, la variabilité intra- et inter-locuteur. Notre but est de créer des bases de données étalonnées qui permettent de prédire les performances d'un système sur une autre base de données. Nous voulons définir des types de locuteurs, éliminer des locuteurs trop semblables. Pour cela nous avons défini une typologie des locuteurs, cela nous permet de diminuer la complexité de l'évaluation.

Pour tester les limites d'un système une autre solution est de créer des tests par déformation temporelle et spectrale de sons de référence. Ceci permet une diffusion plus simple des tests d'évaluation; il suffit alors de transmettre les algorithmes de déformations et un nombre restreint de références.

Systèmes de référence

Une méthode d'évaluation est de se servir de systèmes de réfèrences. Il s'agit d'un système de reconnaissance qui dispose virtuellement d'une mémoire infinie et n'est pas soumis aux contraintes temps réel. Ce système doit refléter l'état de l'art; que ce soit au niveau des algorithmes employés ou du taux de reconnaissance. Il doit offrir une panoplie de représentations du signal de parole et de distances interspectrales. Il doit être facilement implantable pour que de nombreux laboratoires l'adoptent.

Quelques expériences et résultats :

Nous avons developpé un système de référence de reconnaissance de mots connectés. Comme les résultats le démontrent, il reflète bien l'état de l'art. Le langage dans lequel il a été écrit: (C) lui assure une portabilité remarquable; sans modification il a été porté sur trois machines différentes (Vax sous Unix, Vax sous Vms, PC-AT sous Msdos) en moins d'un mois.

Il s'agit de l'extension d'un systéme de référence de mots isolés [5]. La parole est représentée par une suite de vecteurs (MFCC,LAR...) pondérée par leur écart-type. Une grammaire représente la forme des phrases à reconnaitre. Une comparaison dynamique, entre les tests et les réfèrences, conduite par la syntaxe [6] est utilisée pour trouver la phrase prononcée.

Nous ne nous sommes servis, pour l'instant, d'aucun seuil de réjection pour éliminer les mots n'appartenant pas au vocabulaire d'apprentissage [7]. Il n'y a donc que des erreurs de confusion. Il serait souhaitable d'introduire un tel seuil dans les prochaines versions du système de référence.

Ces techniques ont été expérimentées sur la base de données de 20 mots isolés répétés 26 fois par 16 locuteurs (nous n'avons utilisé que 12 locuteurs), enregistée par Texas Instrument et diffusé par le National Bureau of Standards, USA.

Cette base de données de parole est de difficulté moyenne [1] (peu de bruit, prononciation soignée).

- Vocabulaire de la base de données TEXAS.

Mot 1 : Yes
Mot 2 : No
Mot 3 : Erase
Mot 4 : Rubout
Mot 5 : Repeat
Mot 6 : Go
Mot 7 : Enter
Mot 8 : Help
Mot 9 : Stop
Mot 10 : Start
Mot 11 : One
Mot 12 : Two
Mot 13 : Three
Mot 14 : Four
Mot 15 : Five
Mot 16 : Six
Mot 17 : Seven
Mot 18 : Eight
Mot 19 : Nine
Mot 20 : Zero

Pour éviter d'avoir à prendre une décision silence/parole, nous avons considéré que le silence est un mot; ce qui nous donne la gammaire de la fig2.

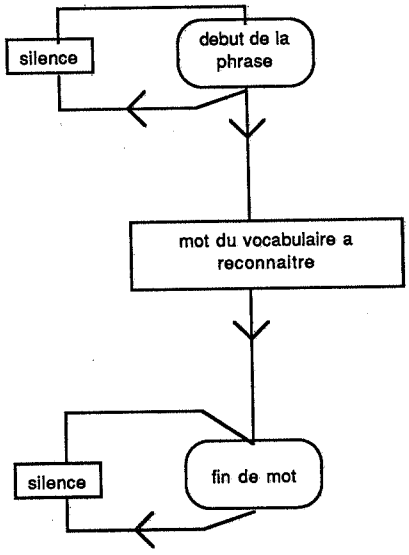


fig2
automate de reconnaissance de mots isolés

Seules les références ont été segmentées par un automate à 4 états sur l'énergie (fig 3) . Ce sont des erreurs de segmentation que proviennent la majorité des échecs du système. Il faut donc soigner l'apprentissage , soit en utilisant une restitution sonore différentielle (écoute du mot segmenté puis du mot non segmenté), soit en utilisant des connaissances phonétiques pour détecter avec précision les débuts et fins de mots.

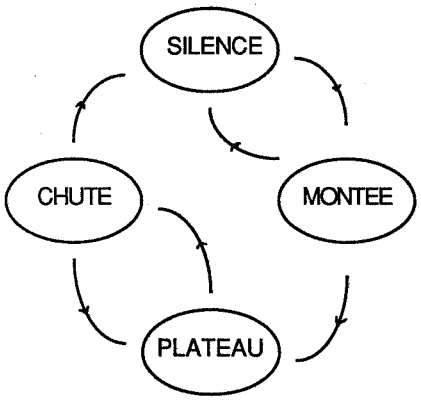


fig3
automate de segmentation de réfèrences
il change d'états en fonction du rapport signal/bruit
et du temps passé dans un même état.

Méthodologie des tests

Nous avons utilisé 9 coefficients MFCC par fenêtres toutes les 20 ms. La largeur des fenêtres est de 41 ms. Le signal a été enregistré à 12.5 Kz et digitalisé sur 12 bits.

Sur les 26 élocutions d'un locuteur, nous avons pris comme référence la première élocution, le mot silence a été créé à partir des trois premiers vecteurs MFCC de l'enregistrement de la première référence. L'ensemble des 21 mots d'une référence (les 20 mots de la base plus le silence) nous permet de calculer une estimation de l'écart-type des coefficients MFCC. Les 25 autres élocutions ont servi de tests, cela pour 12 locuteurs .

Nous avons donc fait 6000 tests, et nous avons enregistrés 109 échecs; ce qui donne un taux de reconnaissance de 98% à 0.5% près. Ce résultat place notre système de référence entre le DP-100 de Nippon Electric et le T-500 de Threshold Technology [8]. La plupart des échecs (fig 4) viennent de la confusion entre le deuxième mot et le sixième mot ("no", "go") due à une mauvaise segmentation des références.

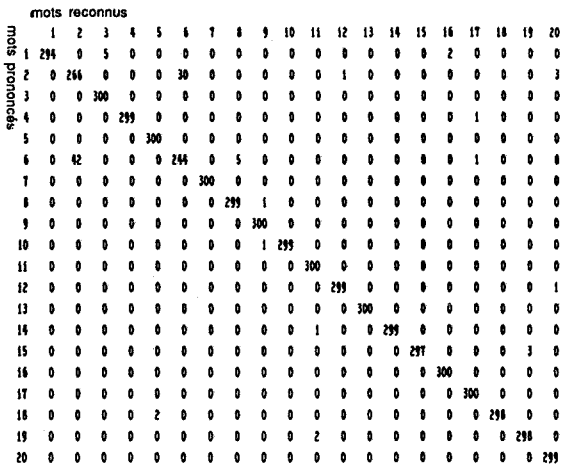


fig4
matrice de confusion

Discussion et Conclusion

Notre souhait est de faire évoluer notre système de référence en y intégrant les meilleurs algorithmes du domaine public.

La méthode de reconnaissance globale, que nous utilisons pour l'instant, a un temps de calcul proportionnel à la taille du vocabulaire. Il nous semble donc souhaitable, pour travailler avec un vocabulaire important (> 100 mots), de nous orienter vers le décodage accoustico-phonétique. Les tests d'évaluation auront lieu sur un PC/AT muni d'une carte traitement du signal.

L'utilisation de tels systèmes de référence va permettre de comparer avec fiabilité les différents systèmes et algorithmes existants, de mesurer objectivement la complexité d'un test.

Il convient pour cela qu'ils soient diffusés largement dans les laboratoires de reconnaissance de la parole.

Références

[1] LEA W. (1982) **Available speech databases for evaluating speech recognizers.** Workshop on Standardization for Speech I-O Technology, Washington

[2] PALETT D.S. (1985), **Automatic Speech recognition performance assessements** National bureau of standart, Wash, USA.

[3] BAKER J.M (1982) **The performing arts - how to measure up.** Workshop on Standardization for Speech I-O Technology, Washington, pp 25-33

[4] LEBART L., FENELON J.P. (1973) **Stastistiques et informatiques appliquées** Dunod, Paris.

[5] CHOLLET G., GAGNOULET C. (1982) **Evaluating speech recognizers and data bases using a reference system.** IEEE-ICASSP, Paris.

[6] BAKER K. (1982) **Assessment of using a reference system for the evaluation of speech and data bases.** Report to National bureau of standart, Wash, USA

[7] SIMPSON C.A., RUTH J.C. (Mars 1987) **The phonetic discrimination test for speech recognizers** Speech Technology, pp 48-93.

[8] DODDINGTON G., SCHALK T. (1981) **Speech recognition: turning theory to practice.** IEEE-SPECTRUM

Bibliographie

CHOLLET G., ASTIER A., ROSSI M. **(1981) Evaluating the performance of speech recognizers at accoustic-phonetic** IEEE-ICASSP.

HIERONYMUS J.L., ENEA H.J. **(1982) Evalutating the performance of commercial recognizers** COMPCON , San Francisco

MOORE K. (1977) **Evaluating speech recognizers.** IEEE-ASSP, pp178-183

PALETT D.S. (1982) **Guidelines for performance assessement of speech recognizers** Workshop on Standardization for Speech I-O Technology, Washington

POOL G.K. (1981) **A longitudinal study of computer voice recognition performance and vocabulaire size** Naval Postgraduate school, Monterey

TAYLOR M.M. (1980) **Issues in the evaluation of speech recognition systems.** Defence & Civil Inst. of Environmental Medicine.

Mécanismes de consultation dans la Base de Données et de Connaissances Parole (BDCParole)

J. Caelen, O. Cervantes, Y. Fernandez
Laboratoire de la Communication Parlée - ICP Unité associée au CNRS
INPG/ENSERG
46, Av. F. Viallet
31038 Grenoble CEDEX

ABSTRACT

The most important problem for based-knowledge ASRS (Automatic Speech Recognition System) is to obtain the "best knowledge". Our aim in this paper, is to specify an expert-aided system through the problem of requests. These requests must allow knowledge reasoning from speech data.

In our approach, expert can interfere with the system at several levels. Therefore, various tasks must be enabled:

-1- selection of informations from the data base according to an experiment plan,

-2- information weighting in terms of probability or annotation,

-3- knowledge entry (quantitative facts or rewritten rules),

-4- selection of kind of reasoning about knowledge: (a) by problem solving in order to assess a priori knowledge, (b) by learning in order to reach a new knowledge,

-5- information filtering before knowledge storage into SDKB (Speech Data-Knowledge Base).

Then, it is possible to define two levels of inquiries:

- "classic" (but complex) requests as SQL language,
- "logic" requests as deductive and inductive processes.

Classic requests provide for each expert a specific experiment base EB (which contains data and rules). From EB, logic requests (expert controlled) provide new knowledge. After filtering, the system stores results into SDKB about experiments (data, knowledge and experiment history).

INTRODUCTION

Un des problèmes parmi les plus importants en décodage acoustico-phonétique (DAP) est de disposer des "bonnes" connaissances (du moins pour les systèmes de reconnaissance automatique de la parole (RAP) basés-connaissance). On peut acquérir ces connaissances de plusieurs manières: (a) à partir du savoir d'un expert, (b) par apprentissage automatique, (c) par des méthodes mixtes mélangeant les approches (a) et (b). C'est cette dernière voie qui nous semble actuellement la plus riche dans la mesure où elle reste contrôlable par l'expert --contrairement à (b)-- et où elle permet de compléter utilement les connaissances acquises par la seule expertise de type (a). Dans cette perspective, un système d'aide à l'acquisition de connaissances [Caelen et al, 86] doit apporter à l'expert des outils pour valider son savoir, pour quantifier certains des paramètres qu'il utilise couramment et pour lui permettre de rechercher plus systématiquement de nouvelles connaissances.

Ce système doit pouvoir:

- gérer des données (sons, paramètres articulatoires, spectres, etc.),
- produire des connaissances à partir de ces données et du savoir (ou du savoir-faire) de l'expert,
- gérer les connaissances produites.

En fait il n'y a pas de véritable frontière entre données et connaissances --appelées ci-après informations-- ce qui permet de les gérer avec un système unique (SGBDCParole) [Cervantes et al, 87] en utilisant un seul formalisme: celui des représentations centrées-objet. Par contre la production de connaissances procède d'un raisonnement [Halpern, 86] sur des faits extraits de la BDCParole, et ce mécanisme dépasse le cadre strict de la gestion de connaissances. Nous proposons donc un système constitué de:

- BDCParole où les données et connaissances sont représentées par des objets et gérées par le SGBDCParole,
- ARCANE-2 qui est un système de production de connaissances à base de règles (Fig. 1).

Les mécanismes de consultation --sujet de cet article-- interviennent donc à plusieurs niveaux dans ce système: (a) pour la gestion proprement dite de la BDCParole, (b) pour l'extraction de faits, (c) pour le contrôle de la production des connaissances.

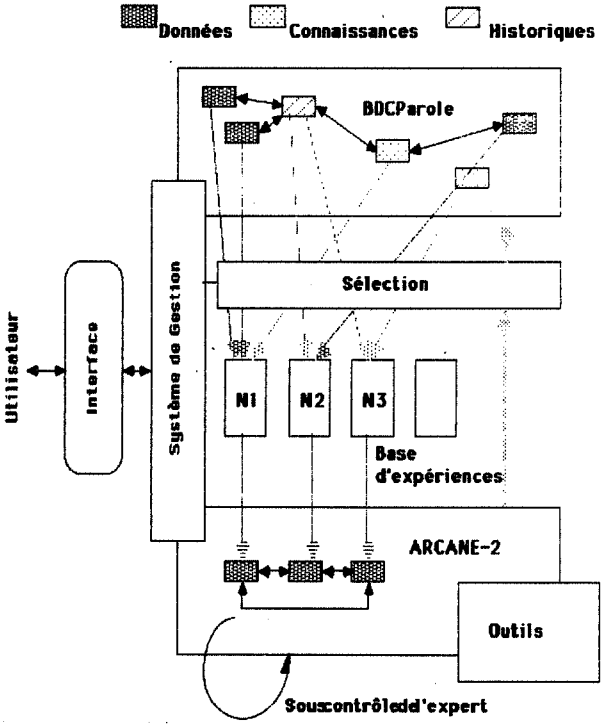


Fig. 1: Structure générale du système BDCParole et ARCANE-2

A- EXPERTISE

Le rôle de l'expert dans un tel système se situe à plusieurs niveaux: (N1) consulter les informations --données et/ou connaissances-- contenues dans la base, (N2) produire de nouvelles informations et les intégrer à la base. Le premier niveau est "classique" dans le domaine des BD, par contre pour le deuxième niveau il est possible de diviser le processus en tâches pour lesquelles:

(T1) il doit extraire les informations de la base selon un certain plan d'expérience: il constitue alors une base d'expériences BE,

(T2) dans cette base BE il peut éventuellement pondérer certaines informations (avec des probabilités, notes, etc.) ou les qualifier par des symboles,

(T3) il peut y introduire ses propres connaissances sous forme de faits (informations quantitatives) ou de règles (informations symboliques),

(T4) il doit préciser le mode de raisonnement du système:

- en résolution de problème: lorsqu'il veut valider sa connaissance --le système tente dans ce cas de faire la preuve de cette connaissance à partir des données. Ce raisonnement est plutôt de type chaînage arrière dans la mesure où l'on connait les buts (conclusions).

