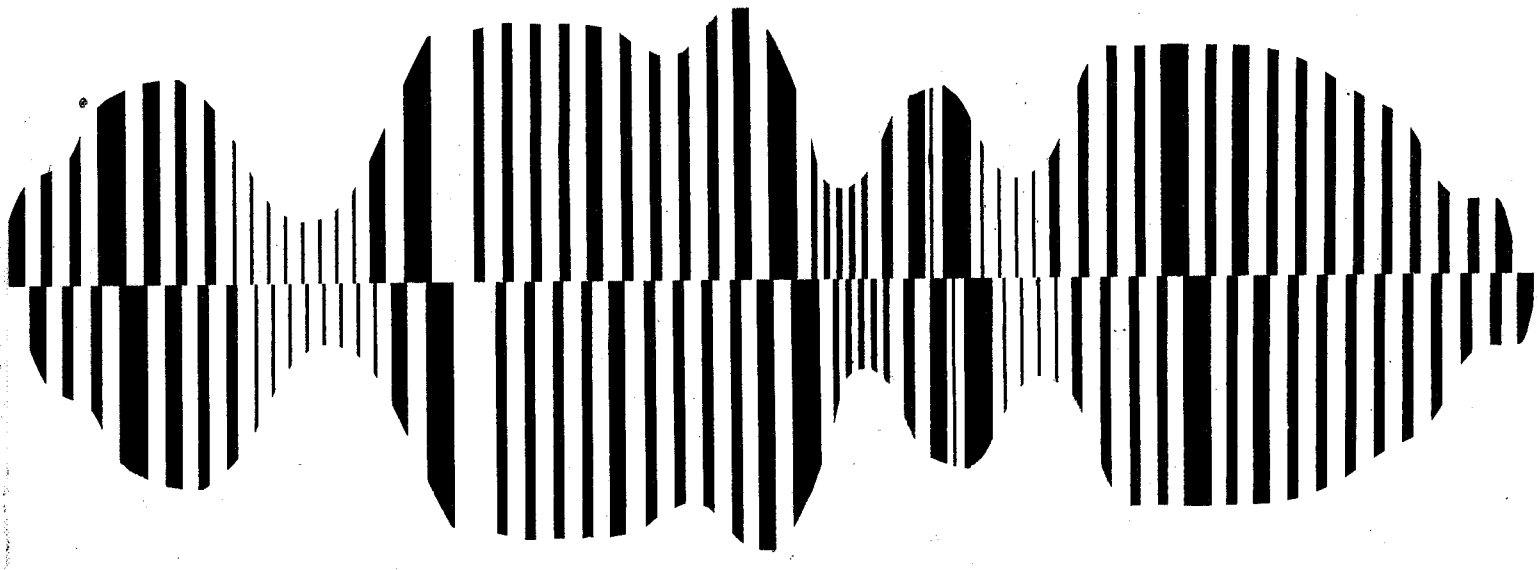


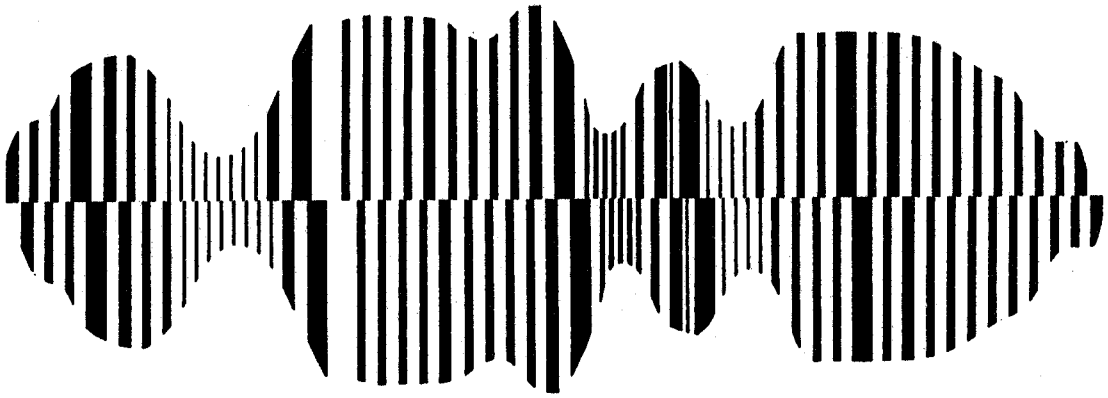
JEP

15e Journées d'Etudes sur la Parole

AIX en PROVENCE 27-30 MAI 86

organisées par le GALF

(Groupe Communication parlée)



JEP

15e Journées d'Etudes sur la Parole AIX en PROVENCE 27-30 MAI 86 organisées par le GALF (Groupe Communication parlée)

Ces journées ont été organisées par l'Institut de Phonétique d'Aix et le Groupe Intelligence artificielle de Marseille-Luminy, dans les locaux de l'Université de Provence à Aix, les 27, 28, 29 et 30 mai 1986, avec le concours du Conseil régional Provence Alpes Côte d'Azur, du Centre National de la Recherche Scientifique et de l'Université de Provence.

TABLE DES MATIERES

SYNTHESE

PHYSIOLOGIE

PRODUCTION

PROSODIE

AUDITION

PERCEPTION

SEGMENTATION

BASES DE DONNEES ET BASES DE CONNAISSANCES

RECONNAISSANCE

ANALYSE

SYNTHESE

Président

Bernard GUERIN

Institut de la Communication parlée de Grenoble

SYNTHETISEUR A FORMANTS NUMERIQUE EN TEMPS REEL. CARACTERISATION.

QUACH TUAN Ngoc et Bernard GUERIN.

Institut de la Communication parlée de Grenoble
I.N.P de Grenoble
46 av. de Félix Viallet, 38000 Grenoble cedex

Abstract: The present built digital formant synthesiser is implanted on a TMS320 signal processing micro-processor. The computations are performed in fixed-point. The structure of the synthesizer itself is a subset of Klatt's. You can either choose a pulse excitation source or an elabored one. The performances in computation time of different components are described in detail precision. They permit to evaluate the feasiblity of real time operations on various synthesizer-structures using the same elementary blocks. As regards the precision problem caused by use of only 16 bits, they are also evoked at the same time as the choice of solution adopted. Finally, an objective comparaison of performances (from temporal and spectral point of view) of the synthsizer , with the same structure and floating-point calculations is presented. It justifies the validity of the implanted synthesizer. The achevement of the synthesizer card, on a mini-computer (the LSI11) will be also presented.

1. Introduction.

De nombreuses études utilisant un synthétiseur à formants nécessitent un système fonctionnant en temps réel. Parmi les applications citons les tests de perception, la synthèse par règle ... Il y a eu plusieurs réalisations de synthétiseur à formants dans le passé. Actuellement, les simulations de synthétiseur sont réalisées sur les ordinateurs numériques afin de créer des sons avec des paramètres variables en fonction du temps. Cette recherche a conduit à réaliser des synthétiseurs avec des circuits analogiques controlés par l'ordinateur |1|, |2| pour atteindre le temps réel. Bien que ces synthétiseurs apportent une solution raisonable pour le temps réel, ils conservent les problèmes typiques des circuits analogiques, c'est à dire de leurs caractéristiques dépendant fortement de la température, du temps, ils demandent une maintenance suivie pour assurer un bon fonctionnement. Pour éviter tous les problèmes associés aux circuits analogiques, les synthétiseurs ont d'abord été simulés sur ordinateur numérique |6|,|7|,|8|. Les avantages pour tous types de synthétiseurs simulés numériquement sont: la précision de contrôle des paramètres (des formants, des bandes passantes, des gains, le raport Signal/Bruit), la souplesse de configuration de la structure du synthétiseur, la facilité de modification des données. Un autre avantage est que les synthétiseurs complexes peuvent être conçus et simulés facilement. Le grand désavantage était que jusqu'à maitenant la plupart des mini-ordinateurs ne peuvaient pas assurer en temps réel pour des sons échantillonnés à 10 kHz, valeur fréquemment retenue.

Avec le dévelopement de la technologie, les calculateurs spécialisés de traitement de signal à grande vitesse ont vu le jour comme le TMS320 du Texas Instrument, 7720 du NEC ... Ils peuvent être connectés aux ordinateurs comme des périphériques pour réaliser des systèmes de traitement de signal en temps réel, le synthétiseur à formants numérique étant un exemple. Le principal problème posé est la précision des calculs effectués en virgule fixe. Dans cet article, nous décrivons un tel synthétiseur qui a été conçu et réalisé à l'Institut de la Communication parlée de Grenoble.

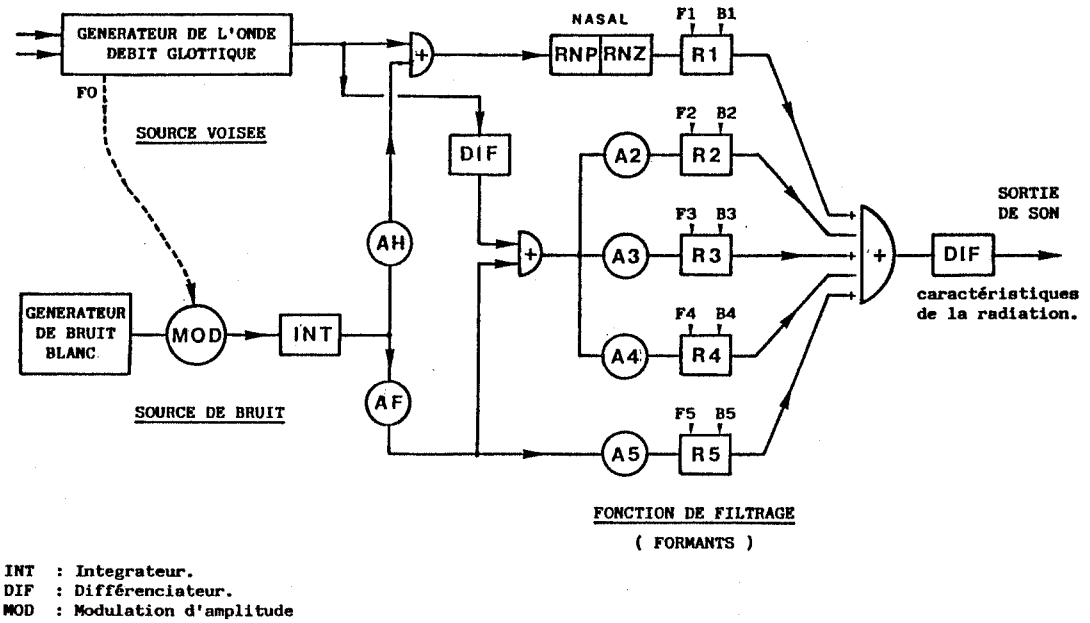
2. Decription du système informatique.

Un micro-processeur spécialisé de traitement de signal à grande vitesse, le TMS32010 a été connécté comme un périphérique à l'ordinateur de la famille LSI11. Les différents types des synthétiseurs sont développés sous forme logiciel (micro-programmé sur le TMS320). Le LSI11 gère les données (des paramètres de contrôle) stockés dans des fichiers, et les fournit au TMS320, il peut récupérer, si besion est, les échantillons calculés par le TMS320.

3.Structure du synthétiseur retenu.

On a retenu une structure parallèle pour le synthétiseur à formants numérique. Cette structure est en fait un sous ensemble du synthétiseur de D. KLATT (1980) ,|7|, avec des simplifications et des améliorations.

Nous avons réalisé plusieurs synthétiseurs micro-programmés comportant plusieurs choix de la source d'excitation. En général, le synthétiseur a deux parties distinctes: la source d'excitation et les filtres (les formants). La figure 1 montre lastructure détaillée du synthétiseur. Nous allons décrire au fur à mesure les différentes parties du



INT : Integrateur.
DIF : Différenciateur.
MOD : Modulation d'amplitude

Figure 1: La structure retenue du synthétiseur à formants numériques.

synthétiseurs.

La partie formants se compose de 5 filtres passe-bas résonateur (5 formants) R1, R2, R3, R4 et R5 connectés en parallèle. Les 4 premiers formants sont éxcités par la source voisée tandis que le cinquième n'est excité que par la source de bruit. Plusieurs recherches de notre laboratoire ont indiqué que cette structure de 5 formants en parallèle était capable de produire tous les sons du français.

A la sortie des filtres de formants, les signaux sont additionnés avec des signes alternatifs pour simuler correctement la fonction de transfert. Bien que la connection en série des formants simule mieux la fonction de transfert du conduit vocal dans le cas des voyelles, la structure parallèle a été choisie car elle permet de produire plus correctement l'ensemble des sons et en particulier les consonnes. Parmi plusieurs combinaisons des sortie des formants connectés en parallèle, celle en signe alternatif est la meilleure.

Pour simuler l'effet de nasalisation, dans la version originale de Klatt (1980), un formant suplémentaire RNP a été ajouté pour créer une paire pôle-zéro par combinaison avec le premier formant et en réalisant l'addition avec le même signe. D'après nos tests, cette simulation ne donne pas perceptivement des sons nasalisés de bonne qualité. Aussi, nous avons connecté en série sur la première branche (premier formant R1) deux filtres: un filtre résonant RNP et un filtre anti-résonnant RNZ afin de simuler l'effet de nasalité par une paire Pole-Zéro.

Quant à la source d'excitation, elle peut être de 2 types: source voisée pour les sons voisés, source de bruit pour les sons bruités comme fricatives, plosives ou aspirés.

La source de bruit est réalisée à l'aide d'un générateur des nombres aléatoires qui fournit un signal de densité de probabilité uniforme avec un spectre plat. Un tel signal simule bien les sons bruités naturels de distribution gaussienne car l'homme ne perçoit pas la différence entre les deux distributions.

La source voisée peut être réalisée de plusieurs façons (le bloc "générateur de l'onde de débit glottique" sur la figure 1 n'est qu'une présentation générale). Ici nous décrivons deux possibilités du choix de la source voisée: la source impulsionnelle classique et une source élaborée.