- en apprentissage: lorsqu'il veut déduire ou induire une nouvelle connaissance. L'expert divise les observations en deux groupes: les exemples et les contre-exemples et le système tente alors de produire des connaissances par apprentissage sur ces deux ensembles d'information. Ce raisonnement est plutôt de type chaînage avant dans la mesure où l'on ne connait pas les buts a priori mais seulement les prémisses.

(T5) il peut, en fin d'expérience, demander l'archivage de ses résultats. Le SGBDCParole doit alors filtrer et organiser les différentes expertises et en contrôler la cohérence, en vue d'une consultation ultérieure.

A-1. Quelques exemples

-a- en résolution de problème:

Soit à démontrer la propriété : *Trait_aigu(phonème /i/)= '+aigu'* il est clair que pour valider cette connaissance à partir des données, il est nécessaire de créer une base d'expérience BE contenant tous les /i/ d'un corpus donné et convenablement paramétrés pour "calculer" le *trait_aigu* (par une table de fonctions). Si celui-ci prend toujours la valeur *'+aigu'* (relativement à un seuil à préciser) sur cette BE, alors le prédicat est localement vrai --sinon le système doit donner la liste des contre-exemples de manière à suggérer à l'expert une nouvelle expérience.

-b- en apprentissage:

Il s'agit maintenant du problème inverse c'est-à-dire de supposer que ce prédicat soit vrai a priori et d'expliquer ce qu'est *trait_aigu* sur les données du corpus. Plusieurs solutions sont possibles ici selon que le trait_aigu est quantifiable ou non. Dans le premier cas l'analyse des données multidimensionnelle offre un cadre d'optimisation dans lequel il est possible de trouver la meilleure combinaison linéaire de paramètres pour définir le trait_aigu. Dans le deuxième cas les méthodes d'appariemment structurel [Ganascia, 87], [Guizol, 85] peuvent convenir --on peut remarquer d'ailleurs que l'on peut passer du quantitatif au symbolique par quantification vectorielle. Dans un même ordre d'idée, les méthodes d'apprentissage stochastiques comme HMM (Hidden Markov Model) peuvent faire partie de la palette des outils offerte par le système.

A-2. Formalisation du problème

Les différentes interventions de l'expert --tâches T1 à T3-- reviennent dans un premier temps à réunir des faits (observations) à partir des données, puis dans un deuxième temps à y introduire des connaissances pour former une base d'expérience, comme représenté sur la fig. 2. Pour cela l'expert sélecte les données nécessaires à son expérience (observations tirées de la base), les pondère éventuellement à l'aide de modulateurs (symboles, notes, etc.) et y ajoute des prémisses (propriétés qui initialisent la recherche automatique de nouvelles connaissances) ou des conclusions (propriétés à démontrer en chaînage arrière). Le système se charge quant à lui, de produire des connaissances et de générer l'historique de l'expérience.

Il est intéressant de constater que les prémisses, les observations et les conclusions peuvent se mettre sous un formalisme de règle de réécriture dans lequel la partie gauche est un symbole et la partie droite est une structure (scalaire, vecteur, matrice, liste, arborescence...):

$Sd_i/Hd_i \dashrightarrow So_i/Ho_i \bmod \pi_i$, pour i =1,n

$Sd_k/Hd_k \dashrightarrow Se_k/He_k \bmod \pi_k$, pour k=1,m

où Sd est une entité de nature symbolique (classe, nom, etc.) définie sous les contraintes Hd,

So est l'entité observée asssociée à Sd sous les contraintes Ho. Se est l'entité hypothétisée par l'expert dans les conditions He,

Hd, Ho, He sont des historiques qui contiennent les contraintes sous lesquelles les entités correspondantes sont obtenues,

mod π est un modulateur posé par l'expert (probabilité, possibilité ou plausibilité, etc.).

Les historiques sont des objets qui répondent à la syntaxe définie dans [Cervantès et al, 87]

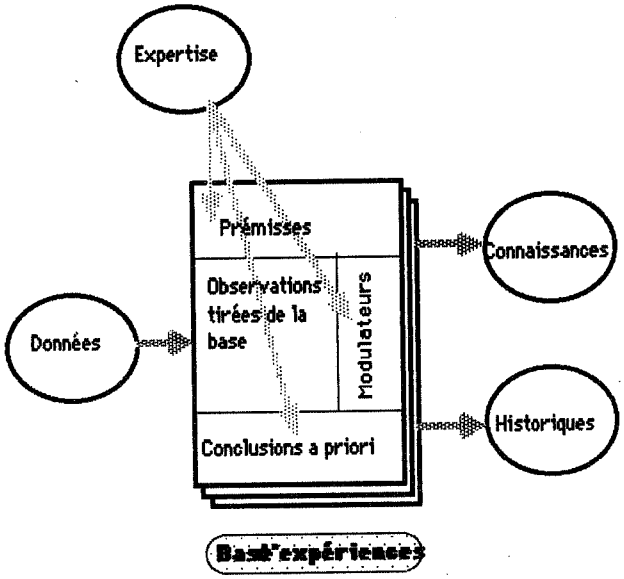


Fig. 2: Structure d'une base d'expériences après sélection des faits par l'expert.

A-3. Exemple

A l'issue de la constitution de la BE, on pourrait avoir:

/i/ --> I=matrice individu/variable (au sens de l'analyse de données),

/autres voyelles/ --> V=matrice individu/variable,

modulateurs = équiprobabilité,

conclusion: trait_aigu(/i/)=+aigu,

$H_{/i/}$=conditions avec lesquelles les /i/ et /autres voyelles/ ont été extraits des corpus,

$H_{/mat/}$=définition des variables et mode de calcul,

He=discriminateur linéaire.

Dans ces conditions le problème est: trouver quelles variables permettent de séparer au mieux /i/ des autres voyelles ou: existe-t-il un hyperplan séparant les classes /i/ et /autres voyelles/. La connaissance produite est en fin de compte une donnée c'est-à-dire la liste des coefficients de la direction orthogonale de cet hyperplan s'il existe.

B- LA CONSULTATION ET LES REQUETES

Le rôle de l'expert dans un tel système est primordial: il faut donc lui offrir une bonne logistique (=ensemble de logiciels) pour gérer au mieux ses expériences. Pour cela il doit disposer de mécanismes de consultation performants mis en oeuvre à travers un langage externe de requêtes. Au sens large du terme, une requête est une question posée au système pour rechercher ou produire une information.

On distingue deux niveaux de requêtes:

(a) la requête de sélection qui conduit à sélecter un objet de la base et à instancier un ou plusieurs de ses attributs --notons ici que le chemin d'accès aux données peut-être particuliérement compliqué,

(b) la requête logique, qui s'accompagne d'un calcul logique (déduction, induction, etc.) voire d'un raisonnement sur les

données et qui suit généralement une série de requêtes de sélection.

B-1. Le 1er niveau: la SELECTION

Par SELECTION on désigne toutes les requêtes ne faisant intervenir que des calculs procéduraux après localisation des informations utiles dans la base. Les exemples donnés ci-après sont significatifs des requêtes formulées par un utilisateur du système. Ils nous permettront de préciser le modèle général défini par la suite. Dans ces requêtes les mots-clefs sont soulignés, ils nécessitent parfois une définition explicite lorsque l'utilisateur désire en préciser le sens. On suppose dans la suite que le système dispose d'un module de dialogue --Q= question du système, R=réponse de l'utilisateur-- géré à travers une interface utilisateur. On distinguera deux cas dans le dialogue selon que les mots-clefs sont des objets déjà définis (attribut global) ou sont des objets qui demandent une définition explicite (attribut local). Nous ne ferons pas de distinction entre les requêtes d'interrogation de la base et celles qui permettent de réunir les faits (le verbe actif de la requête serait *produire* au lieu de *quel est*).

Exemple 1: *Quelle est la valeur du 1er formant de l'échantillon 12?*

Après formulation d'une telle requête le système peut demander des précisions sur les mots-clefs: (a) leur définition ou rattachement à d'autres définitions --nous ne parlerons pas ici des définitions récursives-- (b) les procédures mises en jeu dans l'instanciation des variables, (c) la définition en clair de ces procédures, (d) l'indirection vers une autre formulation. Par exemple, le dialogue peut se dérouler de la manière suivante:

Q1= *valeur* ? R1=instanciation numérique=affectation
Q2= *1er formant*? R2=désigné sur spectre
(ou R2=appel proc. prédéfinie
ou R2=? puis Q3=*procédure*? R3=LPC + crêtes)
Q3=*échantillon 12* ? R3=spectre à la clef 12

Au niveau de l'exécution, les tâches suivantes seront activées:

début
1. sélection du spectre à la clef 12 (ou calcul du spectre s'il n'est pas mémorisé)
2. affichage du spectre et attente curseur
3. calcul de la position curseur et instanciation de l'attribut 1er_formant.
fin

Exemple 2: *Y a-t-il un pic sur la courbe d'intensité des phonèmes /p/ en finale de mot ?*

Cette requête implique une localisation de l'information à l'aide de la forme "pic".
Q1=*pic*? R1=déf. globale=forme prédéfinie
(ou R1=? alors Q2=forme ? R2=triangulaire)
Q2=*courbe d'intensité*? R2=objet=corrélat_prosodique, attribut =intensité
Q3= *phonème /p/*? R3=objet délimité par étiquette 'P'
Q4= *finale de mot* ? R4=déf. locale=suivi_par_pauseApause =etiq. 'X'

Exemple 3: *Quelle est la valeur de Fo du phonème /i/ dans les syllabes accentuées pour le locuteur DG ?*

Cette requête demande la définition du concept "syllabe accentuée" qui peut être celle de "syllabe" et "accentuation" (deux attributs prédéfinis) ou un concept nouveau comme "syllabe ayant un Fo élevé ou une durée longue", etc.
Q1=*valeur*? R1=instanciation numérique=moyenne
Q2=*Fo*? R2=calculé_par=procédure AMDFAmin_pic
Q3=*phonème /i/*? R3=VoyelleAFormants_écartés
Q4=*syllabe accentuée*? R4=2 termes
Q5=*syllabe*? R5=Liste_de phonèmes=(CV,CCV,CVC)
Q6=*accentuée*? R6=?
Q7=critères ? R7=Fo > 2*Fréq_base & Durée > 2*D_ min
Q8=*locuteur DG*? R8=déf. globale=répertoire

Le déroulement des tâches à exécuter pour satisfaire cette requête est ici la suivante:

début
1. sélection du répertoire DG
2. localisation des syllabes dans DG
3. sélection des /i/ dans syllabes accentuées
4. calcul max Fo pour ces /i/ par AMDF sur durée de chaque /i/
5. calcul moyenne Fo
fin

Une extension du langage SQL permet de formaliser ce type de requêtes:

```
SELECT <item> FROM <Espace>
              WHERE <Conditions>
              WITH <item> IS <définition>
                   <espace> IS <définition>
                   <conditions> IS <définition>
PERFORM <résultat> FROM <item>
                   WHERE <conditions> IS <définition>
```

<item>, <Espace>, <Condition> sont des champs qui sont remplis après analyse du dialogue avec l'utilisateur. <Définition> est un objet (ou méthode) ou un ensemble de règles contenant des constructeurs logiques. <Resultat> contient le résultat de la requête.

B-2. Le 2ème niveau: la DEDUCTION et l'INDUCTION

Avec ce deuxième type de requêtes on entre dans le champ de la logique. De manière générale on appelera DEDUCTION toute information nouvelle déduite des informations contenues dans la base. On appelera INDUCTION toute information nouvelle obtenue par généralisation à partir de cas particuliers.

B-2.1.La déduction

Comme ci-dessus quelques exemples permettront d'illustrer les points les plus importants.

Exemple 4: *Sachant que /a/ a un 1er formant d'environ 600 Hz, l'échantillon 12 est-il un /a/ ?*

La logique du 1er ordre permet de formaliser ce problème. Posons:
P=(F1<600+ε & F1>600-ε), ε donné
et Q=(phonème=/a/),
alors le problème peut se résoudre par:
sachant que P => Q, si P alors Q
ce qui revient à vérifier P pour l'échantillon 12 --on peut aussi utiliser les logiques floues comme variantes pour évaluer P.

Exemple 5: *Existe-t-il des /i/ compacts ?*

soient P(x)=(phonème /x/ est compact)
le problème se résoud par:
∃x ? tq P(x) & (x=i)

Exemple 6: *Sachant que l'éch. 1 est un /a/, l'éch. 12 est-il un /a/?*

Ici, contrairement à l'ex. 4, rien ne dit pourquoi l'éch.1 est un /a/. Il faut donc définir l'opérateur "analogue" noté ◇. Le problème peut alors se formuler comme en logique du 1er ordre:
Sachant P ◇ Q, si P alors Q
cet opérateur d'analogie peut recouvrir par exemple, une mesure de ressemblance entre spectres.

Pour prolonger cette idée, on peut utiliser aussi les logiques modales pour pondérer le savoir des experts les uns par rapports aux autres. On peut par exemple introduire une modalité épistémique au moment de l'étiquetage phonétique --qui sous-tend une modélisation phonétique non nécessairement partagée par les autres experts-- : X croit que P noté CX(P) et X sait que P noté SX(P). Si X fait autorité en la matière alors SX(P) => P sinon il est simplement possible que P. On dira généralement que P est vrai avec la plausibilité $\pi(X)$ attachée à l'expert X (noté P mod $\pi(X)$). La

modalité temporelle est aussi intéressante dans la mesure où les données sont évolutives et où l'on veut maîtriser le paramètre temps.

B-2.2. L'induction et l'apprentissage

L'induction et l'apprentissage se rejoignent à travers les méthodes fondées sur les exemples et contre-exemples [Michalski, 83]. Pour produire une connaissance à partir des données --du moins dans une approche contrôlée par un expert-- il faut partir d'un plan d'expérience et d'un noyau de données, ce qui revient à réunir des faits puis à raisonner sur ces faits pris comme exemples, par des mécanismes inductifs. La validation des connaissances ainsi produites est en fait une demande d'explication ou trace déductive sur un autre ensemble de données considérées comme contre-exemples.

Exemple 7: *L'indice FO des voyelles est-il corrélé avec l'aperture?*

Dans ce cas il s'agit bien d'induction puisque l'on cherche à généraliser une propriété des voyelles. La corrélation n'étant jamais strictement égale à 1, il faut un critère de décision dépendant de l'expert, du type de données, etc., pour répondre à la requête. Les éléments du dialogue sont ici:

Définition des items de la requête:
- Q1=*Indice FO*? R1=objet=Indice, attribut=FO
- Q2=*voyelle*? R2=déf. globale=étiq. 'V'
- Q3=*corrélé*? R3=procédure prédéfinie
- Q4=*aperture*? R4=objet=Corrélat, attribut=articulatoire n°2

Sélection des faits:
- Q5= *Exemples* ? R5=1ère moitié du corpus
- Q6= *Contre-exemples* ? R6=2ème moitié du corpus
- Q7= *Prémisses* ? R7=néant
- Q8= *Conclusions* ? R8=Corrélation(Indice_FO,aperture)
- Q9= *Modulateurs* ? R9=équiprobabilité

Choix du type de raisonnement:
- Q10= *Méthode*? R10=objet=Ana_Données, Attribut=corrél.
- Q11= *Décision* ? R11=supérieur_à, variable=seuil

Exemple 8: *Comment définir un indice FO pour les voyelles tel qu'il soit corrélé avec l'aperture ?*

Ce cas diffère du précédent car la proposition "X est corrélé avec Y" est une prémisse. La généralisation à effectuer porte sur l'indice FO qui est considéré comme une fonction que l'on cherche à travers l'une de ses propriétés. Cela revient à inverser dans le dialogue les réponses R7 et R8 et modifier les réponses R1=combinaison_lin(canaux_spectre) et R10=objet=Ana_Données, Attribut=régression_linéaire.

Pour ces deux exemples on voit clairement qu'il faut:

-a- réunir les faits
-b- les répartir en classes (exemples et contre-exemples ou aut
-b'- éventuellement les pondérer
-c- leur appliquer une série de méthodes ou de raisonnemei
(ici l'analyse de corrélation et la régression linéaire)
-d- confronter les résultats obtenus sur les différentes classes
-e- conclure selon un critère de décision.

Sans multiplier exagérément les exemples, il est suffisamment clair maintenant qu'apparait la notion de plan d'expérience: c'est la séquence énumérée ci-dessus que doit définir l'expert de manière explicite. Il ne peut donc plus s'agir ici d'une simple extension d'un langage de formulation de requêtes comme dans le paragraphe précédent (SQL), il faut un interpréteur de schémas logiques. On notera que les requêtes de premier niveau qui permettent de réunir les faits, point -a-, font partie de ce schéma plus général.

Au delà des "expériences", pour rendre utilisable les connaissances acquises par un expert à l'ensemble des utilisateurs, il faut mémoriser les conclusions mais aussi les historiques (ou traces des expériences) pour les replacer dans leur contexte d'obtention.

CONCLUSION

La réflexion faite dans cet article est guidée par une utilisation "intelligente" de la base de données BDSons et plus généralement d'une BDCParole dans le but d'accumuler des connaissances sous le contrôle des experts tout en particularisant leurs rôles respectifs. A travers le problème de la consultation par requêtes se dégage plusieurs fonctionnalités que doit posséder le système de gestion doublé du système d'aide à l'acquisition des connaissances:

1. la possibilité de réunir les faits (interpréteur de requêtes SQL étendues),
2. la possibilité de raisonner sur ces faits par déduction ou induction (couche logique sur le SGBD),
3. la possibilité de choisir un plan d'expérience que doit contrôler l'expert (interpréteur logique),
4. la possibilité de filtrer la connaissance obtenue et de gérer les historiques des expériences.

L'expert conduit donc des expériences qui répondent au schéma d'exécution suivant:

1. interprétation de la requête
2. sélection des données, construction de la base d'expérience
3. raisonnement, apprentissage
4. décision
5. archivage de la connaissance produite, gestion des historiques

BIBLIOGRAPHIE

[Caelen et al, 86] J. Caelen, G. Caelen-Haumont, N. Vigouroux, C. Barrera, J. Malet.
ARCANE: Acquisition et Recherche de Connaissances Acoustico-phonétiques dans un Noyau Evolutif. 15èmes JEP, Aix-en-Provence, pp. 207-211, 1986

[Cervantes et al, 87] O. Cervantes, J.F. Sérignat
Représentation objet dans BDCParole (Base de Données et de Connaissances pour la Parole). 16èmes JEP, Hammamet, 1987

[Ganascia, 87] J.G. Ganascia
Agapé: de l'appariement structurel à l'apprentissage. Intellectica, vol. 1, N°2/3, pp. 6-27, 1987.

[Gascuel, 97] O. Gascuel
Plage: un outil pour construire des systèmes d'apprentissage. Intellectica, vol. 1, N°2/3, pp. 28-47, 1987.

[Guizol, 86] J. Guizol
Apprentissage inductif de règles pour le décodage acoustico-phonétique. 15èmes JEP, Aix-en-Provence, pp. 227-230, 1986.

[Halpern, 86] J.Y. Halpern
Reasoning about knowledge: An overview. IBM Research Report RJ5001, 1986.

[Hayes-Roth, 78] F. Hayes-Roth, J. Mc Dermott
An Interference Matching Technique for Inducing Abstractions. Communication of the A.C.M., Vol. 21, n° 5, pp. 401-410, 1978.

[Michalski, 80] R.S. Michalski
Inductive Learning as Rule-Guided Generalization and Conceptual Simplification of Symbolic Descriptions. Workshop on Current Developments in Machine Learning, CMU Pittsburgh, 1980.

[Michalski, 83] R.S. Michalski
A Theory and Methodology of inductive learning. Artificial Intelligence 20, pp. 11-161, 1983.

[Roulle et al, 87] A. Roulle, J. Quinqueton
Dialogue pour l'apprentissage. Intellectica, vol. 1, N°2/3, pp.178-194, 1987.

REPRESENTATION CENTREE OBJET DANS LA BASE DE DONNEES ET DE CONNAISSANCES PAROLE

O. Cervantes (*), (**), J.F. Serignat (*)

(*) Laboratoire de la Communication Parlée - ICP Unité associée au CNRS
INPG/ENSERG 46, Av. Félix Viallet 38031 Grenoble Cedex

(**) Laboratoire de Génie Informatique - IMAG Université de Grenoble
B.P 68 38402 St Martin d'Hères Cedex

ABSTRACT

The Speech Object concept with a complex and voluminous structure is presented. This concept arised by going deeper into the study of the French Speech Data Base Management System (BDSON) to integrate new data into the base, associated with speech signal, as labels (lexical, syntactic, semantic, prosodic,phonetic, phonologic,...) and processing results. We propose to build a Speech Data and Knowledge Base (BDC-Parole) upon an object oriented model. This model allows to describe and manage the speech objects and to define semantic links between them. Furthermore, it facilitates facts obtaining and rules development from statistic analysis of the base contents. These facts and rules will be incorporated into a Speech Knowledge Base.

I. INTRODUCTION.

L'évaluation de l'expérience réalisée sur BDSON à l'aide d'un Système de Gestion de Bases de Données (SGBD) Relationnel [1], [2] a montré que ce type de modèle de données était bien adapté pour la gestion des DESCRIPTEURS des fichiers-sons. Nous avons développé des applications permettant l'accès aux sons avec des conditions portant sur les caractéristiques des corpus, des locuteurs, des réalisations (enregistrements) ainsi que sur les transcriptions orthographiques et phonétiques de chaque corpus.

Dans l'étude pour enrichir les fonctionnalités d'un Environnement de Recherche sur la Parole [3], nous avons mis en évidence la notion de l'**OBJET PAROLE** avec toutes les caractéristiques d'un objet COMPLEXE [4], [5] possédant :

- une structure complexe (souvent hiérarchique),
- une grande taille,
- des contraintes d'intégrité non triviales,
- des aspects dynamiques,
- des opérateurs spéciaux.

Le modèle relationnel a montré qu'il pouvait assurer la dynamicité des schémas, l'indépendence entre les données et les programmes et un minimum de redondance d'informations [6]. Mais il présente aussi des insuffisances pour l'exploitation de l'information sémantique de l'application. Ses lacunes sont la pauvreté d'expression de contraintes et de liens sémantiques ainsi que l'éclatement de l'information, qui rend difficile la modélisation des objets complexes [7], [8]. Il a donc été nécessaire de proposer un formalisme pour la manipulation des OBJETS-PAROLE (signal numérique, étiquettes syntaxiques, lexicales, phonétiques, phonologiques, prosodiques, résultats des traitements...), ainsi qu'un ensemble de fonctionnalités adaptées à la gestion d'un Environnement de Recherche sur la Parole.

Les caractéristiques d'un modèle de gestion adapté aux données et aux connaissances parole (BDC-PAROLE), sont énoncées dans la section II. Nous décrivons dans la section III, la structure et les éléments de base de cette BDC-PAROLE. Enfin, dans la section IV nous présentons les primitives pour la gestion des objets de la base.

II. CARACTERISTIQUES D'UN MODELE DE GESTION ADAPTE AUX DONNEES ET AUX CONNAISSANCES PAROLE.

Nous présentons ci-après, les caractéristiques que doit posséder un modèle de gestion (représentation et manipulation) adapté aux OBJETS-PAROLE.

a) GESTION DE DONNEES DE NATURE VARIEE

Dans la recherche sur la parole, on est amené à gérer simultanément des données de nature différente

- le signal numérique, résultant de la digitalisation des enregistrements sonores,
- des descripteurs, donnant le détail sur le contenu orthographique et phonétique et sur les caractéristiques des enregistrements,
- des étiquettes, qui repèrent des discontinuités ou des zones stables dans le signal. Elles peuvent être : événementielles, phonétiques, lexicales, syntaxiques, prosodiques, etc.
- des résultats de programmes de traitement : spectres, intensité, énergie, fréquence fondamentale, indices acoustiques, unités linguistiques,etc.

Toutes ces données possèdent des structures et des caractéristiques diverses et leur taille est variable. Elles sont stockées parfois sur des supports différents: bande magnétique, cassette-vidéo, disque compact, etc. D'où leur nature multi-média.

Par ailleurs, les connaissances à manipuler peuvent être déclaratives ou procédurales. Il est donc nécessaire d'avoir un formalisme UNIQUE de représentation pour les données et les connaissances.

b) DYNAMICITE, au niveau :

- de l'objet lui-même, offrant les mécanismes pour l'extension de sa structure et de sa mise à jour, en garantissant l'intégrité et la cohérence de la base.
- du schéma de la base, afin de l'enrichir avec de nouveaux objets, définis selon les besoins de l'utilisateur.

c) GESTION DE LIENS SEMANTIQUES.

Dans un environnement de recherche sur la parole, il existe plusieurs types de liens ou de relations entre les objets qui servent à exprimer la sémantique de l'application. En particulier, il est important de pouvoir modéliser :

- l'aspect évolutif de la création d'objets nouveaux ou dérivés, à partir de ceux déjà existants dans la base.
  => liens de **COMPOSITION**
- la notion d'équivalence conceptuelle entre deux ou plusieurs objets de la base.
  => liens d'**EQUIVALENCE**

d) COMPLEXITE DES REQUETES.

Le phonéticien, principal utilisateur de la base pour l'analyse et l'interprétation des données, n'est pas familier avec les langages de programmation. Il faut donc lui offrir un langage de manipulation des données suffisamment proche du langage naturel et très performant car les requêtes formulées sont souvent complexes. Par exemple, "Sélectionner

les voyelles du corpus PEQ* pour les syllabes accentuées". Une telle requête met en jeu des processus complexes et des mots-clés prédéfinis (voyelles, syllabe, accent) qui sont soit des instances, soit des traitements, soit encore des objets traités avec prise de décision (syllabe accentuée). Il s'agit là de processus complexes pour lesquels il faut mettre en oeuvre une base de règles et un résolveur de problèmes.

e) GESTION DE GRANDS VOLUMES DE DONNEES.

Afin de récupérer des données pour une étude particulière, l'utilisateur est amené à effectuer des recherches qui portent sur de gros volumes de données

- Soit pour la recherche d'objets possédant une sous structure similaire.
- Soit pour la recherche d'objets ayant les mêmes propriétés.

f) GESTION D'OUTILS STATISTIQUES ET SUPPORT POUR LA CREATION DE BASES DE CONNAISSANCES.

Un ensemble d'outils statistiques doit pouvoir opérer sur les objets gérés par la base, en permettant à l'utilisateur d'obtenir des analyses effectuées sur le contenu général de la base. Ces outils doivent permettre l'extraction de faits ainsi que l'élaboration de règles pour la Base de Connaissances.

EN CONCLUSION, un modèle de données adapté aux problèmes de la recherche dans le domaine de la PAROLE doit permettre :

* la modélisation d'objets possédant des structures COMPLEXES et VOLUMINEUSES
* la modélisation des TRAITEMENTS réalisés sur ces structures
* l'expression de contraintes et de liens sémantiques entre les objets.

Nous proposons alors la structuration de la Base de Données et de Connaissances Parole (BDC-Parole) sur un modèle centré "OBJET", qui s'adapte mieux aux besoins de représentation et de manipulation des données et des connaissances parole. Cette orientation s'approche plus fidèlement de la vision qu'a l'utilisateur sur les données qu'il manipule et qui représentent des entités du monde réel. De même, cette approche permet:

- d'exprimer les aspects STATIQUE et DYNAMIQUE de l'application de manière intégrée avec des primitives qui portent sur les différents objets définis dans la base.
- l'utilisation d'un même formalisme pour la représentation et la manipulation des données et des connaisances.

III. DESCRIPTION DE LA BASE DE DONNEES ET DE CONNAISSANCES PAROLE (BDC-Parole)

La BDC-Parole offre un ensemble de fonctionnalités pour la gestion intégrée des corpus de SONS et des résultats de traitements réalisés sur eux. Cette intégration donne à l'utilisateur la possibilité d'avoir une vision globale des données qu'il manipule, en lui permettant de gérer les objets d'une manière indépendante ou en suivant les relations existantes entre eux.

La BDC-Parole est contituée d'un ensemble d'OBJETS auxquels on associe un ensemble de propriétés. Les objets peuvent être génériques ou bien des instanciations de ces objets génériques. Les objets partageant des propriétés et des comportements communs sont regroupés dans des CLASSES. Les relations entre les objets s'expriment à travers des liens sémantiques.

Les objets de la BDC-Parole sont répartis en trois catégories :

---> les OBJETS-PAROLE, qui correspondent aux données décrivant la PAROLE selon une certaine représentation ou aspect (vue) et qui, ensemble, contribuent à approfondir (mieux décrire) les connaissances (acoustiques, phonétiques, phonologiques, lexicales, syntaxiques et sémantiques) sur le phénomène de la Communication Parlée.

Ces divers aspects dépendent du domaine où l'on se place. Pour un phonéticien un signal n'est pas un objet mathématique. Sur ce signal il placera des étiquettes phonétiques et prosodiques. Inversement, le mathématicien établira les propriétés mathématiques du signal pour appliquer les méthodes utiles au phonéticien dans sa tâche d'analyse. En effet, ces VUES sont complémentaires et elles décrivent le même objet.

Les OBJETS-PAROLE contiennent des informations sur :

- les caractéristiques des corpus et des locuteurs,
- les descriptions orthographique et phonétique de chaque corpus,
- les réalisations (signaux) produites à l'enregistrement,
- les étiquettes (phonétiques,phonologiques, lexicales, syntaxiques, prosodiques, ...)
- les résultats produits par l'exécution de programmes de traitement, par exemple FFT, Cepstre, indices acoustiques, etc.