La source impulsionnelle est celle utilisée dans le synthétiseur original de KLATT |7|. Elle se compose d'un générateur impulsionnel suivi des filtres passe-bas RGP et RGS, du filtre anti-formant RGZ pour la mise en forme du spectre de la source. Une telle source ne tient pas compte des caractéristiques de phase du spectre.

La source élaborée est décrite dans le paragraphe suivant.

4. La source élaborée:

Depuis plusieurs années, en vue de l'amélioration de la qualité des sons synthétiques, les chercheurs ont songé à améliorer la source d'excitation en fournissant une forme convenable de l'onde de débit glottique et de ses caractéristiques. Parmi plusieurs modèles de la source vocale, le modèle à deux masses semble nous donner les bons résultats. De plus, B.GUERIN (1978) a montré que ce système peut être commandé par deux paramètres de type physiologique: la pression subglottique Ps et un coefficient Q donnant une bonne idée de la variation de la tension longitudinale et de la diminition de masse linéïque. Les caractéristiques de l'onde de débit fournie par le modèle à deux masses peuvent être

représentées de façons assez simple, aussi les avons-nous retenues pour commander un générateur de l'onde de débit. D'autre part, différentes études (G.FANT 1979, 1980) ont montré que le signal de l'onde de débit glottique peut être approché par des segments des cosinus. G.FANT a proposé un modèle simple de ce signal et nous avons pris cette forme de l'onde dans le générateur de l'onde glottique |5|.

Le principe de la ré alisation de la source élaborée est basé sur l'utilisation des caractéristiques du modèle à deux masses avec ses deux paramètres de commande Ps et Q. Celles-ci sont incorporées au modèle de génération de l'onde de débit proposé par FANT. (cf. figure 2). L'utilisation de cette source est intéressante car elle permettra d'introduire une commande de type physiologique des paramètres de la source de débit, ceci n'étant pas possible dans les modèles proposés jusqu'à présent.

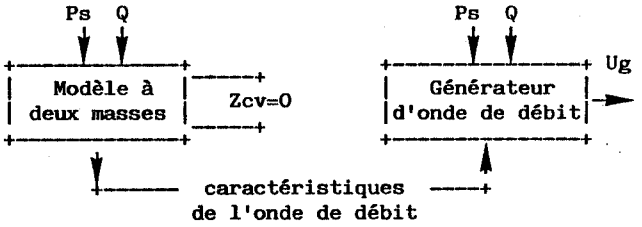


Figure 2: Principe de la réalisation de la source élaborée.

5. Les problèmes liés à la réalisation en virgule fixe.

Les erreurs commises lors de calcul en virgule fixe sont:

- erreurs de la quantification des coefficients des filtres, des échantillons du signal d'entrée x(n);
- erreurs liées à l'opération de multiplication pour garder une longueur constante du mot dans le filtre.
- et enfin, le débordement.

Ces sources d'erreurs causent des phénomènes inévitables comme le cycle limite dans les filtres récusifs.

Le calcul en nombre fractionnaire est une forme couramment utilisée car on peut éviter tout débordement de capacité lors d'une opération de multiplication. Appliquer au calcul du filtre résonant numérique, les coefficients de la cellule de base dans le synthétiseur à formants sont caractérisés par les relations:

$$R = \exp(-PI.BW.T)$$
$$R2 = R.R$$
$$RC = R.\cos(2.PI.Fr.T)$$

où R, R2 et RC sont les coefficients du filtre (sous forme de nombres fractionnaires).

et:

$C = -R2$ avec $-1 \leqslant C \leqslant 0$

$B = 2.RC$ avec $-2 \leqslant B \leqslant 2$

$A = 1-2.RC+R2$ avec $0 \leqslant A \leqslant 4$

et le filtre à formant est calculé comme:

$$y(n) = A.x(n) + B.y(n-1) + C.y(n-2)$$
$$= x(n) - RC.x(n) - RC.x(n) + R2.x(n) + RC.y(n-1) + RC.y(n-1) - R2.y(n-2)$$

Ainsi toutes les opérations de multiplications où interviennt entre autre les coefficients R2 et RC sont faites sous forme fractionnaire. La structure correspondante avec un seul "bloc Q", ce dernier représente l'effet d'erreur dû à la quatification après une mutiplication, est donnée sur la figure 3 ci-dessous. Ainsi, au lieu de faire l'arrondi après chaque mutiplication (il y en a 3 sur la figure), nous gardons le résultat complet en double précision de la mutiplication et ne faisons l'arrondi (le bloc Q|.|) qu'une seule fois après avoir fait la somme des 3 résultats des mutiplications. Cette configuration avec un bloc Q|.| a diminué le bruit de quatification de l'ordre de 3 à 4 fois par rapport à celle avec 3 blocs Q|.|.

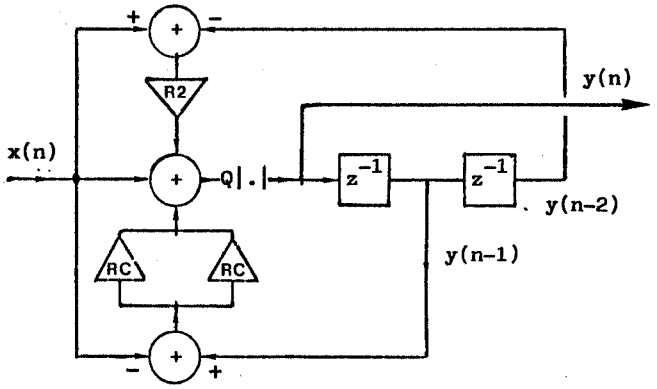


Figure 3: Configuration du résonateur sous forme fractionnaire. (avec "un Q")

La figure 4 donne le résultat de la comparaison entre la réponse fréquencielle du synthétiseur calculée en flottant et celle calculée en virgule fixe. L'évaluation du rapport Signal/Bruit du synthétiseur donne une valeur de l'ordre de 55 dB. Le bruit à la sortie du synthétiseur est environ de 2,5 bits. Ce résultat est appréciable car il n'y a que les 12 bits de poids les plus forts qui sont pris par le convertiseur N/A (16 bits fournis par le calcul).

6. Les performances.

Avec la structure définie ci-dessus, soit avec la source impulsionnelle, soit avec la source élaborée, le synthétiseur fonctionne en temps réel avec une fréquence d'échantillonnage Fe=10 kHz. Par ailleurs avec la même structure on peut atteindre une fréquence d'échantillonnage maximum de 13,5 kHz. Cela veut dire aussi qu'avec Fe=10 kHz, nous pouvons encore ajouter de l'ordre de 4 filtres au synthétiseur si cela est nécessaire. Par exemple, la réalisation du synthétiseur avec une structure mixte comprenant en plus 3 résonateurs connectés en série avec le premier formant.

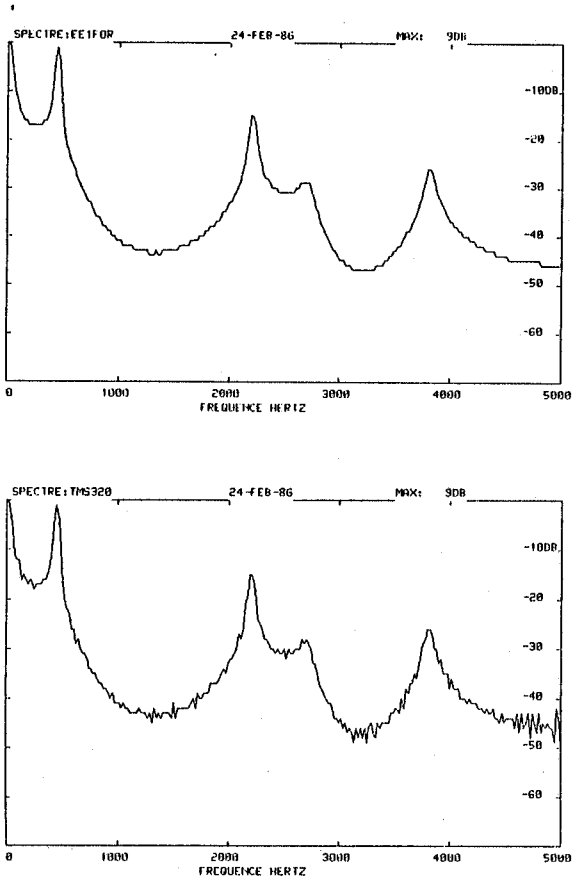


Figure 4: Spectre de la réponse impulsionnelle du synthétiseur calculé en virgule flottante (a) et en virgule fixe (b) avec la voyelle /e/.

Le logiciel du synthétiseur est découpé en module. On a mesuré pendant le calcul pour un échantillon les contributions suivantes:

- Module de la source impulsionnelle. (15 µs)
- Module de la source élaborée. (16 µs)
- Module de la source de bruit. (7 µs)
- Module des formants. (45 µs)

Donc le temps total de calcul pour un échantillon avec une source quelconque est d'environ 70 µs.

La seule charge du LSI11 est de fournir les paramètres de commande au TMS320. Aussi, le TMS320 utilise environ 300 µs sur une fenêtre de 5000 µs (5 ms) pour convertir les paramètres de commande reçus du LSI11 en forme convenable, par exemple calculer les coefficients des filtres à partir de la fréquence de résonnance et de la bande passante des formants.

7. Conclusion.

Le synthétiseur à formants implanté sur un processeur de traitement de signal TMS320 permet d'obtenir une synthèse de bonne qualité en temps réel. Toutes ses caractéristiques ont été validées parcomparaison avec un modèle de référence simulé en virgule flottante. La conception sous forme de bloc: sources, filtre de formants, offre une grande souplesse et permet d'envisager facilement des structures différentes de synthétiseurs adaptées à des besoins de recheche particulier.

BIBLIOGRAPHIE

|1| AWAD S. (1983)
Synthétiseur à formants en temps réel. Etude d'une source d'éxcitation élaborée. Codage optimum des paramètres de commande.
Thèse de Docteur-Ingénieur I.N.P de Grenoble.

|2| GOLD B. & RABINER R.L. (1968)
Analysis of Digital and Analog Formant Synthesizers.
I.E.E.E Trans. Au-16, 1.

|3| GUERIN.B & BOE L.J(1977)
A two mass model of the vocal cords: Determination of control parameters and their respective consequence.
I.C.A.S.S.P, Hartford, 582-586.

|4| GUERIN.B (1978)
Contribution aux recherches sur la production de la parole. Etude du fonctionnement de la source vocale. La simulation d'un modèle.
Thèse Doctorat-d'Etat I.N.P de Grenoble.

|5| FANT.G (1979)
Vocal source analysis. A progress report.
STL-QPSR 3/4, pp 31-53

|6| HOLMES J.N. (1973)
The influence of glottal waveform on the naturelness of speech from a parallel formant synthsizer.
I.E.E.E , Au-21, 298-305.

|7| KLATT. D.H (1980)
Software for a cascade/parallel formant synthesizer
J.Acoust.Soc.Am. 67, 971-996.

|8| RABINER R.L (1968)
Digital formant synthesizer for speech-syntheis studies.
J.Acoust.Soc.Am. 43, 822-828.

|9| QUACH TUAN N. & GUERIN B. (1985)
Module temps réel de la source vocale élaborée pour synthétiseur à formants.
14 eme JEP. Paris.

SYNTHESE A PARTIR DU TEXTE PAR MODIFICATION ET CONCATENATION DE FORMES D'ONDE

F. Charpentier et M. Stella

Centre National d'Etudes des Télécommunications
22301 LANNION CEDEX FRANCE

RESUME

On présente ici une nouvelle méthode de synthèse à partir du texte utilisant les diphones comme unités de parole. Le dictionnaire de diphones est constitué des formes d'onde elles-mêmes, munies d'un étiquetage des périodes de la fréquence fondamentale (pitch). Au cours de la synthèse, les formes d'ondes sont traitées par un système d'analyse-synthèse permettant un contrôle indépendant de tous les paramètres prosodiques, tout en conservant un timbre de parole naturelle. Ce système d'analyse-synthèse part d'une représentation du signal de parole par sa tranformée de Fourier à court terme (TFCT) à une cadence d'analyse pitch-synchrone. L'algorithme de synthèse fonctionne par recouvrement et addition de petits signaux élémentaires à court terme, ce qui garantit une concaténation douce des formes d'ondes. La parole synthétique obtenue par cette méthode possède une qualité plus naturelle que celle obtenue par prédiction linéaire.

INTRODUCTION

La technique la plus répandue pour la synthèse par diphones ou demi-syllabes est le codage prédictif des éléments de parole. Ce procédé est pratique car il permet un contrôle très souple des paramètres prosodiques à travers le signal d'excitation du filtre de prédiction linéaire. Malheureusement, ce type de codage dégrade la parole naturelle, particulièrement pour des sons comme les fricatives voisées. On peut améliorer cette qualité en utilisant des modèles plus sophistiqués du signal d'excitation. Ainsi, plusieurs chercheurs ont tenté d'utiliser le modèle multiimpulsionnel du signal d'excitation /1-3/, mais ce modèle ne permet pas vraiment de varier les paramètres prosodiques de manière souple. Dans une expérience récente /4/, nous avons obtenu une qualité de timbre assez naturelle par concaténation brute de diphones codés par la méthode multiimpulsionnelle. Mais la parole ainsi obtenue devait alors nécessairement subir une correction prosodique, traitement qui introduisait un caractère rauque indésirable, et ce malgré l'utilisation d'un système d'analyse-synthèse de haute qualité fondé sur la synthèse de Fourier à court-terme /5/.