---> les TRAITEMENTS, qui représentent les diverses transformations (par l'exécution d'un programme de calcul, par l'exécution d'une requête de sélection ou par l'intervention de l'expert) que l'utilisateur peut réaliser sur les OBJETS-PAROLE afin d'obtenir de nouveaux objets dérivés.

---> autres OBJETS, qui ne représentent pas en eux-mêmes un aspect de la parole, mais qui aident pourtant à la définition des OBJETS-PAROLE. Certains de ces objets servent à garantir le contrôle de la cohérence et de l'intégrité des objets de la base en général. On inclut ici des objets comme "date", "fichier", "descripteur", etc. Ils suivent les mêmes règles pour leur définition et leur manipulation que les objets-parole et les traitements.

Tout OBJET dans la BDC-Parole peut être défini par 4 sections :

* IDENTIFICATION
  => lui donnant une existance UNIQUE dans la base
* ATTRIBUTS
  => donnant la description de ses caractéristiques structurelles
* LIENS-SEMANTIQUES
  => exprimant ses relations avec les autres objets de la base
* CONTRAINTES D'INTEGRITE
  => restrictions mettant en relation deux ou plusieurs attributs de l'objet ou portant sur lui en tant qu'entité indépendante

Nous présentons ci-après les mécanismes qui permettent de définir chacune des composantes des objets de la BDC-Parole, d'organiser les objets dans des classes et de gérer l'héritage des propriétés entre elles. Les mots-clefs sont précédés par le symbole %.

III.1 CLASSES ET INSTANCES

L'ensemble des objets qui constituent l'univers de la BDC-Parole est composé par des OBJETS-GENERIQUES représentant les CLASSES et des INSTANCES de ces objets génériques.

Dans la partie IDENTIFICATION des objets de la BDC-Parole, on détermine le nom de l'objet (unique dans la base) et s'il s'agit d'un objet générique ou d'une instance.

Ceci est indiqué par les expressions :

%sorte-de : <nom-obj-générique>
-> pour la définition d'objets génériques

%est-un : <nom-obj-générique>
-> pour la création d'instances

Le lien %sorte-de est établi entre une paire de classes et le lien %est-un entre un individu et la classe à laquelle il appartient. Lorsque l'objet en cours de définition n'hérite d'aucune propriété

d'autres objets, il est déclaré tout simplement comme %sorte-de : OBJET. Il suivra ainsi la structure établie par la définition de tout objet, avec les quatre sections possibles.

L'expression %sorte-de sert à réaliser une organisation de concepts représentés par les objets de la base, sous forme d'une hiérarchie. Elle permet l'héritage de propriétés entre CLASSES (objets génériques). Cette possibilité facilite la déduction de nouvelles connaissances à partir des définitions des objets existants dans la base [9],[10].

Il n'y a pas de différence structurelle entre l'objet générique et ses instances. Celles-ci possèdent les mêmes attributs que ceux qui apparaissent dans la description de l'objet générique. Alors que les attributs d'un objet générique sont décrits par une liste de facettes (voir description d'attributs dans III.2), les atttributs d'une instance ne sont précisés que par leurs seules valeurs.

Une instance pourraît être incomplète : tous les attributs de sa classe n'ont pas forcément de valeur. Les valeurs manquantes peuvent éventuellement être obtenues par déduction ou calcul réalisés par le système.

La figure 1 montre un exemple des liens %sorte-de pour un sous-ensemble des objets-parole.

III.2 Les ATTRIBUTS.

Les ATTRIBUTS décrivent les propriétés statiques de l'objet et ils n'ont pas une existance indépendante de celle de l'objet. Leur nombre est variable et dépend de l'objet en question. Les attributs peuvent aussi être simples ou le résultat d'une agrégation. Lorsque l'attribut en cours de définition est structuré et nullement utilisé ailleurs, il ne mérite pas d'être défini indépendamment. Pour cela, il est possible de le définir de la façon suivante :

```
$ <nom-attribut-agrégat>
    ┌─ début-att
    │
    │   liste d'attributs simples
    │    qui composent
    │   l'attribut agrégat
    │
    └─ fin-att
```

Chaque attribut simple est référencé en indiquant:
<nom-attribut-agrégat>.<nom-attribut-simple>

Chaque attribut a un nom et il est décrit par une liste d'aspects ou facettes [10]. Dans un attribut, les facettes peuvent apparaître dans un ordre quelconque. Elles servent à :

- déterminer le type de l'attribut, qui pourraît être réel, entier, chaîne de caractères, booléen ou symbole (nom d'un autre objet).
  -> %un, %liste-de

- déterminer des restrictions sur le type,
  -> %domaine, %intervalle, %à-vérifier, %card-min, %card-max

- déterminer sa valeur,
  -> %valeur-cte, %défaut, %attach-proc

- exprimer des commentaires,
  -> %comm

III.3 Les LIENS SEMANTIQUES

Nous présentons ici les mécanismes de représentation des liens de COMPOSITION et d'EQUIVALENCE.

--> LIENS DE COMPOSITION. Ils s'établissent entre l'objet en question et les objets qui ont été nécessaires pour sa création. Ces liens ne doivent pas être confondus avec ceux qui permettent l'héritage des propriétés entre les classes. Les liens de composition matérialisent les prérequis (ordonnancement) à la création d'un objet.

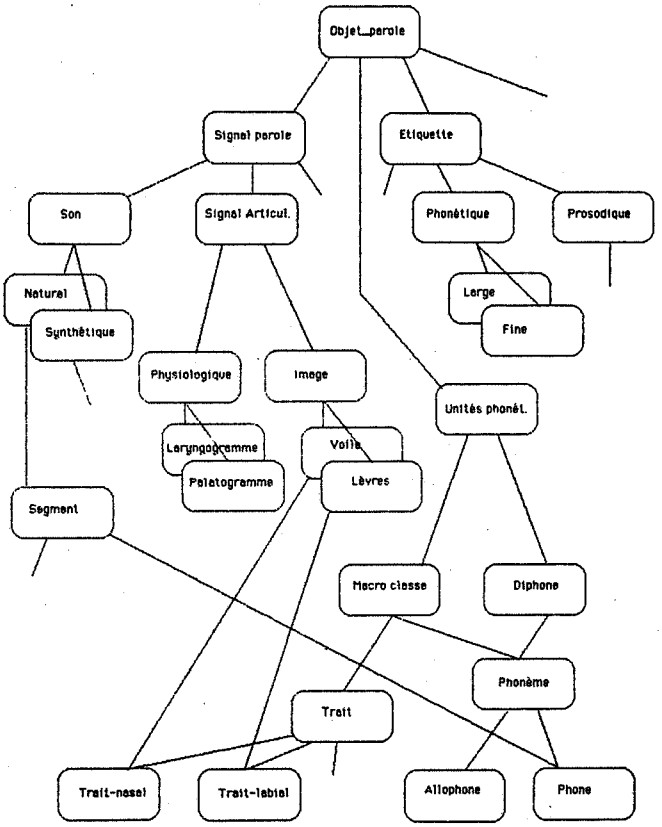


Figure 1
Exemple des liens %sorte-de

Il existe des liens de COMPOSITION DIRECTS (lorsque l'objet est obtenu par l'application d'une méthode de traitement/calcul), et des liens de COMPOSITION D'EXPERTISE (lorsque l'objet est produit avec l'intervention de l'expert)

Ces liens sont notés par :
%lien-comp (<type-comp>, <nom-objet>, <obj-lié>)

Où : .<type-comp> exprime le type de lien de composition
.<nom-obj> désigne l'objet en cours de définition.
. <obj-lié> pourraît représenter :
- un seul objet,
- un choix parmi des objets,
- une liste d'objets ou une combinaison des précédents.

--> LIENS D'EQUIVALENCE. Ils s'établissent entre deux objets représentant un aspect (vue) d'un même objet, pour exprimer la notion d'équivalence entre eux. Ces liens sont aussi utilisés pour exprimer la complémentarité des différentes représentations dans la définition conceptuelle de l'objet.

Ils se dénotent par :
%lien-équiv (<nom-obj>, <obj-équiv>)

Où : . <nom-obj> désigne l'objet en cours de définition.
. <obj-équiv> est le nom de l'objet équivalent.

Dans la section de la déclaration des liens sémantiques de l'objet générique, on peut trouver plusieurs liens d'équivalence pour le même objet.

La figure 2 montre un exemple de LIENS SEMANTIQUES.

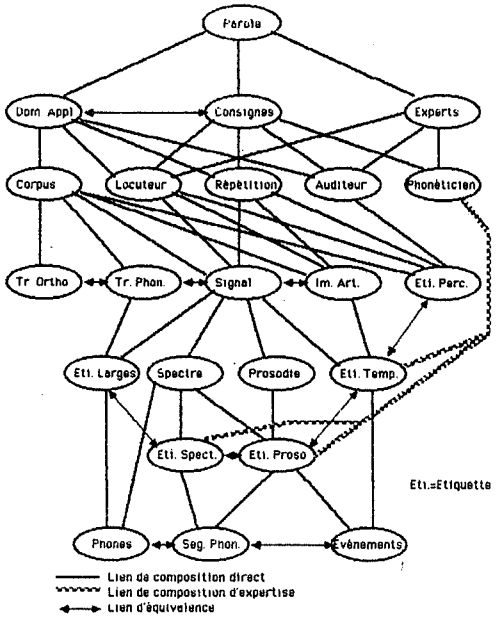


Figure 2
Exemple de liens sémantiques

III.4 CONTRAINTES D'INTEGRITE

Certaines contraintes d'intégrité portant directement sur les valeurs des attributs, d'une manière individuelle sont exprimées à travers les facettes. Pourtant, il reste un autre type de contraintes à déclarer. Celles-ci portent sur l'objet entier en tant qu'unité sémantiquement définie. Elles expriment des relations entre les attributs qui décrivent l'objet générique.

Ce type de contraintes est défini à l'aide de prédicats qui doivent être vérifiés à chaque opération réalisée sur l'objet. Ces prédicats s'ajoutent à la suite des liens sémantiques.

Un module spécialisé dans la gestion de l'intégrité tient compte de ces prédicats et de l'ensemble des contraintes sur les attributs exprimées à l'aide des facettes. Dans une étape postérieure de la BDC-Parole, nous envisagerons l'intégration d'autres facettes pour gérer les réactions (déclenchement d'actions) sur les événements (les opérations de manipulation), qui peuvent avoir lieu sur les objets directement ou sur les attributs qui les définissent.

IV. FONCTIONNALITES DE LA BDC-Parole

Les différents services offerts à l'utilisateur par la BDC-Parole sont disponibles à travers un ensemble de primitives. Ces primitives peuvent être utilisées par des applications, comme des moniteurs-parole, des interfaces graphiques, des systèmes experts, etc. Elles permettent:

- la gestion d'un espace général de recherche et de divers espaces de travail utilisateurs ainsi que le passage d'un espace à l'autre.
- la gestion des objets (CLASSES et INSTANCES): création, modification, suppression, garantissant la cohérence de la base à travers des mises à jour en cascade.
- l'interrogation sur la structure et les liens sémantiques des objets de la base
- la génération de faits visant à la création de connaissances à l'aide de requêtes de nature statistique. Les MECANISMES de CONSULTATION de la BDC-Parole, sont présentés plus en détail dans une autre communication [12] de ces mêmes journées.

Dans toute opération effectuée, la BDC-Parole prend en charge la vérification des contraintes d'intégrité, aussi bien statiques (sur les attributs des objets) que dynamiques (par la gestion de liens sémantiques lors de la mise à jour).

V. CONCLUSION ET PERSPECTIVES

La prise en compte des particularités des données et des connaissances parole nous a conduit à proposer une représentation centrée "objet" pour la Base de Données et de Connaissances Parole. Les caractéristiques du premier noyau de la BDC-Parole ont été définies. Il permet de représenter et de manipuler des structures complexes et de modéliser les traitements appliqués sur elles. Les conditions particulières de mise en oeuvre (logiciels utilisés) doivent encore être précisées.

La structure établie pour la BDC-Parole et le concept d'objet parole complexe qu'elle intègre font de cette base un support pour le développement des Bases de Connaissances Parole [11], [12].

La Communauté Européenne met en place une action de recherche sur la méthodologie d'évaluation et de standardisation des systèmes d'entrée-sortie vocale. Le GRECO Communication Parlée participe à cette action dans le cadre du projet ESPRIT et la Base de Données et de Connaissances Parole constitue l'une des composantes de ce projet dans le contexte multilangue : "Multilingual Speech Input-Output Assessment Methodology and Standardisation" (SAM Project). L'effort de réflexion et de conceptualisation mené dans le cadre de la BDC-Parole doit permettre au GRECO de jouer un rôle pilote pour l'établissement d'un standard au niveau Européen dans la représentation et la manipulation des données et des connaissances sur la parole.

V. BIBLIOGRAPHIE

[1] R. DESCOUT, J.F. SERIGNAT, O.CERVANTES, R. CARRE
BDSON : Une base de données des sons du Français
11th I.C.A., Montreal, 1986.

[2] O. CERVANTES, J.F. SERIGNAT, R.DESCOUT, R.CARRE
Définition et réalisation d'une base de données des sons du français. 15èmes JEP-GALF, 1986.

[3] O. CERVANTES, J.F. SERIGNAT, J. CAELEN
D'une Base de Données-Sons vers une Base de Données-Parole. 3èmes Journées Base de Données Avancées. 20-22 Mai 1987, INRIA-Port Camargue.

[4] M. ADIBA
Modeling Complex Objects for Multimedia Databases. 5th ER Conference, Dijon, 1986.

[5] S. ABITEBOUL, S. GRUMBACH
Bases de Données et Objets Complexes.
Proposé à TSI, 1987

[6] C. DELOBEL, M. ADIBA
Bases de Données et Systèmes Relationnels.
DUNOD - Informatique, 1982

[7] P.P. CHEN
The entity-relationship model, toward a unified view of data. ACM TODS, Vol 1, No. 1, 1976.

[8] E.F. CODD
Extending the database relational model to capture more meaning. ACM TODS, Vol 4, No. 4, 1979.

[9] R.J. BRACHMAN
What IS-A is and isn't: An Analysis of Taxonomic Links in Semantic Networks,IEEE, Comput,Oct 1983

[10] F. RECHENMANN
SHIRKA et la représentation orientée objet.
Intelligence Artificielle et Sciences Cognitives, Grenoble, Février 1987.

[11] J. CAELEN, G. CAELEN-HAUMONT, N. VIGOUROUX, C. BARRERA, J. MALET
ARCANE: Acquisition et Recherche de Connaissances Acoustico-phonétiques dans un Noyau Evolutif
15èmes JEP - GALF, Aix-en-Provence, Mai 1986.

[12] J. CAELEN, O. CERVANTES, Y.FERNANDEZ
Mécanismes de consultation de la BDC-Parole.
16èmes JEP - GALF, Hammamet, Octobre 1987.

LA BASE DE DONNEES DES SONS DU FRANCAIS (BDSONS) PERSPECTIVES DE DEVELOPPEMENT

R. CARRE (ICP GRENOBLE) et al

C.N.R.S. GRECO "COMMUNICATION PARLEE", C.R.I.N. Université de Nancy I
B.P. 239 54506 VANDOEUVRE-LES-NANCY (FRANCE)

ABSTRACT

In this paper, we recall the main characteristics of the French sound database BDSONS and some development perspectives. This database is now proposed as a reference for the European sound database which is studied in the ESPRIT SAM project. A close interaction between the French project and the European project has to be taken into account. New developments must correspond to specific needs and would act as experimental test for the European database.

This paper is the conclusion of a workshop held in GRENOBLE on the 30th of april 1987 with 25 participants from the main French laboratories.

INTRODUCTION

Ce rapport présente les conclusions d'une journée de réflexion sur la base de données des sons du français (BDSONS) du GRECO "Communication Parlée". Environ 25 personnes ont participé à cette journée qui s'est déroulée à GRENOBLE le 30 avril 1987. Les principaux laboratoires français (relevant des Sciences Physiques pour l'Ingénieur ou des Sciences Humaines) étaient représentés. Le texte qui suit se veut être un fidèle rappel des principaux commentaires formulés sur les problèmes rencontrés jusqu'alors, sur l'utilisation de la base et sur les perspectives de développement. Des descriptions détaillées des travaux effectués autour de cette base sont contenues dans les références citées en fin de texte.

Après 6 années de travail plus ou moins soutenu, la base de données entre dans une phase opérationnelle. Il s'agit de mesurer le coût humain et matériel pour un projet de grande ampleur et d'apprecier nos capacités à poursuivre efficacement un tel travail.

LES ACQUIS

Environ 32 heures d'enregistrements ont été effectués stockés sur 250 bandes magnétiques et 18 vidéocassettes soit 3,8 gigabytes. Deux grandes classes de corpus ont été définies : corpus "évaluation" et corpus "acoustique". 32 locuteurs ont participé aux enregistrements des corpus "évaluation" (dont 12 "standards", 10 avec accents régionnaux, 10 avec des caractéristiques réputées difficiles à traiter). Les 12 locuteurs "standards" ont enregistré les corpus "acoustiques". Le choix des 12 locuteurs "standards" a fait l'objet d'une réflexion particulière.

Pour plus de détails sur les choix des corpus "évaluation", on pourra se repporter à la référence MARIANI (1987).

Les corpus "acoustiques" ont été limités à des ensembles CVCV pour toutes les consonnes mais avec seulement les voyelles cardinales /a/, /i/ et /u/. Avec quelques compléments (groupes consonantiques, paires et triplets, mots réels...), on arrive déjà à environ 14 heures d'enregistrement.

Les enregistrements ont été particulièrement contrôlés avec une méthodologie spécifique (présentation à l'écran des directives...). Le studio a été étudié au CNET à Lannion où les premiers enregistrements ont été effectués. Pour plus d'informations, on consultera, en particulier, la référence (DESCOUT et al., 1986 :...).

Un premier travail sur l'étiquetage fin à partir de données temporelles ou spectrales a permis de trouver un consensus (ABRY et al., 1986). Selon la méthodologie retenue, 2000 heures de travail de chercheurs sont financées en 1987 pour étiqueter environ 3 heures de signal soit 1/10 de la base. Un étiquetage large adapté aux besoins de la reconnaissance de parole est en cours d'étude.

Enfin les moyens matériels et logiciels nécessaires au bon fonctionnement de la base ont été développés : postes de travail, logiciels... En particulier, un gros travail a été effectué sur la gestion informatique de la base pour permettre tout type d'interrogation de cette base (selon les locuteurs, caractéristiques des locuteurs, selon les corpus, les constituants de chacun des éléments du corpus...), voir CERVANTES et al. (1986).

Pour faciliter la diffusion de la base, un disque compact est en cours de pressage, contenant une grande partie des corpus "évaluation" (mémoire 0,5 gigabyte). Il sera distribué dans 10 lieux différents pour évaluation.

COUT DE LA BASE

Le développement de la BDSONS s'est concrétisé grâce à un effort financier important des participants. Au 1er janvier 1987, le coût de la BD s'élevait à la somme de 3850 KFrancs se décomposant de la façon suivante :

Equipement.....2800
Missions.....150
Vacations.....200
Salaires.....1200

avec les participants suivants :

GRECO.....1800
CNET.....1200
LCP.....850

Dans ces sommes, on ne compte pas le temps consacré par les membres des groupes de travail. On a retenu la moitié de la somme consacré par le CNET à la construction de son studio et la totalité des sommes engagées par le GRECO pour équiper les laboratoires en Bétamax et convertisseurs PCM (1000 kFrancs).

LE PROJET EUROPEEN ESPRIT "SAM" ET LA BDSONS

Une description détaillée de ce projet est contenue dans la référence (DULMAZIN et al., 1987) citée en annexe. Il s'agit de proposer des méthodes d'évaluation des systèmes de synthèse et de reconnaissance de parole. Naturellement, les bases de données jouent un rôle important comme support de l'évaluation.

Durant la phase de définition qui est en cours, le GRECO doit proposer les caractéristiques d'une station de travail standard. Il participe, par ailleurs, avec les Danois, à une étude préalable sur les bases de données des sons. Les méthodes d'étiquetage de cette base sont étudiées sous la responsabilité des britaniques.

A l'exemple français, un disque compact européen est en cours de pressage avec 5 langues différentes.

DISSEMINATION

Le premier objectif d'une base de données est évidemment d'être utilisée par le maximum de personnes et de laboratoires. Celà implique, d'une part, de disposer de données intéressantes, adaptées, faciles à exploiter et, d'autre part, de pouvoir mettre à disposition rapidement et à un coût réduit les informations. Alors la base de données pourra jouer un rôle normatif. Pour répondre à ces préoccupations, les supports de la base de données sont étudiés pour pouvoir être diffusés facilement (supports vidéocassette et maintenant essai de support CDROM). Par ailleurs, le prix de vente est réduit aux frais de copie. La diffusion n'est pas limitée et fait l'objet d'un contrat où le demandeur s'engage à citer le nom du GRECO dans ses rapports ou publications. Le demandeur s'engage aussi à ne pas diffuser lui-même la base de données et à mettre à disposition, à titre de réciprocité, ses propres corpus aux membres du GRECO. Ces règles s'appliquent aussi bien aux étrangers qu'aux industriels.

PERSPECTIVES

Il s'agit, en tenant compte du travail effectué dans le projet SAM, d'apporter une complémentarité et surtout de défricher un terrain pouvant être repris, ensuite, au niveau européen. Le projet SAM permet de donner toute l'ampleur nécessaire en assurant une véritable existence aux bases de données parole. La BDSONS est aujourd'hui unique en Europe par sa taille et par le consensus réalisé au plan français. On peut supposer que l'expérience et la méthodologie acquises seront prises en compte dans le développement des bases de données européennes. Il s'agit aujourd'hui, de conserver cette avance en utilisant la souplesse du GRECO pour tester de nouveaux développements lesquels, étant donné la relative ampleur des expériences françaises, pourront, à nouveau, être proposées aux européens.

La taille et la complexité de la base nous conseillent vivement le réalisme c'est-à-dire que toute nouvelle extension doit s'appuyer sur une demande réelle des laboratoires.

Le concept de centre serveur a fait son apparition lors de la réunion de travail. Il montre bien l'importance du projet et le type d'organisation qu'il faut mettre en oeuvre.

Un nouveau groupe de travail représentatif de la communauté participera à la définition des extensions de la BDSONS.

Exploitations de la BDSONS.

Rappelons tout d'abord l'étiquetage fin effectué actuellement sur quelques corpus.

Parmi d'autres exploitations de la base, citons celles qui visent à l'apprentissage symbolique automatique. Il s'agit de définir de nouveaux objets à partir d'objets existants.

En exploitation phonétique, à partir de l'étiquetage manuel, il s'agit de tester des méthodes de segmentation automatique, des modèles de rupture..., d'effectuer des études statistiques.

De nouveaux corpus, de nouveaux locuteurs.

La première urgence est d'accroître le nombre de locuteurs. Il s'agit de passer de 32 à une centaine de locuteurs (ou plus), ceci principalement pour les besoins d'évaluation. Cette extension concerne le français dit "normatif". L'utilisation de locuteurs enfants (une cinquantaine) a été évoquée, les applications pour des jeux se développant. Une partie seulement des corpus pourrait être distribuée (les 2/3 par exemple), le tiers restant étant conservé pour des besoins d'évaluation dans un lieu restant à définir.

Pour les études acoustiques, l'extension du corpus CVCV à toutes les voyelles a été proposée.

Enfin, et à la demande, une réflexion doit être conduite pour créer un corpus "grand vocabulaire", pour accroître le nombre de situations d'environnement : bruits, vibrations..., pour enregistrer un même locuteur dans des situations variées..., pour étendre le nombre de locuteurs avec des variantes régionales...

Un corpus représentatif permettant l'étude des dialogues oraux doit aussi faire l'objet d'études.

Naturellement, de nouveaux lieux d'enregistrement devront être retenus pour permettre ces extensions. Ils auront le label "GRECO" et leurs productions seront estampillées avant intégration dans la base. Un nouveau studio est maintenant disponible à GRENOBLE.

Une base de données physiologiques?

Une base de données qui contiendrait des informations sur la forme du conduit vocal (rayons X, scanner...) avait déjà fait l'objet de réflexions. Alain MARCHAL (IP AIX) reprend aujourd'hui ce projet et, après une visite des différents laboratoires, doit rédiger des propositions, lesquelles seront étudiées par un groupe ad hoc et soumises aux décisions du GRECO. D'ores et déjà, on reconnait que des paramètres physiologiques, très utiles pour les recherches acoustiques, sont absents : par exemple, des données glottographiques.

Vers une Base de données Parole.

Il s'agit d'associer aux données actuellement gérées par la BDSONS des informations lexicales, syntaxiques, sémantiques, prosodiques, phonétiques, phonologiques.... La notion d'objet-parole conduit à une structure de base beaucoup plus complexe de type orienté objet. La BDSONS doit donc évoluer vers une BD connaissances parole (CERVANTES et al. 1987).

La liaison avec la Base de Données Lexicales du GRECO "Communication Parlée".

Un séminaire est prévu à l'automne à TOULOUSE qui étudiera, en particulier, ce problème.

PUBLICATIONS CONCERNANT LA BDSONS

R.CARRE, R. DESCOUT, M. ESKENAZI, J. MARIANI, M. ROSSI (1984)
The French Language Database : Defining, Planning, and Recording a large Database.
IEEE ICASSP, 42.11, SAN DIEGO.

T. BOMMART, R. DESCOUT (1985)
Etat d'avancement de la BDSONS.
Réunion GRECO/Industries, GIF sur YVETTE.

C. ABRY, D. AUTESSERRE, C.BARRERA, C.BENOIT, L.J. BOE, J. CAELEN, G.CAELEN HAUMONT, M.ROSSI, R.SOCK, N. VIGOUROUX (1985)
Propositions pour la Segmentation et l'Etiquetage d'une Base de Données des Sons du Français.
14 èmes Journées d'Etude sur la Parole, GALF, PARIS.

O. CERVANTES, J.F. SERIGNAT (1985)
Définition et réalistion d'une base de données informatique des sons du français.
Rapport interne.

R. DESCOUT, J.F. SERIGNAT, O. CERVANTES, R. CARRE (1986)
Une base de données des sons du français.
11ème Congrès International d'Acoustique, MONTREAL.

R. CARRE (1986)
Les Industries de la Langue : Enjeux pour l'Europe ; les bases de données.
Colloque International sur les Industries de la Langue, TOURS.

BDSONS. Base de Données des sons du Français. Description du contenu des corpus de type I : Evaluation
II : Acoustique.
Rapport GRECO, LCP GRENOBLE (1986).

J. CAELEN, O. CERVANTES, J.F. SERIGNAT (1987)
Objets complexes dans BD Parole, Connaissances centrées objet dans ARCANE.
Symposium Intelligence Artificielle et Sciences Cognitives, GRENOBLE.

O. CERVANTES, J.F. SERIGNAT, J. CAELEN (1987)
Evolution de la BDSON vers la BD Parole.
Symposium 3èmes journées Bases de Données Avancées, INRIA, PORT CAMARGUE.

O. CERVANTES, J.F. SERIGNAT (1987)
Représentation centrée objet dans la base de données et de connaissances parole.
16èmes Journées d'Etude sur la Parole, GALF, HAMMAMET.

J. CAELEN, O.CERVANTES, Y.FERNANDEZ (1987)
Mécanismes de consultation dans la Base de Données et de Connaissances Parole (BDCParole).
16èmes Journées d'Etude sur la Parole, GALF, HAMMAMET.

J.M. DOLMAZON, C. BENOIT, J.L. GAUVAIN, G. PERENNOU (1987)
Le Projet Européen "SAM" : Evaluation multilingue des Dispositifs d'Entrée-sortie vocale.
16èmes Journées d'Etude sur la Parole, GALF, HAMMAMET.

J. MARIANI (1987)
Choix des corpus d'évaluation.
Rapport interne GRECO-LIMSI, présenté à la journée BDSONS du 16 mai 1987, GRENOBLE.

LE TRAITEMENT MORPHOPHONOLOGIQUE DANS BDLEX_1

G. Pérennou et M. de Calmès

Laboratoire C.E.R.F.I.A., université P. Sabatier,
118, route de Narbonne, 31062 TOULOUSE Cedex, FRANCE

ABSTRACT

This paper presents the BDLEX project (LEXical Data and Base) developed within the context of the GRECO_CNRS on Spoken Communication. The project is centered upon the phonological and morphosyntactical levels of written and spoken French and is intented for use in applications involving the automatic processing of speech and texts.

INTRODUCTION

Les industries de la langue et plus particulièrement la communication homme machine en langage naturel exigent des lexiques électroniques de plus en plus complexes dont le contenu et l'organisation doit s'adapter aux applications visées. L'idée de constituer des bases de données lexicales destinées à déservir toute une classe d'utilisateurs en lexiques particuliers s'est ainsi développée.

BDLEX (Base de Donnée LEXicale), que nous développons dans le cadre du GRECO-CNRS de la Communication Parlée, est centrée sur les niveaux phonologiques et morpho-syntaxiques du français écrit et parlé.

D'autres projets de recherches visant à créer des lexiques formalisés existent pour le français avec des but différents, le plus voisin étant celui de M. Gross [1] se situant dans une perspective de lexique-grammaire. Il faut également mentionner le lexique développé au CNRS-HESO (Gruaz [2]).

La première version de notre base BDLEX-0 est achevée depuis 1986 et la seconde BDLEX-1 s'achève en 1987 [3]. Différents sous-lexiques sont disponibles sur disquettes PC.

La composante phonologique GEPH se présente comme un système expert appliquant les règles des experts phonologues aux énoncés soumis en entrée.

Nous avons en vue diverses applications telles que les suivantes:
- machines à reconnaître la parole dictée[4],
- synthèse de la parole,
- vérification et correction automatique de fautes orthographiques et typographiques,
- enseignement assisté par ordinateur (EAO).