Dans ce papier, nous proposons de bâtir un nouveau système de synthèse entièrement sur une représentation de Fourier à court terme de la parole. Beaucoup de systèmes fondés sur des modifications de la TFCT ont été proposés pour manipuler certains paramètres de la parole tout en conservant une qualité de timbre naturelle. Les premiers systèmes de cette sorte utilisaient une approche où la TFCT est interprétée comme un banc de filtres, et le signal de synthèse était obtenu par addition des sorties de l'ensemble des filtres. Mais ceci conduisait à des algorithmes trop complexes /5,6/. L'approche des méthodes par recouvrement et addition (en anglais, OverLap and Add, OLA par abréviation) est préférable parce qu'elle permet des algorithmes plus rapides, et plus simples conceptuellement /7,8/. Les méthodes OLA fonctionnent toutes par décomposition du signal en une suite de signaux élémentaires à court terme (signaux à CT), qui se recouvrent partiellement et qui se succédent à une certaine cadence d'analyse. Ces signaux à CT sont ensuite convenablement modifiés puis additionnés à une cadence différente. Une cadence de synthèse variable, comme il a été proposé pour la méthode OLA synchrone (SOLA), permet de conserver une parole synthétique de très bonne qualité tout en modifiant la vitesse d'articulation /9/. L'intérêt de cette cadence non-uniforme est de permettre une resynchronisation des signaux à CT avec les périodes effectives du signal de synthèse.

Nous présentons ici une méthode OLA entièrement pitch-synchrone (PSOLA), qui vient généraliser la méthode SOLA en introduisant le pitch-synchronisme dès l'étape d'analyse. Le principe est de faire coïncider chaque signal à CT avec une période de pitch, que ce soit à l'analyse ou à la synthèse. Une étape préliminaire est donc nécessaire afin de segmenter le signal en périodes de pitch consécutives. Une importante innovation de la méthode PSOLA consiste à intégrer dans le processus de recouvrement et addition un algorithme de modification du pitch et de l'enveloppe spectrale. Ceci est obtenu par des opérations relativement simples dans le domaine fréquentiel, qui ne requièrent pas de calcul de phase. Finalement, le système d'analyse-synthèse PSOLA fournit une outil très souple pour le contrôle indépendant des paramètres prosodiques, et il constitue ainsi une alternative intéressante à la technique de prédiction linéaire pour la synthèse à partir du texte.

DESCRIPTION DU SYSTEME D'ANALYSE-SYNTHESE

Dans le système d'analyse-synthèse PSOLA, le signal de parole est analysé en une succession de signaux à CT, pitch-synchrones. Les signaux à CT sont alors modifiés dans le domaine temporel ou dans le domaine fréquentiel, de facon à obtenir des signaux à CT de synthèse, synchrones d'un contour mélodique modifié. La parole synthétique est obtenue par recouvrement et addition de ces signaux à CT modifiés. Nous allons maintenant décrire avec plus de précision les différentes parties de l'algorithme.

Analyse Pitch-synchrone

On choisit tout d'abord une longueur convenable de la fenêtre d'analyse, de façon à toujours englober au moins trois périodes de pitch. Nous avons utilisé une fenêtre de Hanning de 512 points pour traiter de la parole échantillonnée à 16 kHz, ce qui correspond à une durée de 30 ms environ.

La représentation initiale du signal est sa forme d'onde digitalisée, munie de "pitch-marques", c'est-à-dire de marqueurs de pitch distribués successivement le long de l'échelle des temps. Ces pitch-marques (P-marques) sont disposées de manière pitch-synchrone sur les portions voisées, et de manière relativement arbitraire sur les portions sourdes, pourvu que deux fenêtres successives présentent un recouvrement suffisant. Les signaux à CT sont obtenus en multipliant le signal original par la fenêtre d'analyse centrée sur la P-marque correspondante. Chacun de ces signaux à CT est également muni d'un indicateur de voisement associé à sa P-marque.

De cette façon, le signal est décomposé en une suite de signaux à CT pitch-synchrones et se recouvrant partiellement. Les signaux à CT non-voisés sont conservés dans le domaine temporel car ils ne nécessitent qu'un changement de l'échelle des temps. Quant aux portions voisées, il est nécessaire de passer à une représentation fréquentielle: le spectre à CT est calculé en prenant la DFT du signal à CT et en faisant coïncider l'origine des temps avec la P-marque; le spectre est alors décomposé en une enveloppe spectrale globale, et une "composante de source", qui s'obtient en divisant le spectre par l'enveloppe.

Modifications dans le domaine fréquentiel

L'algorithme de modification du pitch et de l'enveloppe est décrit sur la Fig.1. Ces modifications sont obtenues par un traitement fréquentiel des signaux à CT. Les composantes spectrales d'enveloppe et de source sont interpolées séparément de facon à obtenir deux changements d'échelle différents de l'axe des fréquences. Dans le cas de la source, on effectue une interpolation linéaire sur les parties réelles et imaginaires du spectre complexe. Si nécessaire, des hautes fréquences sont introduites par recopie d'une partie des basses fréquences dans les aigus. Un tel changement d'échelle résulte en une modification du pitch. Afin d'éviter des corrections indispensables du spectre de phase, les intervalles séparant les P-marques sont corrélativenent modifiées par le facteur de modification du pitch, ce qui entraîne une modification concomitante de l'échelle des temps. L'enveloppe spectrale est également susceptible de modifications, ce qui est utile pour varier le timbre de la voix. Ainsi, une voix féminine peut être convertie en une voix à caractère masculin ou enfantin simplement en abaissant ou en relevant le niveau formantique par un facteur moyen de 15%, et en corrigeant le niveau du fondamental de manière appropriée.

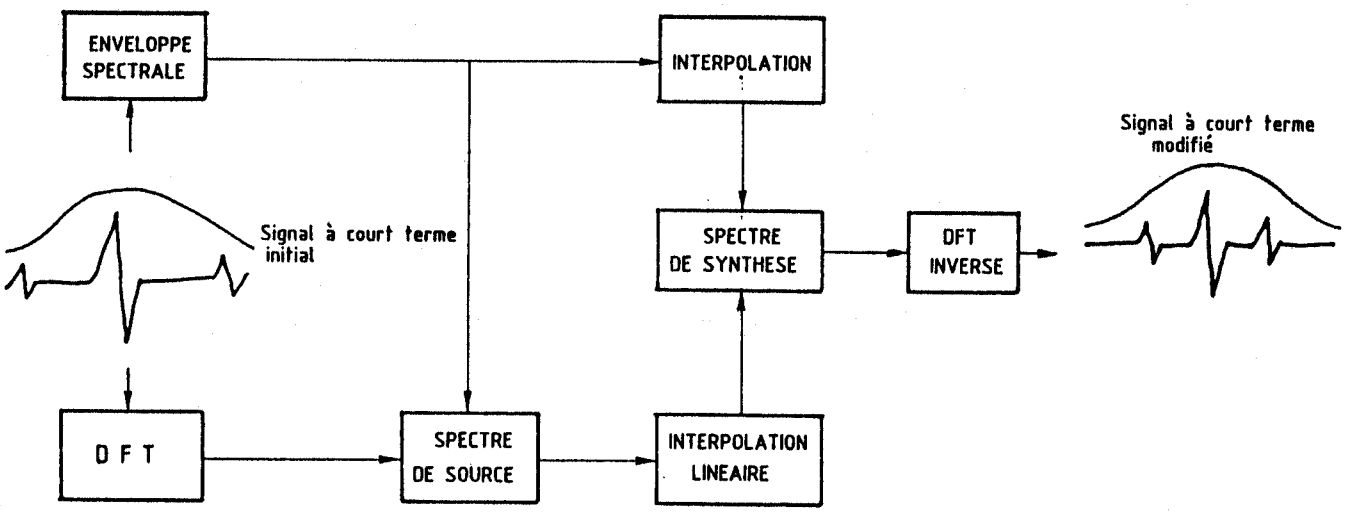


Fig.1 Algorithme de modification de la fréquence fondamentale et de l'enveloppe spectrale.

Modifications de l'échelle des temps

Les modifications de l'échelle des temps sont réalisées entièrement dans le domaine temporel. Les signaux voisés sont traités par cette partie de l'algorithme seulement après avoir subi les traitements dans le domaine fréquentiel. Dans ce cas, l'algorithme doit compenser les modifications de la durée implicites à celles du pitch.

Le fonctionnement de algorithme est illustré sur la Fig.2. Les modifications souhaitées de la durées définissent une fonction de déformation de l'échelle de temps, entre les deux axes temporels d'analyse et de synthèse. A partir de la distribution des P-marques d'analyse, l'algorithme engendre un nouvel ensemble de P-marques de telle façon qu'il conserve le contour mélodique à travers la fonction de déformation. En général, une P-marque de synthèse ne coïncidera pas avec une P-marque d'analyse. Une méthode d'approximation est donc nécessaire pour assigner un signal à CT convenable à chaque P-marque de synthèse. La solution la plus simple consiste à lui assigner le signal à CT d'analyse correspondant à la P-marque d'analyse la plus proche. Ceci revient à une annulation ou à une réplication de certaines périodes de pitch. Un algorithme légèrement plus compliqué a été utilisé, consistant à faire une moyenne entre les deux signaux à CT d'analyse les plus proches.

Synthèse pitch-synchrone par recouvrement

La synthèse d'un signal temporel à partir de spectres à CT se succédant dans le temps à une cadence non-uniforme est possible si l'on utilise la méthode de synthèse OLA des moindres carrés /8/. Le principe de cette méthode est de minimiser les écarts quadratiques entre la suite des TFCT désirées et les TFCT effectives du signal de synthèse. La cadence non-uniforme du traitement est prise en compte par le calcul d'une fonction temporelle de pondération, qui sert à corriger le signal qui serait obtenu par une simple addition des signaux à CT modifiés.

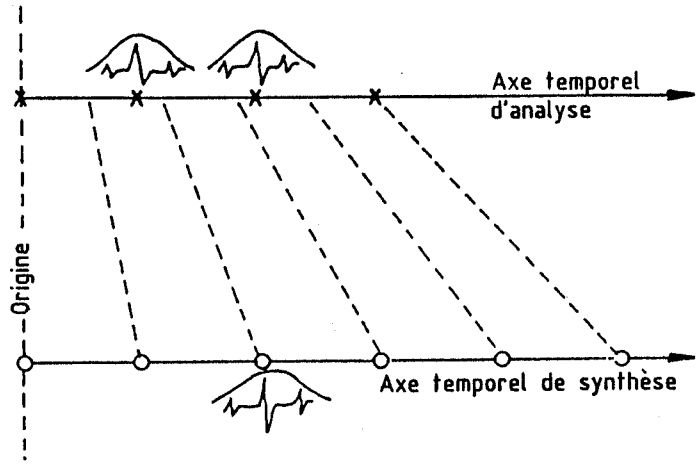


Fig.2 Changement de l'échelle des temps.
Les croix et les cercles désignent respectivement les P-marques d'analyse et de synthèse; les traits tiretés représentent la déformation temporelle.

LE SYSTEME DE SYNTHESE PAR DIPHONES

Ce système d'analyse-synthèse a été appliqué à la synthèse par diphones du français. Le diagramme du système de synthèse est présenté sur la Fig.3. Le dictionnaire de diphones est constitué de deux sources parallèles d'information: les formes d'ondes proprement dites et les P-marques correspondantes. L'étiquetage des périodes de pitch des 1200 diphones a été obtenu par une méthode semi-automatique. Les P-marques ont été posées visuellement d'une manière cohérente de façon à correspondre aux maximas d'amplitude de la forme d'onde de chaque période, ce qui correspond à peu près à l'instant de fermeture de la glotte. Les portions non-voisées ont été découpées en fenêtres de longueurs fixes.

Au cours de la synthèse, les formes d'ondes des diphones sont concaténées comme suit:

(a) chaque diphone est traité par le système d'analyse-synthèse PSOLA de façon à corriger sa durée et à réaliser un contour mélodique souhaité; les paramètres prosodiques sont obtenus par le module prosodique de notre système de synthèse LPC conventionnel /10/; les facteurs de modifications du pitch et de la durée sont obtenus en comparant la prosodie désirée à la prosodie intrinsèque du diphone;

(b) à la frontière entre deux diphones, les signaux à CT du diphone suivant sont tout simplement enchaînés à ceux du diphone précédent, et le lissage est implicitement assuré par le procédé de recouvrement et addition; une correction énergétique est effectuée sur le diphone suivant afin d'éliminer les éventuelles discontinuités d'énergie.

RESULTATS

Deux dictionnaires de diphones, correspondant à deux locuteurs masculin et féminin, ont été préparés afin de tester notre méthode de synthèse à partir du texte. La voix de synthèse obtenue avec les deux voix est plus naturelle avec notre méthode qu'avec la méthode conventionnelle par prédiction linéaire. En particulier, elle ne présente plus le défaut habituel d'un timbre métallique. Il y a toujours un léger caractère artificiel qui vient probablement des discontinuités formantiques. Mais de nouvelles améliorations sont espérées dès lors que le système permet un contrôle souple de tous les paramètres prosodiques ainsi que de l'enveloppe spectrale.

L'algorithme PSOLA est implanté sur un processeur vectoriel et tourne en 20 fois le temps réel pour de la parole naturelle échantillonnée à 16 KHz. Le système de synthèse à partir du texte produit une seconde de parole en une minute environ. La taille en mémoire du dictionnaire de diphones est assez importante (7 Moctets). Par l'utilisation de méthodes appropriées de compression ou de codage de la forme d'onde, il est vraisemblable que l'espace de stockage nécessaire pour les données de parole pourrait être réduite à une taille raisonnable, en tout cas comparable à celle des bases de données lexicales nécessaires pour une synthèse à partir du texte de haute qualité.