Dans cette communication nous indiquerons la structure générale de BDLEX puis nous examinerons les niveaux phonologiques et morphologiques et la traduction dans une base de données des diverses relations morphologiques.

STRUCTURE DE BDLEX

L'organisation de BDLEX en tant que base de données (abrév. BD) est du type relationnel. Actuellement le système de gestion de BD utilisé est M.R.D.S. (Multics Relational Data Store) sous DPS_8 Multics du Centre Interuniversitaire de Calcul de Toulouse. En réponse à une question, une nouvelle relation est générée. Elle peut être vue comme un sous-lexique particulier - voir annexe.

Il est possible d'interroger BDLEX à travers le réseau Transpac ou le réseau téléphonique normal, ce qui facilite la collaboration entre équipes.

Le diagramme de la figure 1 indique les différentes parties prévues à terme dans BDLEX.

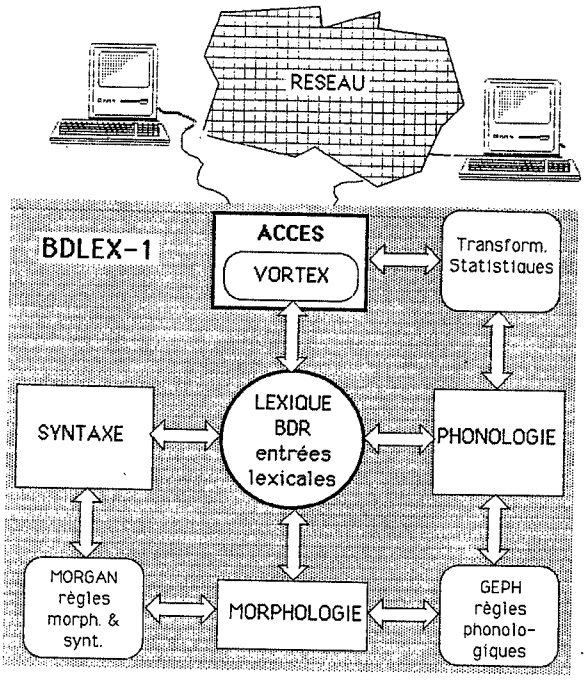


Fig.1 Composantes de BDLEX.

BDLEX-0 comporte:
- 7 000 entrées lexicales et 150 000 mots fléchis,
- une composante morphologique flexionnelle,
- une composante morphosyntaxique,
- une composante phonologique,
- un module d'interrogation et de mise à jour.

BDLEX-1 comporte 25 000 entrées lexicales qui génèrent 350 000 mots fléchis et possède de plus une composante morphologique dérivationnelle.

L'accès peut se faire à travers le sytème VORTEX; il est alors tolérant aux fautes orthographiques et typographiques.

COMPOSANTE PHONOLOGIQUE

A chaque entrée lexicale est associée une représentation phonologique sous jacente. Le système expert GEPH à pour rôle de leur associer les représentations phonétiques qui déterminent leur prononciation dans un contexte donné.

Représentation phonologique sous-jacente

Quelle forme sous-jacente associer aux entrées lexicales? La réponse à cette question est souvent délicate, comme suggéré plus loin, et de toute manière dépendante des traitements phonologiques prévus, que ce soit avant ou après l'insertion dans la phrase. Dans le projet BDLEX, les divers fragments de composantes phonologiques - consonnes terminales, nasalité et semi-vocalisation (Dell & Plénat [5], Dell [6]), voyelles à double timbre (Lambert & Ross [7]) - fournis par un groupe de phonologues sont à la base de la détermination des formes sous-jacentes.

Certaines entrées lexicales sont sujettes à des variations morphophonologiques. Par exemple, en français l'adjectif "neuf" devient "neuve" au féminin. Deux solutions peuvent être envisagées dans de tels cas:

- l'entrée est polymorphique, chaque morphe étant soumis à une condition de sélection (ex. pour l'adjectif "neuf": /nœf/_ masc; /nœv/_ fém),
- la composante phonologique comporte les règles permettant de rendre compte des variations (ex. la forme sous jacente unique de "neuf" est /nœv/ et la composante phonologique comporte la règle de dévoisement des fricatives finales, règle qui ne peut s'appliquer en présence d'un suffixe ou d'une flexion commençant par une voyelle ou une semi-voyelle:

(DEVOI2) [-son +cont] ---> [-voix] / __# ou __+[+cons]

Dans BDLEX la deuxième solution a été adoptée lorsque cela ne supposait pas l'introduction de règles phonologiques ad hoc. Ainsi, la forme sous-jacente est unique lorsque les formes de surface peuvent en dériver par application des règles suivantes:

a) Consonne latente et consonne mixte

Une consonne latente, notée C", se réalise comme la consonne C dans un contexte non-consonnantique postérieur; elle est tronquée dans les autres cas. Par exemple la consonne latente /t"/ de /pətit"/ («petit») permet de rendre compte des deux formes différentes du masculin et du féminin: /pəti/ («petit»); /pətitə/ («petite»), des dérivations comme /pətitɛsə / («petitesse») et des cas de liaison (resp. non-liaison) comme «petit_âne» (resp. «petit pois»).

Une consonne semi fixe C' se réalise comme une consonne fixe C en finale de groupe (c.-à-d. en contexte __##). Dans les autres cas elle se comporte comme la consonne potentielle C". Elle permet de donner une seule représentation sous-jacente à des mots comme /sis'/ («six») ou /ɔs'/ («os») etc...

Les consonnes latentes et mixtes sont régies par l'ensemble ordonné des règles suivantes:

(FIX1) C" ---> C / __ +[-cons]
(FIX2) C' ---> C / __ ##
(DEFIX) C' ---> C"
(LIAI) C" ---> C / __ #[-cons]
(TRONC) C" ---> Λ .

Cet ensemble doit être complété par des règles dévoisant les occlusives et voisant les sifflantes quand elles font liaison. Ainsi «doux» et «grand» pourront avoir les représentations sous-jacentes respectives /dus"/ et /grãd"/ nécessaires pour former «douce» et «grande» alors que les liaisons sont en /z/ et /t/.

b) La dénasalisation régulière des finales lexicales

En français, lorsque (TRONC) prend effet sur une consonne nasale latente placée derrière une voyelle orale, celle-ci est nasalisée. Les mots comme «moyen» peuvent ainsi être représentés par une seule forme sous-jacente /mwajɛn"/ malgré l'alternance [mwajɛ̃ / mwajɛn] correspondant à «moyen/ moyenne». La règle (NAS) suivante, ordonnée entre (LIAI) et (TRONC) rend compte de cette nasalisation

(NAS) V ---> [+nas] / __ N" ,
avec: {i,ɛ} [+nas] ---> ɛ̃,
{y,œ} [+nas] ---> œ̃,
ɔ [+nas] ---> õ,
a [+nas] ---> ã

Toutes les finales nasales ne fonctionnent pas comme précédemment. C'est très clairement le cas pour les consonnes nasales fixes, comme dans «rhum», dont la représentation sous-jacente ne pose pas de problème _ tout simplement /rɔm/ dans notre exemple.

Il existe un cas plus délicat concernant une liste, à vrai dire réduite, de mots outils fréquents comme "selon", "on", "rien", "un", "mon", etc..., caractérisé essentiellement par le fait que c'est la voyelle nasale qui est prononcée en toute position.

Pour ceux de ces mots qui ne sont pas soumis à des variations morphologiques, la solution la plus simple est de prendre la voyelle nasale en forme sous-jacente: /səlõ/ pour le premier qui ne fait pas la liaison, /õn"/, /rjɛ̃n"/ pour les deux mots suivants.

Pour les autres, une possibilité est de prévoir deux allomorphes: une forme masculine et une forme féminine, ainsi: /œ̃n"/ et /yn / pour «un» et «une», /mõn"/ et /ma/ pour «mon» et «ma» (le lexique devant de plus spécifier la supplétion de «ma» par «mon» dans le contexte __#[-cons]).

c) Alternance /ə;ɛ/

Elle se produit pour des verbes tels que "congeler"; une seule racine phonologique /kõʒəl/ suffit, malgré l'alternance [kõʒəle / kõʒɛlə] de "congelé/congèle", moyennant la règle (EC) suivante:

(EC) ə ---> ɛ / en syllabe fermée.

d) Semivocalisation

Dans les cas où des variations entre semi-voyelle et voyelle fermée apparaissent dans les diverses formes d'un même mot, c'est la voyelle fermée qui figure en forme sous-jacente. Un ensemble complexe de règles permet de prévoir les diverses réalisations de surface. Par exemple les racines sous-jacentes de «plier» et «rouer» sont /pli/ et /ru/. Elles donnent lieu aux formes suivantes: [pliə] («plie»), [plie] ou [plije] («plié»), [rue] ou [rwe] (roué). Pour «hier» la forme sous-jacente sera /iɛr/ d'où dérive aussi bien [jɛr] que [ijɛr].

Le choix des formes sous-jacentes du français soulève entre d'autres difficultés que nous ne pouvons examiner ici, comme: l'alternance de timbre des voyelles moyennes de [ɑ/a] et de [ɛ̃/œ̃], la prononciation, les mots d'emprunts...

Indiquons que dans une variante actuelle de BDLEX1, destinée au TAP, la représentation phonologique sous-jacente n'oppose pas les voyelles à double timbre.

Le système expert GEPH

Muni des règles phonologiques placées dans sa base de connaissance, le système expert GEPH (de GEnérateur PHonologique) traite les énoncés en forme sous-jacente pour déterminer les représentations phonétiques de surface des variantes de prononciation - voir Fig.2 -.

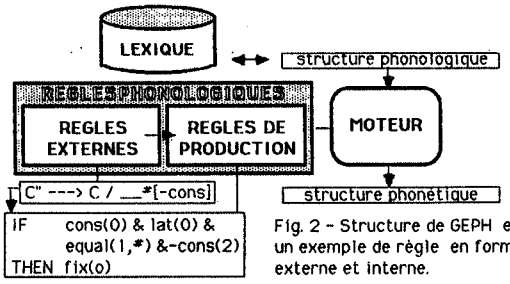


Fig. 2 - Structure de GEPH et un exemple de règle en forme externe et interne.

Les règles peuvent être introduites dans le formalisme des phonologues. Elles sont ensuite traduites en forme interne - pour un exemple, voir Fig.2 -. Ci-dessous on trouvera un exemple de traitement par GEPH en mode génératif.

```
quelle phrase ?   :Arbre;mAlAde:
                  ---------------
regle applicable numero:     48 nom: e_AB
phoneme:  A
-------------------------------------------

regle applicable numero:     26 nom: E_FINb
phoneme:  e
-------------------------------------------

regle applicable numero:     27 nom: EPEN
phoneme:  r                                  Formes phonetiques
-------------------------------------------  ------------------
                                             solutions:
regle applicable numero:     35 nom: LIQUEF_a $$$$$$$$$$$$$$$$$$$
phoneme:  r                                  $                   $
-------------------------------------------  $  **  ArbmAlAd    $
                                             $  **  ArbremAlAd  $
regle applicable numero:     37 nom: NAS_REG $  **  ArmmAlAd    $
phoneme:  b                                  $                   $
-------------------------------------------  $$$$$$$$$$$$$$$$$$$

regle applicable numero:     26 nom: E_FINb
phoneme:  e
-------------------------------------------
```

Inversement, partant d'un énoncé phonétique, il permet de trouver la forme sous-jacente.

Les règles phonologiques ont été fournies par un groupe de phonologues (Dell et Plénat[5] [6] [8]; Lambert et Rossi[7]). Elles doivent être complétées par des règles visant à la description plus fine de la prononciation. C'est à ce niveau que doit s'opérer la jonction avec la base de données BDSONS des sons du français développée également dans le cadre du GRECO_CNRS de la communication parlée[9].

LA COMPOSANTE MORPHOSYNTAXIQUE

Morphologie flexionnelle

Les alternances morphonologiques auxquelles donnent lieu les flexions ne justifient pas en général le traitement polymorphique. Ainsi, l'entrée lexicale «vif» («vive» au féminin) est sujette à une règle d'alternance morphologique, notée ici $[\ ]_{v\text{ə}}$, que doit spécifier le lexique. La représentation sous-jacente unique /vif/ sera valable pour le masculin $[vif + masc]_{v\text{ə}}$ («vif» [vif]) comme pour le féminin $[vif + fém]_{v\text{ə}}$ («vive» [vivə]) si la règle $[\ ]_{v\text{ə}}$ appliquée à une racine de la forme Xf est définie comme suit:

$[Xf + masc\ ]_{v}$ ---> [Xf]
$[Xf + fém\ ]_{v}$ ---> [Xvə]

Par ailleurs, une telle entrée est sujette à la règle de formation du pluriel régulier, notée $[\ ]_{z"}$, suivante:

$[X + sing\ ]_{z"}$ ---> [X]
$[X + plur\ ]_{z"}$ ---> [X z"]

Ainsi /vifz"/ («vifs») est formé à partir de $[[vif + masc\ ]_{v\text{ə}} + plur\ ]_{z"}$.

D'une manière générale les noms, adjectifs et verbes se traitent de cette manière dans BDLEX. Dans des cas exceptionnels comme pour le nom "empereur" dont le féminin est "impératrice" (les parties en alternance sont soulignées), deux entrées séparées mais chaînées l'une à l'autre ont été créées -- on notera que ce choix peut aussi se justifier au plan sémantique puique "une impératrice" a deux acceptions:"femme d'empereur" ou "occupant le trône impérial".

Morphosyntaxe

La composante morphosyntaxique permet l'analyse (non déterministe) et la génération des formes fléchies. Le caractère éventuellement défectif en temps et/ou personnes, lorsqu'il existe, est pris en compte.

Les traitements se font graphiquement et/ou phonétiquement.
Exemples:
• Pour l'entrée «vif», le système génère au mode graphique:
masc. sing. : vif /
masc. plur. : vifs /
fém. sing. : vive /
fém. plur. : vives /
• Au mode phonétique on obtiendra:
masc. sing. : vif /
masc. plur. : vifz" /
fém. sing. : viv /
fém. plur. : vivəz" /
• Le verbe «faire» au passé composé sera conjugué au mode graphique:
je suis fait ou j'ai fait
tu es fait tu as fait
etc... etc...
•L'analyse de «cours» en mode graphique fournit trois groupes de solutions:
1) «cour + s» nom fém. au plur.
2) «cours» nom masc. au sing. ou plur.
•Si [moz] est soumis à l'analyse phonétique le système proposera deux solutions qui sont: «mot+s» (nom masc. plur.) et «m+aux» (nom masc.plur.).

Lorsque l'on interroge la base pour effectuer des conjugaisons ou pour générer les formes fléchies des noms et adjectifs il est nécessaire de faire appel aux règles morphophonologiques. Les alternances qui ne dépendent ni du contexte dans la phrase, ni du locuteur, sont traitées à ce niveau.