CONCLUSION

Un algorithme OLA pitch-synchrone a été présenté dans ce papier pour des modifications prosodiques de haute qualité de la parole naturelle. Il requiert une présegmentation des périodes de pitch par utilisation d'une méthode semi-automatique, mais cette tâche est abordable pour la synthèse par diphones étant donnée la taille relativement limitée de la base de donnée de signal du corpus des diphones. On a pu donc concevoir un nouveau système de synthèse à partir du texte, qui produit une parole synthétique d'une meilleure qualité que la prédiction linéaire. Grâce à la méthode PSOLA, les discontinuités de pitch et d'énergie peuvent être complètement éliminées. Cependant, on a besoin d'une meilleure connaissance des discontinuités formantiques entre les diphones, et de techniques pour en réduire l'effet perceptuel. Mais d'ores et déjà, le principal avantage de notre méthode est de retenir les détails fins du spectre de la parole sans présenter les effets de lissages ou de distortion inhérents aux méthodes de synthèse par formants ou par prédiction linéaire.

REMERCIEMENTS

F. Emerard nous a apporté une aide précieuse par sa segmentation du corpus des diphones pour les deux voix féminine et masculine.

REFERENCES

/1/ B.E. Caspers, B.S. Atal, "Changing pitch and duration in LPC synthesized speech using multipulse excitation", JASA, 73, S5, Spring 83

/2/ J.P. Van Hemert, "Multipulse Excitation: the possibilities and restrictions of a new speech synthesizer", IPO annual progress report No.19, 20-24, 1984

/3/ M. Stella, "Modification des paramètres prosodiques en analyse-synthèse multiimpulsionnelle", Actes des 13e JEP, Mai 84, 151-152

/4/ M. Stella, F. Charpentier, "Synthèse par diphones utilisant le codage prédictif multiimpulsionnel et un vocodeur de phase", Actes des 14e JEP, Juin 85, 101-104

/5/ S.S. Seneff, "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction ", IEEE Trans. ASSP, 30(4), Aug.82, 566-578

/6/ J.L. Flanagan, R. Golden, "Phase vocoder", Bell Syst. Tech. J., 45, Nov.66, 1494-1509

/7/ J.B. Allen, L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis" Proc. IEEE, 65(11), Nov.77, 1558-1564

/8/ D.W. Griffin, J.S. Lim, "Signal estimation from modified short-time Fourier transform", IEEE Trans. ASSP, 32(2), Apr.84, 236-243

/9/ S. Roucos, A. Wilgus, "High quality time-scale modification for speech", Proc. ICASSP, Mars 85, 493-496

/10/ J.L. Courbon, F. Emerard, "SPARTE: a text-to-speech machine using synthesis by diphones", Proc. ICASSP, Mai 82, 1597-1600

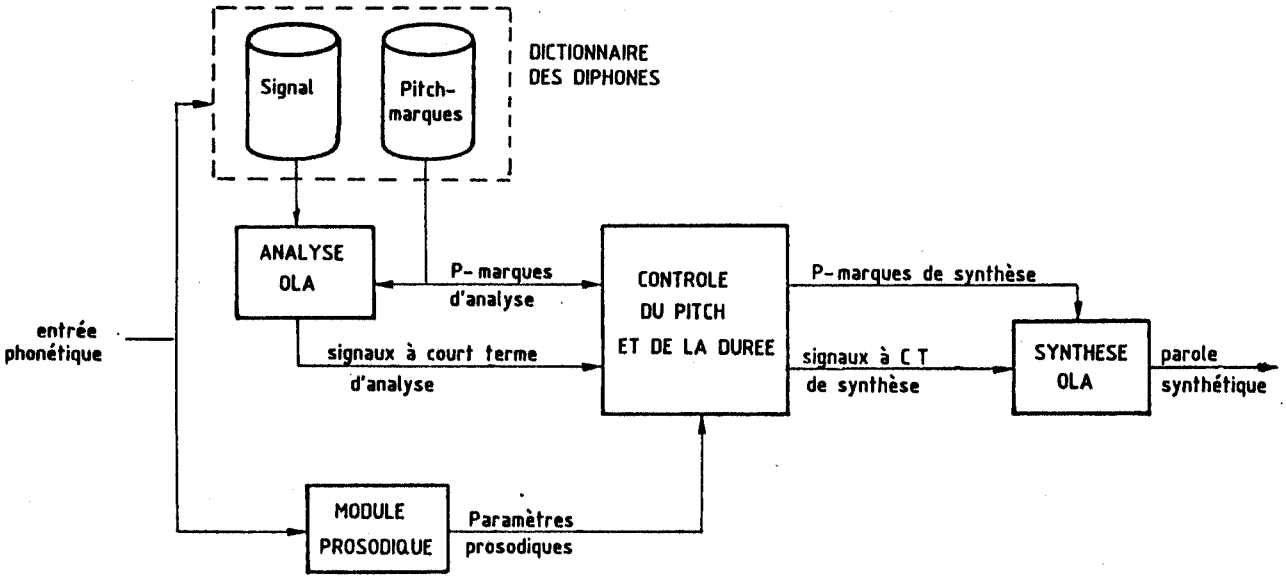


Fig.3 Organisation générale du système de synthèse par diphones.

ANALYSE-SYNTHÈSE ET ÉTUDES DE RÈGLES ACOUSTIQUES DE PRODUCTION AVEC UN SYNTHÉTISEUR À FORMANTS

F. Laferrière, D. O'Shaughnessy

INRS-Télécommunications, 3 Place du Commerce, Ile des Soeurs, Québec, H3T-1H6, Canada

Abstract

A complete and straightforward interactive analysis-synthesis system for a formant synthetizer has been developed. The synthetizer architecture used is an enhanced version of the Klatt synthetizer ([1]). The extraction of parameters is done interactively on short windows. The natural signal can be completely analysed and resynthesized frame by frame, but can also be synthesized using a description by targets and trajectories for the parameters. This is achieved by using a dedicated language for this description, through a software interpreter created for this purpose. This description of speech by targets and trajectories is used for the acoustic studies and elaboration of production rules for high quality rule-based speech synthesis. The method is used on French nasal vowels.

1. Introduction

L'élaboration d'un système de règles de production pour faire de la synthèse de parole *ab initio*, c'est-à-dire n'utilisant aucune information venant directement d'une base de données verbale (contrairement au cas de la synthèse par diphones par exemple) constitue un défi important. LOQUAX, le système de synthèse de parole développé à l'INRS-Télécommunications permet déjà de faire de la synthèse *ab initio* à partir du texte [2]. Néanmoins, la recherche de l'excellence dans ce domaine suppose l'amélioration d'un ensemble de règles déjà fort complexe.

Au niveau de production acoustique des phonèmes, la synthèse par règles utilisée dans LOQUAX peut être schématisée de la façon suivante :

$$\left\{\begin{array}{c}\text{description}\\ \text{par règles}\end{array}\right\} \Rightarrow \left\{\begin{array}{c}\text{description par}\\ \text{trajectoires}\end{array}\right\} \Rightarrow \left\{\begin{array}{c}\text{description}\\ \text{par trames}\end{array}\right\} \Rightarrow \left\{\begin{array}{c}\text{parole}\\ \text{synthétisée}\end{array}\right\}$$

Les règles de production permettent de générer une description du phénomène de phonation en termes de cibles et de trajectoires pour les différents paramètres acoustiques servant à contrôler le synthétiseur. À partir de cette description par cibles et trajectoires, on génère une succession discrète de trames de paramètres, qui serviront à contrôler le synthétiseur proprement dit[†]. Quand on va d'une étape à l'autre du processus de synthèse décrit ci-dessus, l'information va en se complexifiant en allant de gauche à droite. Néanmoins, chaque étape peut être considérée comme générative de l'étape la suivant à droite et la synthèse est faite de façon univoque à partir des règles.

Dans le cadre du projet LOQUAX, nous nous sommes proposés de développer une méthodologie systématique permettant d'élaborer des règles de production à partir d'un corpus de parole naturelle, dans le but de faire de la synthèse *ab initio* de très haute qualité.

2. Le synthétiseur utilisé

Le synthétiseur que nous avons utilisé est une variante de l'architecture proposée par Klatt [1]. La principale différence architecturale qui le distingue du synthétiseur originellement proposé par Klatt consiste en deux paires pôle-zéro supplémentaires. En effet des tests préliminaires tendent à démontrer que l'emploi d'une seule paire pôle-zéro (comme le propose Klatt) est en général insuffisant pour un modélisation fidèle du spectre des voyelles nasales[‡].

Un simulateur de cette architecture a été créé en langage PASCAL. Dans cette version, tous les paramètres contrôlant les filtres (fréquence centrale f_c et largeur de bande Δf) et les amplificateurs (gain externe en dB) sont externes, par opposition à la vingtaine de paramètres que Klatt propose de faire varier pendant que les autres restent fixes. Le nombre de paramètres externes est maintenant de 42: f_c et Δf pour 15 filtres du deuxième ordre, gain de 10 amplificateurs, valeur de préaccentuation représentant les caractéristiques de rayonnement et fréquence fondamentale F_0.

---

† Il est à noter que dans cette représentation simplifiée, on ne fait figurer ni le système de génération de la ligne prosodique, ni le système de règles permettant de faire la transcription orthographique-phonémique. On considère que les données sur lesquelles travaille le système de règles de production acoustique contient déjà ces informations.

‡ Le fait que Klatt ait d'abord développé cette architecture pour faire de la synthèse de l'anglais, langue ne présentant pas de voyelles portant intrinsèquement le trait phonétique [+nasal], n'est probablement pas étranger à cette déficience apparente.

Les 42 paramètres sont renouvelés a toutes les 10 ms (longueur de la trame) ce qui assure environ un renouvellement par période du fondamental, les voix étudiées ayant un F_0 de l'ordre d'une centaine de Hz. Les coefficients des filtres et les gains linéaires des amplificateurs (paramètres internes correspondant aux paramètres de contôle externes) sont interpolés à l'intérieur de la trame de façon à prévenir les « clics » et la distortion pouvant se produire lors d'un changement brusque des paramètres. Cette technique produit un signal de bien meilleure qualité qu'avec le renouvellement période-synchrone (« pitch synchronous ») préconisé par Klatt, et moyennant quelques astuces d'implantation, ne nécéssite qu'environ 30 à 40% de plus de temps de calcul que le renouvellement période-synchrone.

3. Analyse segmentale

L'architecture même du synthétiseur utilisé suppose implicitement une décomposition du signal de parole en un certain nombre d'éléments correspondant à autant de phénomènes physiologiques et acoustiques distincts, soit :

1– *partie voisée du signal :* elle-même décomposée en:
   a) *la source de voisement*
   b) *les caractéristiques conduit vocal*
   c) *les caractéristiques de rayonnement*

2– *la partie non-voisée du signal*
   a) *l'aspiration*
   b) *la frication*

Cette décomposition exige, en pratique, lors de l'analyse, l'intervention de l'expérimentateur pour associer les bons paramètres aux bons phénomènes acoustiques.

De plus, notre modèle du conduit vocal contient plusieurs paires pôle-zéro; or la présence de zéros dans la fonction de transfert rend impossible l'analyse linéaire du signal.

Pour ces raisons, une analyse automatique directe est virtuellement impossible pour une architecture de synthétiseur de la complexité de celle que nous avons retenue.

Devant ces difficultés nous avons décidé d'adopter une approche entièrement interactive pour l'analyse spectrale segmentale. L'analyse du signal de parole est faite en visualisant la fonction de transfert(calculée à partir des paramètres du synthétiseur) affichée sur un écran vidéo, *en même temps* que le spectre du signal naturel à analyser, obtenu par transformée de Fourier rapide (comme on le voit à la figure 2). Les paramètres du synthétiseur apparaissent simultanément à l'écran du terminal, en mode plein écran, et il est possible de les éditer pour visualiser immédiatement le changement apporté à la fonction de transfert du synthétiseur par le changement de paramètre. Cette méthode permet d'arriver assez rapidement à trouver un ensemble de paramètres qui colle bien au spectre du segment analysé. La parole naturelle doit avoir été préalablement découpée en segments spectralement stables ou en segments suffisamment courts pour être considérés stables.

La fonction de transfert est évaluée directement dans le domaine fréquentiel en calculant la valeur complexe de transformée en z du signal sur le cercle unité (sans donc qu'une convolution implicite avec le train d'impulsion du fondamental).

4. Description par trajectoires

L'essentiel de la démarche d'analyse supra-segmentale consiste à rechercher quelles trajectoires il faut donner à chacun des paramètres du synthétiseur pour reproduire le phénomène de phonation étudié. Cela finalement s'apparente beaucoup, dans le domaine temporel, à la démarche employée dans le domaine fréquentiel, à savoir, ajustement de paramètres (qui sont alors des descripteurs de trajectoire) jusqu'à l'obtention d'une bonne représentation du signal naturel. L'analyse automatique des trajectoires est assez difficile d'une part, et d'autre part on désire l'intervention intelligente de l'expérimentateur pour fixer des contraintes concernant le type de trajectoire pour chacun des paramètres, le groupement de paramètres qui doit être fait pour que des paramètres décrivant le même mouvement articulatoire bougent de concert etc.

Pour décrire les transitions des différents paramètres acoustiques (ayant leur équivalent dans le synthétiseur) lors de la phonation, on doit avoir un ensemble génératif de trajectoires possibles.

Trajectoire	symbole	degré	continuité			
			$f(t_1)$	$f(t_2)$	$f'(t_1)$	$f'(t_2)$
saut	=>	0	-	-	-	-
linéaire	- >	1	√	√	-	-
parab. droite	(>	2	√	√	-	√
parab. gauche	) >	2	√	√	√	-
biparabolique	~>	2	√	√	√	√
cubique	] >	3	√	√	√	√

Tableau 1 Trajectoires possibles des paramètres du synthétiseur avec continuité au début (t_1) et à la cible (t_2).