Morphologie dérivationnelle

Les dérivations peuvent s'interpréter comme des relations. Ainsi «inégal» et «inégalement» se décrivent par des tuples dans les relations
PREFIX(préf,entrée-1,entrée-2) et
SUFFIX(entrée-1,suf,entrée-2),
à savoir:
$(IN_{11}$, égal,inégal)
(inégal, $MENT_{11}$, inégalement)

Dans ces tuples on doit tenir compte du fait que des affixes peuvent être homophones et donner lieu à des variations morphophonologiques. Ainsi «in-» recouvre en français deux préfixes. IN_{11} est le préfixe négatif d'adjectif dans sa variante «in-» (il peut aussi se réaliser «il-» devant un «l», «ir-» etc.), $MENT_{11}$ est le suffixe formant des adverbes sur des adjectifs dans sa variante «-ement».

On peut penser que "entrée-2" est redondant dans ce genre de formulation, et c'est ce qui paraît ressortir de l'exemple précédent.

D'autres cas plus complexes montrent que la simplification envisagée n'est pas possible en général sans que soit prévu un dispositif de traitement d'irrégularités ou d'allomorphies.

Par exemple, «éli» étant la racine du verbe «élire» débarassée de la désinence «re» de l'infinitif et semblablement «croi» la racine du verbe «croire», les couples (éli,IBLE) et (croi,IBLE) ne permettent de prédire totalement les formes «éligible» et «crédible». Les deux triplets suivants ne sont donc pas redondants:

(éli, IBLE, éligible)
(croi, IBLE, crédible).

Les opérations classiques des systèmes de gestion de base de données permettent de retrouver les dérivations complexes - voir Fig.3.

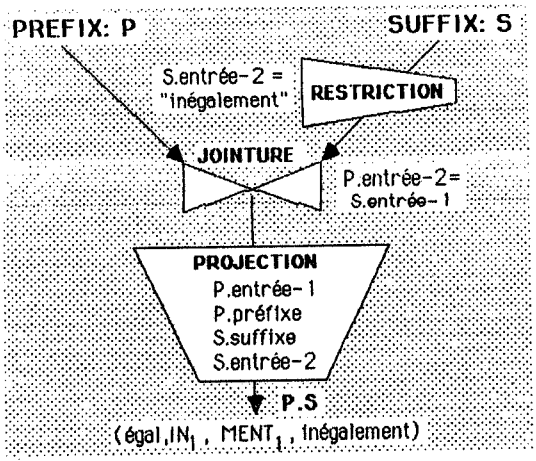


Fig.3 Un exemple de requête sur une dérivation complexe

Dans BDLEX nous avons choisi un premier stade de représentation morphologique à finalité descriptive. Chaque entrée lexicale est munie de marques de frontière spécifiant la nature du terme placé immédiatement après. Lorsque une partie du mot est une entrée, celle-ci n'est pas décomposée. Les substitutions (notées >), ajouts (notés +) ou suppression (notée -) de lettres qui sont nécessaires pour obtenir l'identificateur de cette entrée sont placés entre parenthèse. Ceci est illustré par les exemples qui suivent où:

```
:X renvoie à l'entrée autonome X,
=X renvoie à l'entrée non-autonome X,
.s désigne s comme suffixe,
+p désigne p comme préfixe,
;d désigne d comme désinence.

    égal      +lin:égal              :inégal.lement
    éli;re    :élig(-1)(+re).ible
    croi;re   :créd(>2oi)(+re).ible
    +bi=game
```

Lexique général– Lettre F –

GRAPH_ACC	HG	PHON_SYLL
...		
feindre	11	fæ̃/dr
feinte	11	fæ̃t
feld-maréchal	11	feld/ma/re/šal
fêler	11	fe/le
félibre	11	fe/li/br
félicitations	11	fe/li/si/ta/sjõ
félicité	11	fe/li/si/te
féliciter	11	fe/li/si/te
félin	11	fe/li
fellaga	11	fe/la/ga
fellagha	11	fe/la/ga
fellah	11	fe/la
fellation	11	fe/la/sjõ
félon	11	fe/lo
...		

Références

[1] Gross, M., "Méthodes en syntaxe," Paris:Hermann, 1975.
[2] Gruaz, C., "La dérivation suffixale en français contemporain dans les familles monosyllabes de haute fréquence," Thèse de doctorat d'état, université de Paris, 1984.
[3] Pérennou, G., de Calmès, M., "BDLEX Lexical Data and Knowledge Base of Spoken and Written French," European Conference on Speech Technology, Edinburgh, 1987.
[4] Pérennou, G., "Lexique et phonologie - Parole et texte," monographie AFCET, (à paraître).
[5] Dell, F. & Plénat, M., "Semi-voyelles et consonnes finales en français," rapport interne du GRECO communication parlée du CNRS,1985.
[6] Dell, F., "Les règles et les sons," Paris:Hermann, réédition 1986.
[7] Rossi, M. & Lambert, M., "Représentation et traitement des voyelles à timbre multiple," Actes du séminaire GRECO-GALF, 1986.
[8] Plénat, M., "La loi de Littré," Cahier de grammaire 2, 45-135, 1980.
[9] Descout, R., Sérignat, J.F., Cervantès, O. & Carré, R., "BDSONS: a Data Base of Sounds of the French Language," 11e ICA, 1986.

ANNEXE: Sous-lexique extrait de BDLEX_1

GRAPH_ACC graphie accentuée de l'entéee lexicale
HG numéro homographe (nombre, numéro d'ordre)
PHON_SYLL représentation phonologique sous-jacente syllabéee
FPH fonctionnement phonologique de la finale
HP numéro homophone (nombre, numéro d'ordre)
VA indicateur variante (G: graphique, V: graphique et phonétique,...)
CL_PHON représentation phonologique en classes majeures
NS nombre de syllabes
FREQ fréquences
CS catégorie syntaxique

FPH	HP	VA	CL_PHON	NS	FREQ	CS
e	11		SVn/DR	2	B0	V
e	11		SVnT	1		N
	11		SELD/NA/RE/SAL	4		N
r"	11		SE/LE	2		V
e	11		SE/LI/DR	3		N
z"	11		SE/LI/SI/TA/SJVn	5	B0	N
	21		SE/LI/SI/TE	4		N
r"	22		SE/LI/SI/TE	4	B0 TR	V
n"	11		SE/LI	2		J
	21	G	SE/LA/DA	3		N
	22	G	SE/LA/DA	3		N
	11		SE/LA	2		N
	11		SE/LA/SJVn	3		N
n"	11		SE/LE	2		J

LANGAGE DE MANIPULATION DES INFORMATIONS ACOUSTICO-PHONETIQUES

Nadine Vigouroux

Laboratoire CERFIA UA-CNRS N°824, Université Paul Sabatier - TOULOUSE

ABSTRACT

This is an introduction to an Acoustic Phonetic Data-Base (APDB) that is constructed on an object-oriented model of representation. This model allows both to express and to handle acoustic and phonetic objects that are either multiple or complex. Any approach to vocal message characteristics --based on an observation of acoustic phenomena-- is bound to be complex. This observation phase of acoustic data cannot mean much, unless both a fairly important sampling of data, and a language that is adapted to representation of this data are available.

In order to meet these needs, it is necessary to define an acoustic analysis system that adequately corresponds to the criteria involved. In the present article, a presentation is made of the object formalism that is used to design the base. This is followed by a description of how this formalism is set to work both in managing acoustic phonetic data and in consulting this data.

INTRODUCTION

La parole possède une structure complexe qui comporte plusieurs centaines de types de réalisations acoustiques et différents contextes phonologiques, si l'on ne considère que l'étape acoustico-phonétique. Dans l'état actuel de la connaissance, on suppose que c'est l'expression du message parlé (forme phonologique), essentiellement sa projection sur les plans de la substance (contenu acoustique et articulatoire) et son articulation dans l'axe des temps, qui déterminent ses propriétés et ses fonctions acoustiques [1]. L'obtention de connaissances détaillées sur la forme et sur les structures complexes de la réalisation acoustique vis-à-vis des niveaux supérieurs (phonétique, phonologique, lexical et même prosodique, ...) est l'objectif de nombreux chercheurs dans le domaine.

Dans ce contexte, nous proposons de définir un ensemble d'outils d'aide à la détermination de la structure et du comportement de la réalisation acoustico-phonétique de l'onde vocale. En France [2, 3, 4], aux Etats-Unis [5], on s'est orienté de nouveau vers des études systématiques de la structure du message parlé par l'élaboration et l'exploitation des bases de données et de connaissances multi-médias. L'étude des caractéristiques du message vocal est complexe et repose sur l'observation des phénomènes acoustiques. Cette fonction d'observation des données acoustiques nécessite pour être significative un échantillonnage important de données et un langage adapté à leur représentation et à leur manipulation. L'environnement que nous présentons doit permettre l'accès à un certain nombre de données sur la parole telles que les représentations spectrales, les descriptions acoustico-phonétiques et articulatoires [6], etc, ainsi que l'accès à des résultats de multiples techniques d'analyses (experte, reconnaissance, analyses statistiques, ...). L'objectif est ici de décrire le système de consultation de la base de données acoustico-phonétiques (BDAP). Cela revient à définir un langage adapté à la fois à la représentation et à la manipulation d'entités acoustico-phonétiques.

LE MODELE DE REPRESENTATION

La description des informations acoustiques et phonétiques

La BDAP est munie d'un ensemble de fonctions qui permettent à l'utilisateur d'analyser le signal vocal, de l'étiqueter selon un modèle cognitif, de le visualiser, de gérer les résultats de ces tâches ... Le système de BDAP voit les informations produites comme une collection d'attributs qui caractérise une entité de parole au niveau acoustico-phonétique. Les attributs de ces entités sont numériques ou symboliques. Certains sont uni-dimensionnels d'autres multi-dimensionnels (vecteur). Les informations acoustiques et phonétiques que nous considérons sont limitées aux informations suivantes, à savoir :

- les données quantitatives [7] résultant de l'application de modules d'analyse acoustique tels que bancs de filtre, Fast Fourier Transform et de fonctions de calculs d'indices spectraux, formantiques, prosodiques, ... ,
- les données qualitatives qui représentent l'identification des événements conceptuels linguistiques [8] décelés par le phonéticien lors des sessions d'étiquetage.

Cette succincte description au niveau du décodage acoustico-phonétique (DAP) nous montre déjà que les informations utilisées sont de natures diverses (acoustique, phonétique, articulatoires) et de types (structures) différents (numérique et conceptuelle).

Le problème est ici de définir un modèle de données adapté à la représentation et la manipulation de ces entités dans une base de données ainsi qu'un environnement convivial pour la consultation.

La modélisation

Des recherches menées conjointement dans les domaines des bases de données et des bases de connaissances nous proposent des représentations de type structurée (Script, Frame, Objet) qui permettent de prendre dans une même structure les diversités de type et l'aspect sémantique des données. Nous avons opté pour une modélisation objet, car ce formalisme permet de définir la sémantique des objets concernés dans une application de base de données. Cette sémantique correspond aux propriétés de structure et de comportement des objets utilisés. Notre approche est donc de concevoir le modèle de données comme une collection d'objets structurés munis de liens d'ordonnancement et de liens d'équivalence (canonique). De ce fait, la base de données acoustiques et phonétiques se compose :

- d'un schéma conceptuel qui représente l'ensemble des contraintes et des règles de structuration des informations,
- de la base de description contenant l'image des objets modélisés.

Le schéma conceptuel

Pour faciliter la maintenance de la base, nous utilisons un formalisme externe qui permet d'exprimer les entités acoustico-phonétiques de la manière suivante :

- les descriptions permettent de décrire en termes d'attributs chaque objet ainsi que leurs domaines de définition;

Exemple : **Description** : Phonème (valeur). La valeur d'un phonème appartient à l'ensemble des phonèmes d'une langue.

Description : Echantillon : (durée, nom-fichier, adresse).

- les définitions structurent les objets en arbres et décrivent comment produire les entités "fils" à partir d'une entité père (en spécifiant à chaque noeud les fonctions à appliquer). A partir de ce formalisme externe proche des usagers et de l'application, il est facile de passer à une conceptualisation en objets. Aux concepts d'entités et d'actions, la conceptualisation en type d'objet des données, impose cependant, d'ajouter les mécanismes d'instanciation et d'héritage permettant de créer

puis d'ajouter entre eux les différents objets du système (Cf. Figure 1).

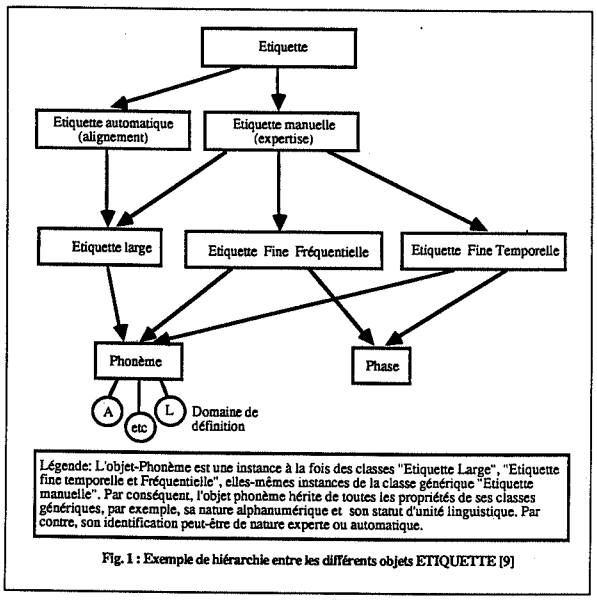


Fig. 1 : Exemple de hiérarchie entre les différents objets ETIQUETTE [9]

La définition d'un objet

Un objet est défini [10] par :
- son identification (nom, nom-classe),
- l'ensemble des attributs qui vont lui conférer un état,
- l'ensemble des fonctions que l'on peut appliquer sur lui.

Plus formellement, nous définissons les objets avec la syntaxe suivante :
OBJET=(Nom-Objet, Nom-Classe, Nom-Etat, Nom-Action)
avec Nom-Objet : Nom de l'objet,
Nom-Classe : Nom de la classe dont il hérite,
Nom-Etat : Nom de la table qui définit l'objet,
Nom-Action : Nom de la table des méthodes de l'objet.

La table Nom-Etat contient les caractéristiques des attributs : leur nombre, leur type, leur domaine de définition. Ces attributs peuvent aussi être des pointeurs vers des objets d'autres classes. Il y a autant d'enregistrements dans cette table que d'attributs à spécifier.

La table Nom-Action contient la liste de toutes les fonctions que l'on peut appliquer globalement sur l'objet (pour le créer et le détruire) ou localement sur un attribut. Chaque enregistrement comporte le nom de la fonction, la liste et le type des paramètres d'entrée et de sortie.

Ces tables sont définies sous l'éditeur de texte, ce qui rend facile toute mise-à-jour. La figure 2 illustre quelques attributs et quelques fonctionnalités de l'objet Phonème.