Ces trajectoires (tableau 1) constituent notre métamodèle de production, c'est-à-dire l'« alphabet » de base servant à décrire tout type de phonème. Elles sont toutes de type polynomial avec une continuité des dérivées successives aux limites de la trajectoire qui évidemment s'améliore quand on augmente le degré du polynome.

saut : Simple saut de la valeur courante à la cible. Les sauts sont utilisés lors de transitions très rapides, en particulier lors de l'explosion d'une consonne occlusive.

linéaire : Transition linéaire simple (rampe) depuis la valeur courante, jusqu'à la cible.

parabolique gauche : Trajectoire parabolique avec dérivée libre à droite (à la cible) et dérivée continue avec les valeurs précédentes à gauche.

parabolique droite : Trajectoire parabolique avec dérivée libre à gauche (à la valeur initiale) et dérivée nulle à droite (en arrivant à la cible). Ce type de trajectoire est particulièrement utile pour modéliser la relaxation des paramètres après l'explosion se produisant à la fin d'une occlusive.

biparabolique : Trajectoire constituée de deux segments de paraboles avec dérivée initiale continue (à gauche), dérivée continue à la jonction des segments (point d'inflexion), et dérivée nulle à droite (à la cible). Le moment auquel se produit l'inflexion à partir du départ est fixé par l'usager, ou par défaut, se produit au milieu de la durée de la transition. On peut démontrer la pertinence de ce genre de trajectoires pour les mouvements des formants lors de la coarticulation.

quadratique : Trajectoire du troisième degré avec dérivée continue à gauche et nulle à droite.

Pour pouvoir mettre en pratique notre métamodèle de production, nous avons dû créer un *langage de description de trajectoires acoustiques*, en abrégé, LDTA.Un interpréteur permettant de transformer un « programme » écrit en LDTA en une séquence continue de trames de paramètres a été implanté en langage PASCAL. Grâce au LDTA, il est possible de modifier la dynamique de n'importe quel paramètre ou groupe de paramètres du synthétiseur, à n'importe quelle trame sans être obligé de noter explicitement ni toutes les trames, ni tous les paramètres des trames considérées. Supposons, par exemple que l'on veut, à la trame no 14, faire partir F_2 (la fréquence centrale f_c du 2^e formant) de sa valeur courante et le porter à 1200 Hz, avec une trajectoire *biparabolique* d'une durée de 100 ms, avec le point d'inflexion à 30 ms du début de la transition, ceci en même temps que B2 (Δf du 2^e formant) montera vers 200 Hz avec une trajectoire linéaire de même durée. Cela s'écrira en LDTA:

```
14 : (F2~>1200 B2->200)\100\30  ;
```

Ce langage permet donc de décrire de façon à la fois claire et compacte les trajectoires des différents paramètres du synthétiseur. Le fichier en LDTA peut être structuré et commenté pour en faciliter l'interprétation et d'éventuelles modifications.

5. Exemple d'application : étude des voyelles nasales du français montréalais

Notre méthodologie est actuellement appliquée à l'étude des caractéristiques acoustiques des voyelles nasales (ɛ̃ , ɑ̃ , ɔ̃ et œ̃) du français montréalais.

Les voyelles nasales du français montréalais se distinguent tout d'abord de celles du français parisien par leur nombre effectif. En effet, en français parisien, l'opposition œ̃ ↔ ɛ̃ est aujourd'hui neutralisée, ce qui n'est pas du tout le cas en français montréalais. De plus, en français montréalais, /ɛ̃/ se réalise acoustiquement par [ẽ], et ɑ̃ se réalise par [ã]. Enfin, les voyelles nasales du français montréalais se caractérisent par un *retard de nasalité* qui peut aller jusqu'à la diphtonguaison ([3]), par exemple /ɛ̃/ peut souvent se réaliser [ɛẽ]. Ce phénomène de retard de nasalité est illustré par le spectrogramme de la figure (1) où on voit bien la dynamique des formants.

Pour illustrer la technique d'analyse employée, on donnera ici l'exemple de l'analyse du phonème /ɛ̃/ réalisé dans le contexte du logatome [agɛ̃pa] (figure 1).

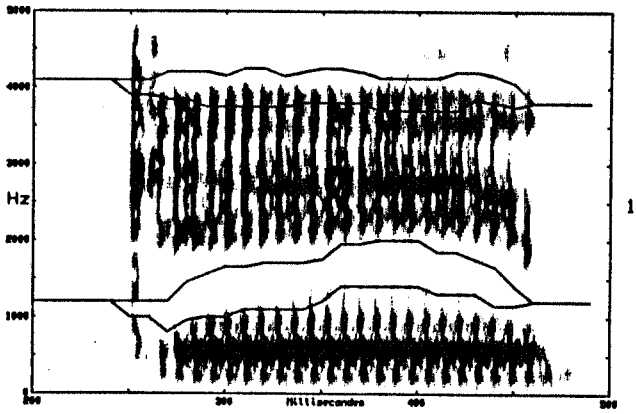


Fig. 1 Spectrogramme avec localisation des paires pôle-zéro pour /ɛ̃/ dans le logatome [agɛ̃pa]

Dans ce cas, on a procédé à une analyse fine, trame à trame, du signal. À la figure (2) on a un exemple de résultat d'analyse fine.

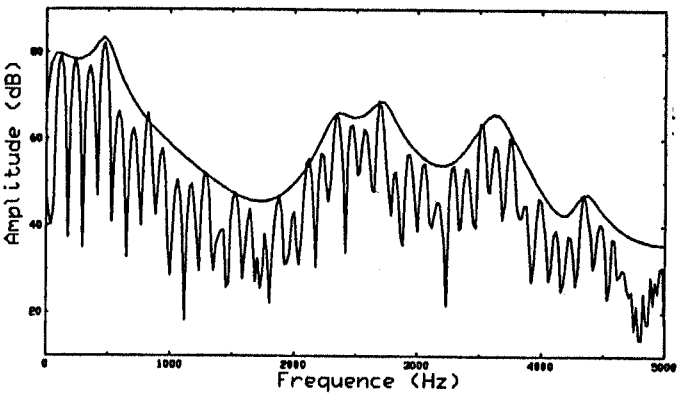


Fig. 2 Analyse d'un segment de [ẽ]

La trame analysée se situe à un moment où la nasalisation est maximum; le spectre du signal est approché en utilisant pour la fonction de transfert du conduit vocal, les

5 résonnateurs formantiques habituels, plus 3 paires pôle-zéro. L'effet de ces paires (se manifestant peu ou pas dans le cas des voyelles orales) est de créer une dénivellation spectrale en faveur des basses fréquences. Typiquement le zéro de la paire se manifeste à une fréquence plus élevée et avec une largeur de bande plus petite que le pôle. Par sa largeur de bande plus étroite, le zéro domine la plupart du temps, comme c'est le cas pour les paires du [ẽ] considéré. À la figure (1), on voit, superposées au spectrogramme, les trajectoires des paires, chacune d'elles se divisant après l'explosion du [g], le zéro passant au-dessus du pôle. Les largeurs de bande varient énormément. La différence de largeur de bande entre le pôle et le zéro est déterminante sur l'importance de leur manifestation. Chose qui est impossible à voir sur le spectrogramme, les paires augmentent considérablement leur importance spectrale, environ 80 ms après l'explosion, au moment même où F_2 semble entreprendre une deuxième ascension.

La distance subjective entre le signal naturel et la synthèse réalisée à partir des résultats bruts (sans lissage) peut être pratiquement réduite à néant, ceci à condition de faire une analyse soigneuse trame à trame. La réduction à une représentation LDTA peut être faite en ne sacrifiant que très peu de naturel.

6. Conclusions

La méthodologie que nous avons développée permet de mettre en évidence des caractéristiques acoustiques difficiles à saisir par d'autres approches. De plus la réduction de la parole naturelle à une description par cibles et trajectoires (utilisant le LDTA) permet de réaliser un nombre virtuellement illimité d'expériences psychoacoustiques pour évaluer l'importance perceptuelle des caractéristiques acoustiques observées. Ces expériences devraient mener à l'élaboration d'un modèle de production plus complet qui pourra augmenter considérablement la qualité de la synthèse par règles.

De plus, cette méthodologie reste très générale et pourrait être appliquée à d'autres architectures de synthétiseur et même servir à étudier d'autres classes de phénomènes acoustiques que la voix parlée (voix chantée, instruments acoustiques etc.).

Dans le cadre de notre étude, nous avons pu mettre en évidence les lacunes du synthétiseur de Klatt pour la synthèse du français et apporter les correctifs nécessaires.

Les premiers résultats de l'étude des voyelles nasales mettent en évidence un fait assez intéressant : le « formant nasal » devant apparaitre à une fréquence inférieure à F_1 et que l'on considère comme une caractéristique importante de la nasalité ([4]), n'est pas universellement présent chez tous les locuteurs. Non seulement il est invisible sur le spectrogramme, mais souvent, même une analyse fine ne parvient pas à mettre en évidence de perturbation spectrale ponctuelle notable sous F_1, et ceci chez des sujets ne présentant aucun problème dans la production des voyelles nasales.

Il semble que dans la plupart des cas, la paire pôle-zéro la plus importante apparaisse en fréquence intermédiaire, c'est-à-dire (dépendemment de la voyelle et du locuteur) quelque part entre 1000 et 2200 Hz, ceci sans créer forcément l'apparition « pseudo-formant » supplémentaire, mais en créant un biais spectral important entre les fréquences inférieures et supérieures à cette paire. Une généralisation hâtive est fort hasardeuse, car à la lumière de quelques expériences préliminaires, il semble que de nombreux traits acoustiques concourent au trait phonologique [+nasal]. Il y a bien sûr des traits de caractère statique, à savoir les cibles pour les fréquences et largeurs de bande des résonateurs et des antirésonnateurs. Mais il y a aussi des caractéristiques importantes liées à la dynamique de phonation. En français montréalais, cette dernière constatation est possiblement compliquée par le fait que les 4 voyelles nasales portent toutes le trait [+long] en plus du trait [+nasal]. Il reste à évaluer de façon plus systématique l'importance perceptuelle et la variabilité inter-locuteur des traits observés, mais on peut prévoir que pour atteindre un plus grand *naturel* en synthèse, il faudra réunir plusieurs de ces traits, alors que la simple *intelligibilité* n'en eût pas demandé autant.

La poursuite de nos objectifs de recherche devrait mener à court terme à une meilleure compréhension des caractéristiques acoustiques de la nasalité en général, et dans le cas du français montréalais en particulier.

BIBLIOGRAPHIE

[1] Dennis H. Klatt, « Software for cascade/parallel formant synthesiser », *J. Acoust. Soc. Amer.*, vol. 67(1), pp. 971-995, Mars 1980.

[2] D. O'Shaughnessy, « Design of a real-time French text-to-speech system », *Speech Com.(North-Holland)*, vol. 3, pp. 233-243, 1984.

[3] J.D. Gendron, *Tendances phonétiques du français parlé au Canada*, Presses de l'Université Laval, Québec, 1966.

[4] S. Maeda, « The role of the sinus cavities in the production of nasal vowels », *Int. Conf. Acoust., Speech and Signal Proc.*, 1982, pp. 911-914.

ETUDE DE LA GEMINATION DES OCCLUSIVES EN ARABE

A. Rajouani, M. Najim, A. Mouradi, D. Chiadmi & M. Ouadou

L.E.E.S.A, Faculté des Sciences, BP. 1014, Rabat,MOROCCO.

ABSTRACT

The purpose of this paper is to study duration of closure as a cue factor in the discrimination of single versus geminate stop consonants in Arabic. Emphasis is made on the correlation between the closure duration and the following parameters :
- nature of the stop consonant : voiced/unvoiced, pharyngealized/non pharyngealized.
- spectral structure of the stop consonant,
- lenght of the adjacent vowels (short/long).

INTRODUCTION

La gemination est un processus fréquent dans le système de dérivation de l'Arabe. Par ailleurs l'opposition simple/geminée joue un rôle majeur pour la differenciation morphologique et semantique . Le signe diacritique (chadda) placé au dessus de la consonne dans un texte orthographique pour marquer la gemination est transcrit phonétiquement par un dedoublement de cette consonne. En Arabe, toute consonne non initiale est suceptible d'être geminée. Etant souvent considerée sur le plan phonologique comme une succession de deux consonnes identiques simples la gemination soulève quelques problèmes au niveau de la syllabisation. Dans ce travail nous nous sommes particulièrement interessés aux aspects acoustiques et phonétiques.

POSITION DU PROBLEME

La représentation spectrographique montre que l'opposition simple/géminée pour une consonne non occlusive est réduite essentiellement à l'opposition temporelle courte/longue. Les tests de synthèse permettent de valider cette remarque. En effet, par la seule variation de la durée de la partie stable, on synthètise avec le même ensemble de paramètres de contrôle la consonne non occlusive simple et la geminée correspondante. Une gemination naturelle,est obtenue avec les facteurs multiplictifs de durée suivants : 1.7 pour les fricatives sourdes et les nasales, 1.6 pour les fricatives sonores , les semi-voyelles et les liquides.

Des travaux antécedents /1/, /2/ decrivent à partir d'observations spectrographiques l'occlusive geminée comme une occlusive simple réalisée avec une plus grande durée de silence de fermeture (SF). Une approche par synthèse realisée par O'Brecht/3/ pour l'occlusive /b/ confirme ce résultat. Dans ce travail, nous présentons une étude quantitative du phénomène en essayant de détecter d'éventuels effects des facteurs suivants : le voisement, la pharyngalisation, les propriétés spectrales de l'occlusion, la nature quantitative (longue/courte) des voyelles adjacentes.

PROCEDURE EXPERIMENTALE ET RESULTATS

Synthétiseur et stratégie de synthèse

Nous avons utilisé un synthétiseur à formants série/parallèle (synthétiseur de Klatt) que nous avons développé dans le cadre d'un projet pour la synthèse par règle de l'Arabe. La structure adoptée est similaire à la version présentée par O'Shaughnessey /4/. Chaque phonème (ou segment) est représenté dans la table d'entrée par 25 paramètres relatifs aux caractéristiques temporelles, spectrales, et phonétiques (figure 1.). Les trajectoires des paramètres de contrôle sont calculées automatiquement par des règles de transition en fonction des traits phonétiques et phonologiques des segments de la séquence.

La fréquence d'échantillonage des paramètres de contrôle est 100 Hz. Une occlusive est représentée dans la table d'entrée par le segment correspondant a 10 ms de la plosion. Les paramètres relatifs à la zone de transition occlusive-segment adjacent sont calculés par règles alors que les paramètres correspondant au silence de fermeture sont automatiquement déduits à partir des paramètres de la plosion. Seuls les paramètres suivants changent :
- les paramètres 1,2 et 3,
- les paramètres 22,23,24,25 sont annulés pour le SF,
- le paramètre 21 est mis à 1 (1=SF).

paramètres n°

1,2,3	: durée de la partie stable, durées des transitions avec les segments adjacents
4	: fréquence fondamentale
5,...,9	: valeurs des formants F1,..,F5

10,11,12	: valeurs des bandes passantes des 3 premiers formants
13	: fréquence du resonateur nasal
14,..,18	: gains des resonateurs parallèles
19	: amplificateur des hautes fréquences
20	: indicateur de l'accent
21	: indicateur de la classe phonétique
22,..,25	: amplificateurs des sources d'excitation.

Fig.1. Liste des paramètres

Matériau phonétique

L'ensemble des 8 occlusives de l'Arabe est entièrement décrit par les traits : voisé/non-voisé, pharyngalisé/non pharyngalisé. Afin de pouvoir examiner à fond les effets de chaque trait nous avons limité cette étude à 3 occlusives :/b/ voisée, non pharyngalisée; /t/ non voisée, non pharyngalisée; /ṭ/ non voisée, pharyngalisée.

Guidés par le souci de rendre les comparaisons plus exactes nous n'avons introduit qu'un seul timbre vocalique (a) avec ses deux variantes quantitatives longue (180 ms)/courte (90 ms). Les variantes pharyngalisées sont utilisées dans les séquences comprenant /ṭ/. Par ailleurs la consonne geminée étant toujours en position intervocalique, les logatomes utilisés sont de type sVCVsa. La durée du silence de fermeture varie de 40 ms à 200ms avec un pas de 10 ms (17 occurences). Des investigations preliminaires ont montré qu'au niveau des paramètres spectraux, seules les bandes passantes des formants sont suceptibles d'être modifiées sans que la qualité des occlusives synthétisées ne soit profondément atteinte. Dans cette étude on s'est intéressé à l'étude des effets du rapport simple (B.P)/moitié (B.P/2) sur la perception et la qualité de la gemination. Afin d'étudier l'effet de la quantité vocalique les 4 combinaisons possibles sont considérées : /saCasa/,/saaCasa/,/saCaasa/, /saaCaasa/, où aa indique la variante longue. Le corpus est ainsi constitué de 408 occurences synthétisées, chaque occlusive étudiée étant représentée dans 136 occurences (17x2x4). Les 408 séquences sont reparties sur 24 listes, 8 listes pour chaque consonne étudiée. Une liste contient 17 occurences correspondant aux 17 durées du SF (40 ms à 200 ms, pas=10ms) considérées.Pour la clarté de l'exposé on adoptera la numérotation suivante :

liste	contexte vocalique	valeur des bandes passantes.
1	sabasa	BP
2	sabasa	BP/2
3	sabaasa	BP
4	sabaasa	BP/2
5	saabasa	BP
6	saabasa	BP/2
7	saabaasa	BP
8	saabaasa	BP/2

Les listes 9 à 16 (resp. 17 à 24) sont dans le même ordre relatives à /t/ (resp. à /ṭ/).

Expérience

Les 17 occurences de chaque liste sont écoutées dans un ordre aléatoire et separément par chacun des 4 sujets qui ont participé à l'expérience. Chaque occurence peut être écoutée deux fois de suite. Toute occurence écoutée et classée ne peut être reécoutée par le même sujet. La durée de pause entre deux occurences est laissée au choix du sujet. Pour chaque occurence le sujet a le choix entre trois réponses :"géminée", "non géminée", "je n'arrive pas à juger".

Résultats

Nous considérons que pour une application de type synthèse, il est préférable d'éviter dans la mesure du possible toute approche statistique. Par ailleurs on considère qu'une occurence est géminée lorsqu'elle est perçue comme telle par les 4 sujets , de même pour la non géminée. Les résultats sont présentés dans la figure 2. Ts et Tg sont tels que (40-Ts) est l'intervalle de SF correspondant à la consonne simple et (Tg-200) correspondant a son opposée géminée.

Liste	Ts	Tg	liste	Ts	Tg	liste	Ts	Tg
1	60	140	9	70	140	17	70	120
2	60	140	10	80	150	18	70	130
3	50	130	11	70	130	19	60	120
4	70	130	12	70	130	20	70	110
5	60	150	13	50	150	21	60	130
6	80	110	14	80	140	22	60	110
7	60	150	15	50	150	23	80	130
8	80	120	16	70	140	24	80	120

Remarques

- Les valeurs de Tg pour les listes correspondant a BP sont (pour un même contexte) identiques pour les deux occlusives non pharyngalisées /b/ et /t/. La durée du silence de fermeture necessaire pour la gémination serait indépendante de l'opposition sourde/sonore.
- Dans un même contexte, les valeurs de Tg relatives à /ṭ/ sont inférieures à celles de /t/, la pharyngalisation tend à diminuer le seuil de la gemination pour l'occlusive sourde.
- L'introduction d'un facteur multiplicatif 0,5 pour les bandes passantes tend à augmenter Ts. Le fait est plus notable pour les consonnes non pharyngalisées /b/, /t/.
- Pour la consonne voisée /b/, l'introduction du facteur multiplicatif ne peut , dans un même contexte vocalique, que diminuer Tg. Pour les consonnes sourdes /t/ et /ṭ/ cet effet n'est pas evident.
- D'après les listes 1,3,5,7 (resp. 9,11,13,15) (resp. 17,19,21,23) on remarque que le contexte vocalique /aCaa/ tend à diminuer Tg alors que les contextes /aaCa/ et /aaCaa/ tendent à l'augmenter. Cependant cet effet n'est pas évident pour les listes correspondant à BP/2.
- Dans les contextes /saaCaasa/ et /saaCaasa/les sujets ont souvent noté que la voyelle /aa/ post-consonantique supposée longue tend à être

confondue avec son opposée courte lorsque la conne C est gêminée.
- Il n'existe pas de rapport évident entre Ts et Tg.
- Nous avons repris l'expérience sur la paire xabara/xabbara étudiée par O'Brecht /3/. Pour une durée de SF égale à 140 ms, les 4 sujets marocains jugent à 100% que /b/ est géminée reconnaissant /xabbara/ alors que O'Brecht rapporte pour la même durée un score de 15%. O'Brecht remarquait que "the data were gathered from subjects who were both linguistically and experimentally naive, and who were in no sense trained observers other than of the phonemic and morphemic contrasts of their language". Cependant la non concordance des résultats des deux expériences serait dûe en partie aux effets des substrats dialectaux, les sujets pour l'experience de O'Brecht étant originaires du Moyen Orient.

EVALUATION

Dans quelle mesure, les résultats obtenus sont ils exploitables pour le développement d'une synthèse par règle de l'Arabe ? Au stade actuel, des tests intensifs sur des mots réels ne contenant que le timbre /a/ ont permis de valider les aspects suivants :

- les valeurs de Tg déterminées sont indépendantes de la position, dans le mot, de la syllabe contenant l'occlusive;
- les valeurs déterminées pour /t/ (resp./b/) sont parfaitement valables pour la synthèse de la gemination de /k/ et /q/ (resp./d/);
- les valeurs de Tg utilisées pour la consonne voisée pharyngalisée /ḍ/ sont déduites à partir de celles déterminées pour /b/ sous l'hypothèse que la pharyngalisation diminue le seuil de la gemination indépendement de la nature sourde/sonore de l'occlusive. La bonne qualité de la gemination synthétisée atteste la validité de l'hypothèse.
- les valeurs de Tg obtenues pour /b/ ne permettent pas la synthèse de la gémination de l'occlusive voisée /ʔ/. Cela pourrait être dû à la nature glottale de /ʔ/ qui tend à se confondre avec la voyelle. Les valeurs de Tg pour /ʔ/ sont supérieures de 10 ms à celles de /b/ quoique la représentation spectrographique montre systématiquement que le silence de fermeture de l'occlusive simple /ʔ/ est plus court que celui de /b/. Ce fait confirme l'absence d'un rapport entre la durée du SF pour l'occlusive simple et la valeur correspondant à son opposée geminée.

PERSPECTIVES

Des tests intensifs sont menés afin de définir les effets probables des voyelles /u/ et /i/. L'influence des facteurs prosodiques, essentiellement le rythme, sera étudiée dans une étape ultérieure.

REFERENCES

(1) S. Al Ani, Arabic Phonology, Mouton, The Hague, 1970.
(2) H. Yoshioka, A. Lofqvist et H.Hirose, "Laryngeal adjustments in the production of consonant clusters and geminates in American English", J. Acoust. Soc. Am. vol-70,n°6,1981,pp.1615-1623.
(3) D.H. O'Brecht,"Three experiments in the perception of geminate consonants in Arabic", Language and Speech, vol-8,part I, 1965, pp. 31-41.
(4) D.O'Shaughnessy, "Design of a real-time French text-to-speech system", Speech Communication, vol-3,n°3, 1984,pp. 233-244.

L'EVALUATION DE LA QUALITE DE L'ALLEMAND SYNTHETISE PAR DIPHONES

Betina Schnabel

Institut de la Communication Parlée
Institut de Phonétique
Grenoble

ABSTRACT

A large number of studies have been concerned with the determination of speech quality in synthesis, but still no generally accepted methods of evaluating speech intelligibility and quality have been developed so far.

In a first approach we confined ourselves to the quantification of speech intelligibility only and we developed tests to measure auditive perception based on rhyme tests [1]. These evaluation tests enabled us to obtain concrete percentages of correct perception with regard to the identification of phonemes, and to apply modifications to the dictionary of dyads for the purpose of eliminating errors due to segmentation and concatenation [2].

INTRODUCTION

De nombreux travaux ont été consacrés à la détermination de la qualité de la parole codée, mais il n'existe pas de tests d'évaluation universellement reconnus pour la synthèse. Afin de juger dans un premier temps de l'intelligibilité de la parole synthétique nous avons conduit nos expériences perceptives, en nous fondant sur la méthode de test de rime [1].

Les tests ont été établis : d'une part pour obtenir des taux précis sur l'intelligibilité de la synthèse au niveau de l'identification des phonèmes; d'autre part, pour effectuer des modifications dans le dictionnaire de diphones, et corriger les fautes de segmentation et de concaténation [2].

1. ELABORATION DES TESTS DE RIME

Cette démarche nous a conduit à élaborer des tests de rime d'après la méthode du "Modified Rhyme Test" (MRT) [3]. Le MRT se compose de six mots naturels monosyllabiques qui forment deux à deux des paires minimales. La tâche de l'auditeur consiste à sélectionner l'élément prononcé parmi l'ensemble de choix possibles, comme dans l'exemple de l'unité :

wir vier Bier Pier Tier mir

L'avantage du MRT réside dans la facilité de son application et dans la rapidité de l'évaluation des résultats :

- le test est constitué de mots naturels pris dans un corpus phonétiquement équilibré et représentatif pour la langue donnée;
- aucun entraînement particulier n'est demandé aux auditeurs pour les tests;
- les résultats des tests peuvent être doublement exploités sous forme de courbes de pourcentages concernant la reconnaissance des sons, et sous forme de matrices de confusion pour les phonèmes;
- et finalement, l'application étendue du MRT permet de comparer les résultats tirés des tests à d'autres systèmes de synthèse [4, 5].

L'analyse des résultats des tests permet non seulement de définir les traits distinctifs les plus sensibles au codage, mais aussi de déterminer les facteurs favorisant certains défauts de synthèse lorsque la référence n'a pas été correctement perçue.

2. LES AUDITEURS ET LES CONDITIONS DES TESTS

A partir d'un corpus de tests de rime modifié, existant déjà à Berlin [4], nous avons établi 4 séries de tests, constituées chacune de 55 lignes de 6 mots. Chaque série de tests consiste en un ensemble de 22 mots tests pour les noyaux vocaliques, un ensemble de 18 mots tests pour les consonnes initiales, et un ensemble de 15 mots pour les consonnes finales.

Les 11 auditeurs qui ont passé les tests étaient tous des auditeurs "naifs", dans la mesure où aucun d'entre eux ne travaille dans le domaine de la phonétique ou de la linguistique. Naturellement, l'allemand est la langue maternelle de tous ces auditeurs.

Les tests ont été appliqués à des petits groupes de 2 à 4 sujets dans un endroit calme, mais pas dans une chambre sourde. Chacun des 11 auditeurs est venu 4 fois avec un décalage de 1 à 3 jours entre les différentes séances. Avant la fin de toutes les séries de tests, aucun renseignement sur la qualité de la synthèse n'était donné aux auditeurs.

3. LES RESULTATS DES TESTS

Le taux de reconnaissance des sons pendant toute la série de tests atteint en moyenne 89,2 %, ce qui est comparable aux taux correspondants à la parole transmise dans la bande téléphonique (300-3400 Hz).

Les figures 1 à 3 montrent qu'en moyenne nous avons obtenu un taux d'identification des sons de
90,3 % pour les noyaux vocaliques,
89,0 % pour les consonnes initiales,
88,2 % pour les consonnes finales.

En regardant les courbes de pourcentage de plus près, il est évident que la plus faible reconnaissance se trouve pour le /h/- initial, qui n'obtient que 30 % en moyenne. La mauvaise identification du /h/- initial est due à une faute de segmentation : dans un contexte intervocalique la consonne sourde peut être transformée en une fricative sonore. La concaténation de /silence + h/ et /h + V/ sonorisé rend la combinaison inintelligible. En supprimant le voisement du /h/ la faute de segmentation a pu être éliminée. De la même manière, la faible reconnaissance de /e/-long et de /OE/-bref a été améliorée, car elle était produite par le /h/- initial dans le contexte des mots test, comme dans les exemples : Heer, heb, höre, Höhle.