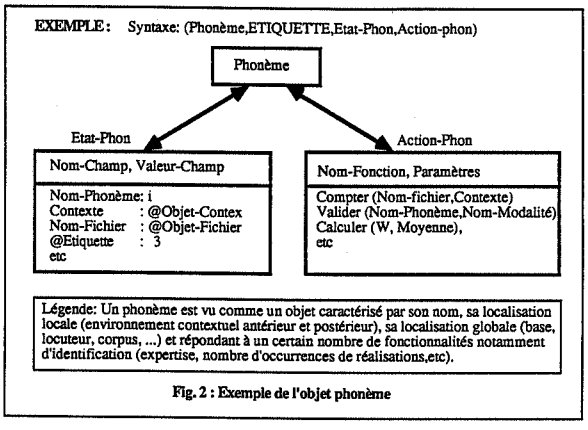


Fig. 2 : Exemple de l'objet phonème

La base de description des objets

L'analyse des données et des cahiers des charges des utilisateurs montrent qu'il est nécessaire de disposer de deux types d'attributs :

- les attributs de type "données" qui représentent des propriétés (exemple, pour un échantillon spectre, sa description est : énergie, durée, fondamental, nom du phonème, ..),
- les attributs de type "relations". Nous distinguons dans ces relations :
  - des liens d'ordonnancement relatifs au mécanisme de production. Ces liens traduisent le déclenchement et le contrôle de l'avancement de génération d'un objet (donc de ses instances). Ils représentent la hiérarchie qui existe entre les différents objets.
  - des liens d'équivalence relatifs à la mise en correspondance. Ces liens permettent de mettre en "bijection" une entité parole vue dans différents plans : par exemple, associer l'ensemble des échantillons-signal et la description phonétique, les données acoustiques et articulatoires.

La figure 3 illustre quelques liens de base entre les données acoustico-phonétiques.

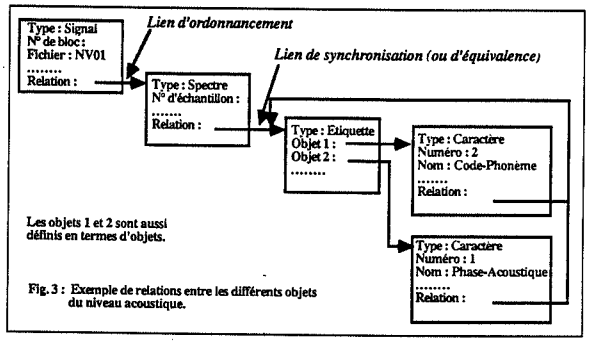


Fig. 3 : Exemple de relations entre les différents objets du niveau acoustique.

Le formalisme objet satisfait pleinement aux exigences du DAP puisqu'il permet dans une même structure d'associer des données de type et de sources différentes et d'établir un "maillage" entre elles.

Par création d'une nouvelle structure de données (i.e. d'un objet), nous pouvons établir des correspondances entre les numéros des blocs signaux et spectre, l'étiquetage et a transcription phonétique. Cette structure de données peut être définie comme l'objet des "relations" dans un système de type relationnel de bases de données. L'instanciation de cette structure donne le dictionnaire des données de la BDAP.

LA CONSULTATION

Objectifs

L'objectif est ici d'offrir à l'utilisateur un ensemble de moyens (objets et méthodes) pour qu'il puisse construire et analyser sa base d'expérimentations. Il ne sera pas traité ici de procédures d'apprentissage automatique. Le phonéticien suit un certain protocole d'expérimentation. Le but à atteindre est de traduire au mieux son protocole de tests et de lui permettre :
- soit d'acquérir une "connaissance" sur une entité quelconque de la base (locuteur, unités linguistiques, ...),
- soit de vérifier une hypothèse qu'il formule a priori à partir des corpus et des locuteurs enregistrés dans la base.

Le système de consultation s'intègre dans un environnement plus général (Cf. Figure 4).

Le système de consultation

Nous proposons d'organiser le système de consultation de la BDAP de la manière suivante :
- la base des objets,
- la base des méthodes,
- la base de travail, résultat de l'opération de consultation sur laquelle l'utilisateur pourra appliquer un deuxième niveau d'interaction [11]. La construction de cette base s'effectue avec des conditions portant sur les attributs des objets et sur les fonctions que l'on peut appliquer sur ces attributs.

Exemple de requête : **SELECTIONNER** l'ensemble des réalisations de l'objet-**ETIQUETTE** instancié par les valeurs des attributs : Phase= **T** et Phonème= **/i/** dont l'instanciation de Objet-**SPECTRE** associé, est une analyse FFT.

Toute requête formulée par l'utilisateur est une commande envoyée au système pour rechercher et/ou produire une information. Toute exécution de requêtes nécessite les phases suivantes :
- la spécification de la requête par l'utilisateur,
- l'exécution proprement dite de la requête.

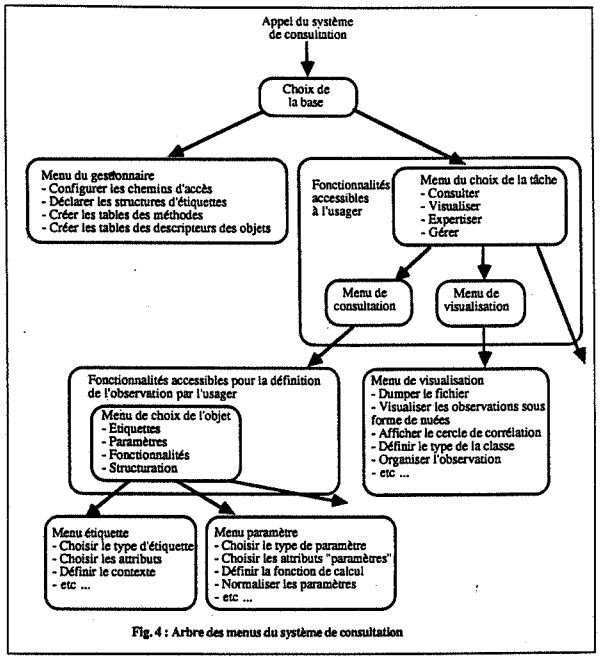


Fig. 4 : Arbre des menus du système de consultation

La figure 5 illustre l'architecture du système de consultation. Il est doté d'un ensemble de fonctionnalités de base permettant la création et la mise-à-jour :

- de nouveaux objets,
- de nouvelles méthodes que l'on peut appliquer sur les attributs des objets pour les instancier.

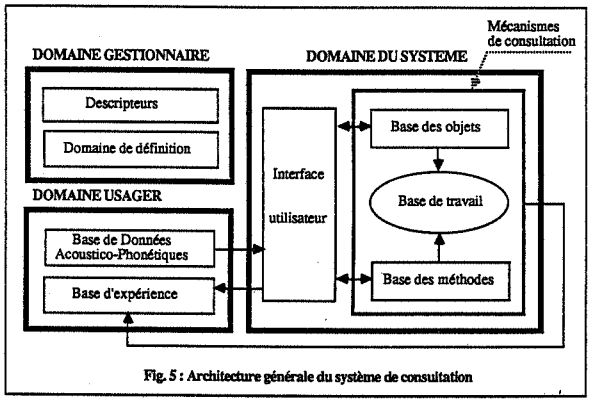


Fig. 5 : Architecture générale du système de consultation

Ce système de consultation est en permanence sous le contrôle de l'expert. Le rôle de ce dernier est primordial dans la spécification de la requête (choix des corpus, des locuteurs, des unités conceptuelles, des systèmes de représentation, ...).

Nous distinguons différents niveaux de complexité de requêtes :

- les requêtes simples qui consistent à sélectionner un objet prédéfini et à lui appliquer un certain nombre de méthodes,
- les requêtes complexes qui font intervenir plusieurs objets déjà définis dans la base des objets. Ces objets produits sont de structure plus complexes. Leur construction nécessite des opérateurs de logique.

Principe d'exécution de la requête

Dans cet article, nous ne considérons que les requêtes faisant intervenir des fonctions de recherche d'informations dans le dictionnaire de la base et des fonctions d'instanciation des attributs.

L'analyse du cahier des charges de l'usager montrent que trois entités apparaissent souvent dans la formulation des requêtes :

- l'environnement global (i.e. la définition du corpus, du locuteur, ...),
- l'unité acoustico-phonétique,
- le système de représentation.

Les exemples donnés ci-dessus sont significatifs des souhaits formulés par les utilisateurs du niveau acoustique. Les requêtes peuvent porter à la fois sur des données numériques ou conceptuelles. Par exemple :

- Lister tous les phonèmes /i/ du fichier x qui ont une modalité fricative;
- Donner la valeur moyenne du paramètre énergie pour les voyelles;
- Donner l'adresse des bloc-signaux des phonèmes qui ont un contexte antérieur nasal ?
- Lister les modalités associées au phonème /i/ pour le locuteur JM
- etc.

Spécification de la requête

Par l'intermédiaire de menus déroulants, l'utilisateur spécifie :

- le choix de la sous-base (locuteur, corpus, type de données),
- la définition des attributs de l'objet-Observation grâce à des menus successifs aux relatifs aux étiquettes répertoriées, aux paramètres, aux fonctions ou plus généralement aux tâches (Cf. Figure 4).

Schéma d'exécution de la requête

Soit la requête : "Donner la valeur moyenne du paramètre énergie pour les voyelles"

Au niveau de la spécification:
Sous-base: Acoustique
Objet: Phonème
Objet: Paramètre
Fonction: Valeur-Moyenne

Au niveau de l'exécution, les fonctions suivantes sont activées :

Instanciation de l'objet Phonème par les attributs de la macro-classe VOYELLE;
Recherche de toutes les réalisations des voyelles dans la base acoustique;
Exécution de la fonction "Moyenne";
Instanciation de l'attribut résultat et **affichage.**

Soit la requête : "Donner l'adresse des blocs-signaux des phonèmes qui ont un contexte antérieur nasal.

Au niveau de la spécification:
Sous-base: Acoustique
Objet: Phonème
Fonction: Filtrage et Localisation
Objet: Adresse

Au niveau de l'exécution, les fonctions suivantes sont activées :

Instanciation de l'objet Phonème par les attributs de la macro-classe NASALE;
Recherche de toutes les réalisations des nasales;
Exécution de la fonction "Filrage" ===> Création d'un objet temporaire de travail;
Accès à l'unité suivante par le parcours des contraintes sémantiques;
Instanciation de l'attribut résultat et **affichage.**

Mise en oeuvre

Dans le système BDAP et de consultation associé, il existe deux formalismes. Le formalisme externe qui est le plus proche de l'usager et le formalisme interne de l'application qui est le langage de l'application : FORTRAN. Ce formalisme présente un double inconvénient :

- la difficulté d'exprimer clairement l'accès aux informations (i.e. leur production). Chaque requête a la forme d'une procédure.
- l'insuffisance en structure de contrôle et en environnement de programmation. La conceptualisation en objet, nous a permis de programmer la séquence de fonctions sur l'objet en entrée pour produire l'objet spécifié en sortie grâce aux parcours des différentes relations. De même, les liens d'équivalence nous permettent de localiser les réalisations des objets définis dans la phase de spécification.

L'implémentation en langage orienté objet permettrait d'éviter l'écriture fastidieuse des règles pour exprimer le "savoir" sur les données acoustico-phonétiques et/ou les fonctions à appliquer [12]. En effet, toute connaissance qui n'est pas liée à un savoir-faire humain mais qui découle des liens logiques existants entre les entités physiques peut-être pris par le mécanisme de l'héritage.

CONCLUSION

Nous avons procédé :

- à l'étude d'un formalisme externe permettant aisément l'introduction de la connaissance des entités manipulés,
- à la structuration des entités du domaine acoustico-phonétique à l'aide du formalisme objet,
- à la prise en compte des requêtes de consultation.

Sur le plan de la réalisation, ce système se présente comme une couche construite autour de la BDAP. Il est souhaitable de coupler cette couche de logiciel avec des interfaces qui permettraient à l'utilisateur de poursuivre son expérimentation avec des procédures d'apprentissage de type stochastique ou inférentielle. Les perspectives sont aussi, d'optimiser le système de requêtes et d'enrichir l'interface homme machine. Parmi les tendances visant à étendre le système de consultation, nous visons à intégrer l'usage du graphique pour améliorer la représentation des objets manipulés sur un poste de travail IBM-PC avec un kit de développement MetaWindow [13]. On tente ainsi, par le biais de l'introduction de l'intelligence graphique des systèmes, à l'élargissement des bases de connaissances couplées avec des bases de données où le processus d'application se mélange avec la gestion elle-même de la base.

REFERENCES

[1] M. Rossi, Niveaux de l'analyse phonétique: nature et structuration des indices et des traits, (Speech Com. Vol. 2, N°2-3, 1983, pp 91-106).

[2] G. Pérennou, M. De Calmes, BDLEX: Lexical Data and Knowledge Base of Spoken and Written French (European Conference on Speech Technology, Edinburgh, 1987).

[3] J. Caelen, G. Caelen-Haumont, N. Vigouroux, C. Barrera, J. Malet, ARCANE: Acquisition et Recherche de Connaissances Acoustico-Phonétiques dans un Noyau Evolutif, (15 e JEP, 1987, pp 207, 211).

[4] O. Cervantes, J.F. Serignat, J. Caelen, D'une Base de Données-Sons vers une Base de Données-Parole, (PRC-BD3, Port-Bacarès, Mai 1987).

[5] V.W. Zue, D.S. Cyphers, R.H. Kassel, D.H. Kaufman, H.C. Leung, M. Randolph, S. Seneff, J.E Unverferth, T. Wilson, The Development of the MIT Lisp-Machine Based Speech Research Workstation, (Proceedings of IEEE-ICASSP, Tokyo, 1986)

[6] C. Barrera, R. Espesser, G. Pérennou, M. Rossi, B. Teston, N. Vigouroux, Acoustic and Articulatory Information in Speech Data Base, (European Conference on Speech Technology, Edinburgh, 1987).

[7] J. Caelen, N. Vigouroux, Producing and Organizing Phonetic Knowledge from Acoustic Facts in Multi-Level Data-Information, (Proceedings of IEEE-ICASSP,Tokyo,1986)

[8] C. Barrera, J. Caelen, G. Caelen-Haumont, J.F.Malet, N. Vigouroux, Towards an Automatic Labelling System, (XI th ICPHs, Talinn,1987).

[9] C. Abry, D. Autessere, C. Barrera, C. Benoit, L.J. Boé, J. Caelen, G. Caelen-Haumont, M. Rossi, R. Sock, N. Vigouroux, Propositions pour la segmentation et l'étiquetage d'une base de données des sons du français, (XIV JEP. ENST-PARIS, 1985, pp 156-163).

[10] P. Cointe, Une Introduction à la Programmation par Objet, (2ièmes Journées Bases de Données Avancées, Giens 1896, pp 335-366).

[11] C. Barrera, A Knowledge-Based System for Voiceless Plosive Decoding, (European Conference on Speech Technology, Edinburgh, 1987).

[12] T. Martelli, L. Miclet, J.P. Tubach, REMORA: a Sofware Architecture for the Collaboration of Different Knowledge Sources in Phonetic Decoding of Continuous Speech, (IEEE-ICASSP, Dallas1987, pp 387-390).

[13] MetaWindow Reference Manual, Metagraphics, SOFWARE CORPORATION.

Les Presses Parisiennes - Tél. : 43 22 39 24