Les taux d'intelligibilité de /M/- syllabique, environ 70 %, et de /ng/- final, environ 30 % seulement, s'expliquent par la perception des consonnes nasales en général. La distinction entre les consonnes nasales finales constitue déjà un problème de perception dans la langue naturelle. En synthèse, et sans contexte, ces consonnes amènent certainement des confusions avec d'autres consonnes nasales.

L'amélioration de certains points précis, comme l'élimination des fautes de segmentation de /h/- initial, /e/- long et /OE/- bref, devrait nous permettre d'atteindre rapidement des taux d'intelligibilité comparables à ceux obtenus grâce à un synthétiseur de haute qualité (taux compris entre 95 % et 98 %).

REFERENCES

[1] W.P. VOIERS, Diagnostic Evaluation of Intelligibility in Present-Day Vocoders. IEEE. Trans. Audio. Electr., Au-16, 275-279, 1972.

[2] B. SCHNABEL, La synthèse de l'allemand : segmentation des diphones et évaluation de la parole codée. D.E.A. Institut de la Communication Parlée, Grenoble, 1985.

[3] A.S. HOUSE, C.E. WILLIAMS, M.H.L. HECKER & K.D. KRYTER, Articulation-Testing Methods : Consonantal Differentiation with a Closed-Response Set. J. Acous. Soc. Am. 37, 158-166, 1965.

[4] J. SOTSCHECK, Ein Reimtest für Verstandlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachgüteübertragung. Der Fernmeldeingenieur 36, 1-84, 1982.

[5] H. KAESLIN, Systematische Gewinnung und Verkettung von Diphonelementen für die Synthese deutscher Standartsprache. Thèse de Docteur Ingénieur, ETH Zurich.

(Ce travail a été réalisé dans le cadre du contrat SYNTALIT (synthèse multilangue) financé par l'ADI et le CNET Lannion).

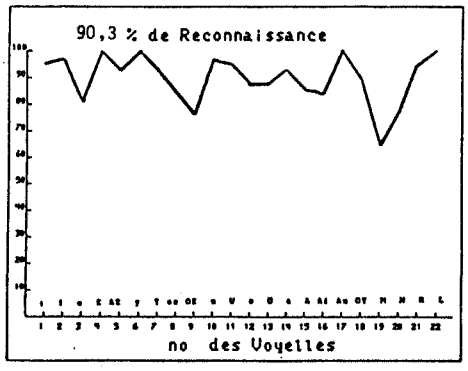


figure 1

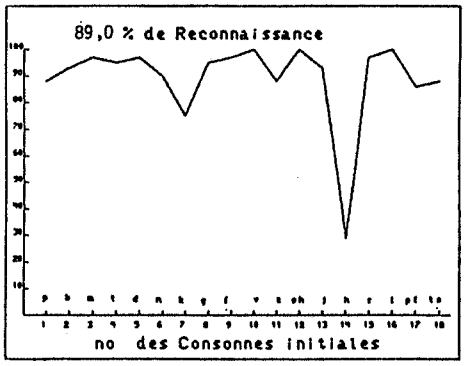


figure 2

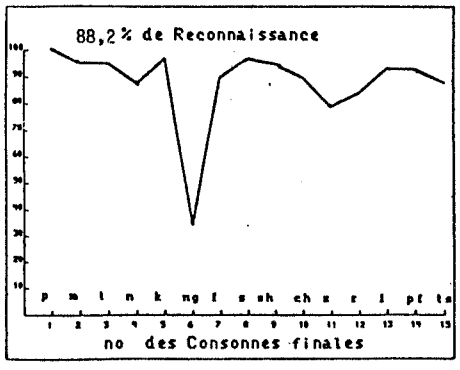


figure 3

- Figures 1 à 3 -

Les pourcentages de reconnaissance des voyelles, des consonnes initiales et des consonnes finales (moyenne totale de reconnaissance : 89,2 %).

PHYSIOLOGIE

Président

Denis AUTESSERRE

Institut de Phonétique d'Aix

A PILOT STUDY OF THE ARTICULATION OF /n/ IN ITALIAN USING ELECTROPALATOGRAPHY AND AIRFLOW MEASUREMENTS

E. Farnetani

Centro di Studio per le Ricerche di Fonetica del C.N.R., Padova

ABSTRACT

The production of the dentoalveolar nasal consonant /n/ was studied in one Italian subject by simultaneous application of palatography and airflow measurements. The scope of the research was to infer the dynamics of lingual movements (from patterns of linguopalatal contact) and of velar movements (from the course of egressive nasal airflow) and to study the spatiotemporal changes due to the context. The data indicate that: a) the timing of velar opening/closing gesture depends systematically on the context preceding or following the nasal, but in similar context carryover velar coarticulation has greater temporal extent than anticipatory coarticulation; b) tongue body movements are affected by adjacent vowels and consonants, tongue-tip movements only by following consonants.

INTRODUCTION

The present paper reports the results of a study of the dentoalveolar nasal stop /n/ in various contexts by means of elecropalatography and measurements of egressive oral and nasal airflow. The simultaneous application of the two techniques makes it possible to visualize patterns of linguopalatal contact as well as airflow patterns and to infer the temporal coordination and the spatial characteristics of the movements of two different articulators involved in /n/ production, the tongue and the velum.

As far as velar articulation is concerned, the determination of how velar position and velar gestures are controlled in speech is of great importance for speech production theory. One limitation of the aerodynamic technique is that it can record the timing of the velar port opening and closing, the amplitude of the velar movement when the velar port is open, but cannot record velar height nor the onset/offset of velar movements when the velar port is closed. Its validity rests, however, on the fact that it can reliably measure the temporal extent of nasality in speech and compare its anticipatory and carryover effects. Moreover, cross-language comparisons are possible not only with data obtained with the same technique (Benguerel [1]), but also with data obtained with the "nasograph" (Ohala [5], Clumeck [3]).

As for lingual articulation, the recording of the linguopalatal contact pattern in synchrony with the airflow makes it possible to assess how lingual and velar movements are coordinated in the production of /n/ , to infer if changes in egressive nasal flow are due to velar movements or are side effects of lingual movements and finally it makes it possible to examine the context effects on tongue dynamics.

In the present research the following were studied: i) the temporal course of lingual and velar movements during the production of word-initial and word-medial /n/ in the context of /a/;

ii) the contextual effects on the pattern of lingoupalatal contact, on the onset/offset time of egressive nasal flow and on the maximum value of nasal flow volume velocity;

iii) the temporal extent of anticipatory and carryover velar coarticulation;

METHOD

Speech material consisted of words containing the /n/ phoneme in different position and different environment, e.g. nVCV, VnV, VnnV, VnCV,CVVn(C)V, where V stands for /a/ or /i/ and C following the nasal stands for any dentoalveolar or mediopalatal consonant or glide permissible in Italian. One subject, the author, repeated the list of words for three times, with a small pause between each other word.

The electropalatographic system is the RION Electropalatograph. An ample illustration of the

system and of the processing of the acoustic and of the EPG signal is in Farnetani et al. [4]. An example of the computer print-outs is shown in Fig. 1 (bottom). For recording the airflow volume velocity we used the Frokjaer-Jensen Electroaerometer, calibrated against a rotameter. The Aerometer is provided with two facial rubber masks, one for the oral and the other for the nasal flow. A microphone is inserted in the oral mask so that the oral audio signal can be recorded with the oral flow.

The oral audio signal, the egressive oral and nasal airflow volume velocity signals were fed into a multichannel FM recorder,together with the microphone signal from the palatograph. This procedure allowed us to synchronize the airflow and the EPG traces. The analysis of the course of the oral and nasal airflow was made on mingograms (paper speed:100 mm/s), where also the audio signal was displayed. The data for the item /an'na/ are illustrated in Fig. 1. From top to bottom are: the audio signal from the mouth microphone, the oral flow course,the nasal flow course, the computer print out of the audio signal and of the temporal evolution of the linguopalatal contact, a few selected EPG frames (bottom). As shown on the frames, the surface of the artificial palate has been divided into two zones, corresponding to the alveolar area (upper portion) and to the prepalatal/palatal area (lower portion): this allowed to trace separately the contact pattern of the tongue tip and of the tongue body. The present data refer to the second reading of the speech material.

Results

ARTICULATION OF WORD-INITIAL AND WORD-MEDIAL /n/ IN THE CONTEXT OF /a/

In word-initial position (after pause) the consonant is characterized by longer duration of oral closure, more extensive tongue body contact and higher peak of nasal flow.

As for the temporal course of the nasal flow, in word-initial nasals the airflow starts to rise from the baseline after closure has been completed; the onset of periodic signal occurs 10-15 ms later. This would suggest a rather high velar position at the beginning of an utterance, even when it starts with a nasal sound, in agreement with Benguerel et al. [2], and that the opening of the velopharyngeal port after oral occlusion seems to be responsible of initiation of vocal fold vibration, as revealed by the onset time of the periodic signal. In intervocalic nasals instead, onset of nasal flow occurs before oral implosion and its timing (the temporal interval between onset of flow and the first palatographic frame indicating attainment of complete closure) depends on the context that precedes the nasal, but what is common in all items is a sudden increase of nasal flow at the implosion moment, which coincides with a sharp drop of oral flow This has to be interpreted as the effect of the closing gesture of the oral structures.

During the closure the nasal flow course reflects quite well the velar movements - provided that no changes occur at the glottal level -. In both initial and intervocalic /n/ the nasal flow reaches its maximum at the release, suggesting that the velum continues lowering for the whole interval of oral closure. The sudden drop of nasal flow at the release of occlusion, timed with a sudden decrease of the tongue tip contact and with a sharp rise of the oral flow has to be interpreted as a consequence of the opening of the oral structures rather than the effect of the rising gesture of the velum; this is also shown by the course of nasal flow after the release: it reaches a nearly steady state when also the tongue tip lowering gesture has been concluded. These strong effects of the movements of the oral articulators on the nasal flow do not allow to assess with precision at what point in time the velum starts rising; however it is reasonable to suppose that the velum rising movement starts in most cases at the release of occlusion: the contribution of the velar component to the drop of the nasal flow at the release is revealed by the values of the nasal flow volume velocity: after the oral release, the nasal flow reaches the low value it had before the occlusion: if the velum had not started its rising movement in concomitance with oral release, the flow values would have been greater during the following than during the preceding vowel, since during the phase of oral occlusion there has been a substantial increase of nasal flow due to the continuation of the velar lowering gesture. During the vowel following a nasal, the nasal flow never reaches the baseline before the acoustic offset of the vowel, if it is word final, most often the nasal flow continues after the end of the vowel. Similar extensive carryover effects are observable in vowels following word-initial nasals: in these sequences the nasal flow reaches the baseline only at the onset of the following oral consonant.

EFFECTS OF VOWEL QUALITY ON /n/ ARTICULATION

As for the effects of vowel quality on the tongue movements, the trends shown in the present study are very similar to previous results on VtV sequences produced by the same subject (Farnetani et al.[4]): the closure tends to be longer and the tongue body elevation tends to be greater in /i/ than in /a/ context, while the

tongue-tip contact during the closure does not differ systematically as a function of the vowel; what seems interesting from the point of view of a speech production theory, is that also during intervocalic geminated nasals, whose duration is at least twice the duration of a single /n/, the effects of the vocalic context on the tongue body elevation persist, suggesting that the tongue-body configuration in the production of dentoaveolar stops depends mostly on the quality of the adjacent vowels even when there is a sufficient amount of time to reach a specific - :context free - position and confirming the results on /t/ production, e.g. that during the tongue tip/blade occlusion, the tongue body performs a vowel-to-vowel gesture.

As for the effects of the vowel context on the airflow pattern, the data indicate that the vowel has systematic effects on the onset time of the nasal flow: in all productions the airflow begins to increase later in the context of /i/ than in /a/ context (mean difference =73 ms).

As for the peak value of the nasal flow no systematic differences were observed as a function of the vowel, while remarkable differences were observed between single and geminated nasals: peak nasal flow is always higher in geminated than in single nasals, indicating that the velum continues to lower until the end of the oral occlusion whatever is its duration.

During the post-consonantal vowel (be it /a/ or /i/), the nasal flow never reaches the baseline, and tends to have higher peak values during /i/ than during /a/. This can be explained as an effect of the different degree of mouth opening for the two vowels: a closer configuration of the mouth can in fact induce a greater amount of air to escape through the nose even if no variations occur in the size of the velopharyngeal port.

EFFECTS OF THE FOLLOWING CONSONANT ON THE ARTICULATION OF /n/

In clusters with stops or affricates (e.g. in /'anda/, /'anta/, /'antsa/, /'antʃa/), the tongue always anticipates the position of the following stop (or the following closure in affricates) already at the beginning of the oral occlusion for the nasal, so that the whole cluster,or the whole portion of the cluster characterized by complete closure is produced as one articulatory closing and opening gesture. See Fig. 2, which compares the patterns of linguopalatal contact in /'anna/, /'anta/, /'antʃa/, /'anja/. The figure shows that in items b) and c), already during the first portion of the oral closure), both tongue tip and tongue body have taken the configuration of the following stop closure. Item c) is a clear instance of anticipatory palatalization of the nasal, a contextual effect which is different, in terms of articulatory strategies, from the vowel-to-vowel(or vowel-to-glide) tongue-body diphthongal gestures observed in items a) and d).

As far as the nasal airflow, the stop consonants following the nasal are the best environment for the study of the velum rising movement, since during complete oral closure the nasal flow course reflects variations in the size of the velopharyngeal port, if no changes occur at the glottal level. In all our instances of /nt/, /nts/, /ntʃ/ clusters, devoicing never occurs before the closure of the velopharyngeal port but occurs in concomitance or immediately after it, and could be a consequence of the closure, owing to the equalization of the subglottal and supraglottal airpressure (Rothenberg [6]).

Consequently in all /n/+stop clusters it is possible to detect from the nasal airflow pattern at what point in time after implosion the velum starts rising and to determine the duration of its closing movement. The data indicate that the rising movement of the velum tends to occur as late as possible, is a relative fast gesture and and is timed differently, as a function of the articulatory/perceptual requirements of the the following consonant: /nd/ clusters are systematically produced with an open velopharyngeal port, which closes only at the release of oral occlusion.

TEMPORAL EXTENT OF VELAR COARTICULATION

We examined the temporal course of the nasal airflow in VnV, CVnV and CVVnV (the vowel immediately preceding the nasal is /a/). The data clearly indicate that any oral consonant preceding /an/ has the effect of delaying the opening of the velopharyngeal valve, thus restricting the temporal range of anticipatory coarticulation. Analogous delay occurs when two vowels intervene between the oral and the nasal consonant. This can be interpreted as a perseveration of the high velar position of an oral consonant during the production of a following vowel. Moreover, while the onset time is quite stable in C(V)VnV sequences (15 to 40 ms), it is subject to remarkable variations (15 to 165 ms) in VnV items, suggesting that the velum may delay the lowering movement even when no constraints on its position are imposed by the articulation of a preceding oral consonant.

As seen in the preceding paragraphs, the data relative to the onset time of the velum rising and to the time of velum closing indicate that both tend to occur as late as possible, whatever is

the context (vocalic or consonantal) that follows the nasal. Altogether the results of the present research indicate a greater temporal extent of carryover than of anticipatory nasalization.

REFERENCES

[1] A.P. Benguerel,"Nasal airflow pattern and velar coarticulation in French" Speech Comm. Seminar, Stockholm, Vol.2, 105-112, 1974

[2] A.P. Benguerel, H. Hirose, M. Sawashima and T.Ushijima, "Velar height and its timing in French: a fiberoptic study" Ann. Bull. Logopedics and Phoniatrics, Tokio, 9, 67-78, 1975

[3] H. Clumeck, "Patterns of soft palate movement in six languages" J. Phonetics, 4, 337-351 1976

[4] E. Farnetani, K. Vagges and E. Magno Caldognetto " Coarticulation in Italian VtV sequences: a palatographic study" Phonetica, 42, 78-99, 1985

[5] J. Ohala, "Monitoring soft palate movements in speech" Project on Linguistic Analysis Report, Dep. of Linguistics, Univ. of California, Berkeley, 13, 1-15, 1971

[6] M. Rothenberg, "The breath-stream dynamics of simple-released-plosive production" Biblioteca Phonetica, 6, 1968.

FIGURES

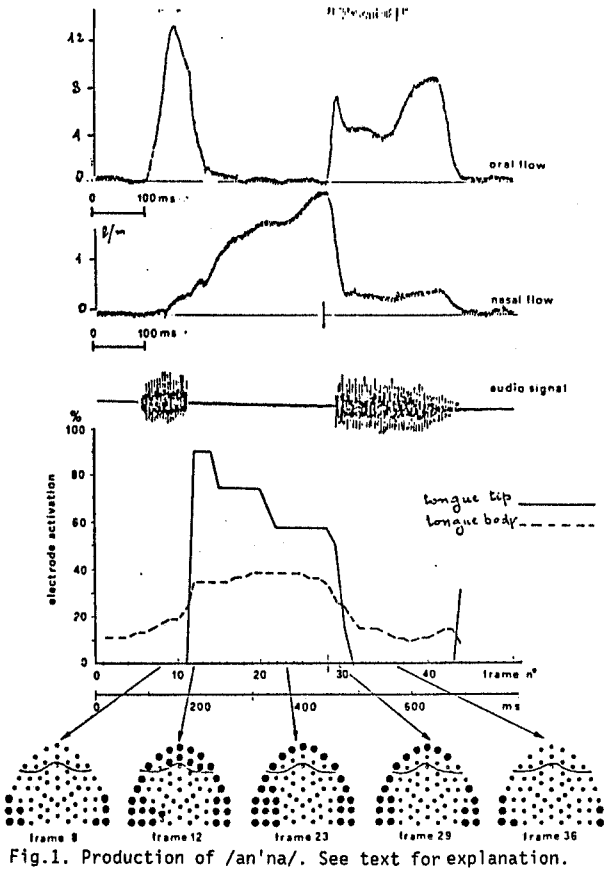


Fig.1. Production of /an'na/. See text for explanation.

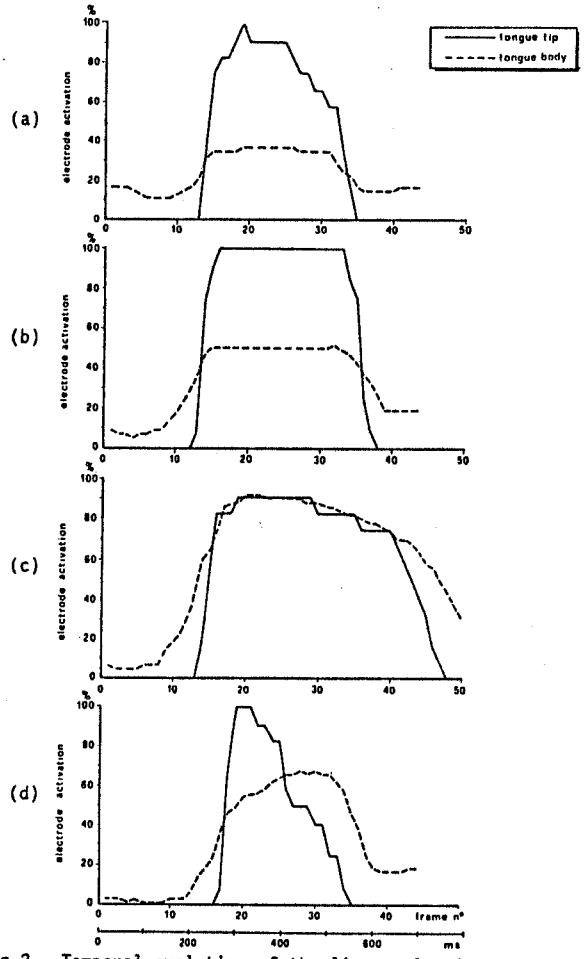


Fig.2. Temporal evolution of the linguopalatal contact in: /'anna/ (a), /'anta/ (b), /'antʃa/ (c), /'anja/ (d).

EVOLUTION DES PARAMETRES AMPLITUDE - FREQUENCE DU SIGNAL E.G.G. LORS DES CHANGEMENTS DE MECANISME VIBRATOIRE LARYNGE

B. Roubeau, C. Chevrie-Muller, C. Arabia

INSERM U3 - Paris

ABSTRACT

The laryngeal vibratory mechanisms during the shift from one register to another were the object of an EGG study of 10 men and 8 women during glissandos and held notes. Computerized analyses demonstrated two types of phenomena. A rapid one corresponds to the change of the vocal fold contact. The others which are slow and succeed each other are represented by a progressive change of the amplitude of the EEG waves followed by a loss of control of the frequency. These phenomena seem to represent two different neural control loops of the vibratory mechanisms of the vocal folds.

INTRODUCTION

Des observations déjà anciennes (Garnault [1]) et d'autres plus récentes (Hirano [2] [3], Askenfelt [4]) effectuées par des physiologistes ont montré clairement l'existence de 2 principaux mécanismes vibratoires laryngés résultant de 2 confrontations différentes possibles de la glotte. Bien que des études systématiques n'aient pas été réalisées, il semble bien que cela soit aussi le cas chez certains mammifères.

Des explorations E.M.G. ont permis de corréler ces changements morphologiques avec des modifications de l'activité des muscles intrinsèques du larynx (Gay [5]).

Le mécanisme vibratoire souvent appelé registre de poitrine et que nous nommerons mécanisme I permet la réalisation des sons les plus graves de l'étendue vocale.

L'autre mécanisme que nous nommerons mécanisme II souvent appelé registre de fausset permet d'émettre les fréquences les plus élevées de l'étendue vocale.

Une "zone" plus ou moins importante de fréquences peut être réalisée par l'un ou l'autre mécanisme.

Afin d'approcher un des aspects de la régulation de l'activité laryngée, nous nous sommes intéressés par l'intermédiaire de l'électroglottographie aux modifications de la vibration cordale lors du changement de mécanisme vibratoire.

MATERIEL ET METHODE

10 hommes et 8 femmes chanteurs et non chanteurs, d'âge et de tessiture différents ont participé à cette étude.

Tous les sujets ont été enregistrés en chambre sourde en utilisant un microphone L.E.M. type DU 70 et un électroglottographe FROKJAER-JENSEN type EG 830, muni d'une ou de trois paires d'électrodes. Le signal acoustique et l'électroglottogramme (E.G.G.) sont enregistrés par un magnétophone stéréophonique REVOX type A77.

Un mélographe de type FIES CNET reproduit la courbe d'évolution de la fréquence du signal EGG en effectuant toutes les trois ondes la moyenne des fréquences instantanées. Les mesures obtenues sont fiables entre 75 et 600 Hz.

La représentation graphique est réalisée par un traceur de type GOULD ES 1000.

Un mini-ordinateur de type MODCOMP IV est utilisé pour effectuer la digitalisation du signal analogique EGG qui est suivie d'un filtrage en fréquence et en amplitude.

L'analyse de l'amplitude et de la fréquence de chaque onde est réalisée à l'aide d'une méthode de détection des passages par zéro (Guidet et Chevrie-Muller, [6]). Pour ce qui est de la fréquence, chaque onde est représentée par un point défini par les coordonnées suivantes : fréquence en ordonée (échelle logarythmique) et temps en abcisse. L'amplitude normalisée de chaque onde est définie en ordonnée suivant une échelle linéaire, le temps est en abcisse.

Ce traitement informatique fournit également une revisualisation du signal EGG avec une dilatation de l'échelle de temps.

Protocole vocal

Nous avons demandé aux sujets d'effectuer des glissandos ascendant et descendant, c'est à dire des sons dont la fréquence augmente puis diminue de façon continue à la manière d'une sirène, du grave au plus aigu de leur étendue vocale, et ceci sur les phonèmes A, O et I.

Nous leur avons également demandé d'effectuer des sons tenus dont la fréquence est constante sur ces mêmes voyelles tout en changeant de mode d'émission ou de "registre" sans interrompre la production sonore.

Ces notes tenues ont été réalisées à différentes hauteurs dans la zone de fréquence des sons pouvant être émis suivant l'un ou l'autre des mécanismes vibratoires, c'est à dire aux environs de 293 Hz, 311 Hz et 329 Hz correspondant respectivement à Ré3, Re#3 et Mi3.

Une analyse du timbre vocal qui a parallèlement été réalisée (Castellengo et al [7]) à partir des enregistrements effectués au cours de l'étude électroglottographique ne figure pas dans ce rapport.

Changement de mode vibratoire durant l'émission de glissandos

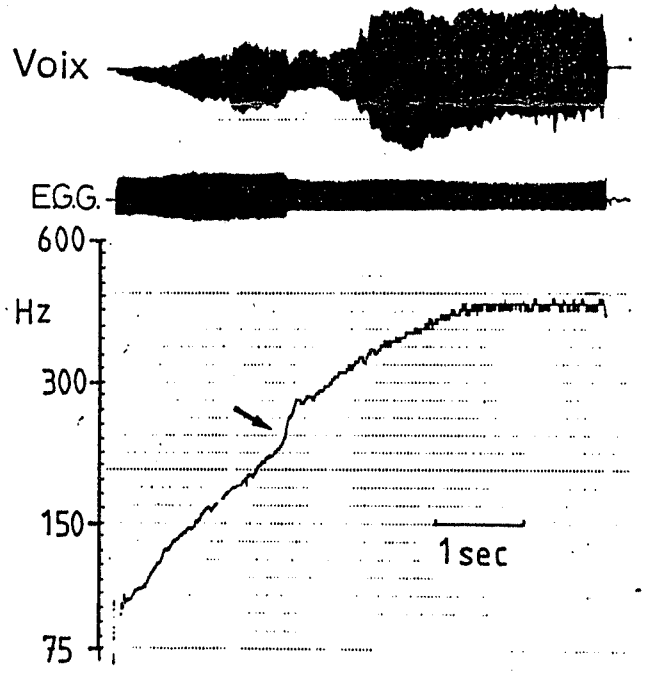


Figure 1 : Glissando ascendant sur la voyelle "A", sujet masculin, chanteur amateur, baryton.

L'analyse des enregistrements après visualisation sur le traceur GOULD (Fig. 1) met en évidence lors de l'émission d'un glissando ascendant, une diminution brutale de l'amplitude du signal EGG. Celle-ci coïncide avec une diminution brutale de l'amplitude du signal acoustique et une brève modification de la pente de la courbe mélodique. Cet évènement a été observé de manière identique chez tous les sujets excepté deux pour chacune des trois voyelles utilisées. Un glissando déscendant fournit une représentation tout à fait symétrique : l'amplitude du signal EGG et du signal acoustique augmente brutalement, la pente descendante de la fréquence s'accentue momentanément.

Ce phénomène déjà observé et corrèlé avec les résultats obtenus grâce à d'autres techniques d'investigation telles que l'électrophotoglottographie (Kitzing et al [8] ; Dejonckere [9]), l'électromyographie (Vennard et al [10]) ainsi que la cinématographie ultra-rapide (Childers [11]) traduit le changement brutal d'accolement des cordes vocales, correspondant à des modes vibratoires différents.

Les fréquences vibratoires au moment des changements de mécanisme ayant été mesurées nous avons noté que le "passage" se produit à un niveau plus élevé lors de l'émission de glissandos ascendants que de glissandos descendants et cette différence semble plus prononcée chez les sujets féminins que chez les sujets masculins. De plus, au cours des glissandos descendants, le niveau du "passage" est très souvent voisin pour les hommes et pour les femmes.

Analyse du signal digitalisé

La revisualisation de l'EGG digitalisé pour 5 hommes et 5 femmes montre la rapidité avec laquelle s'effectue le changement de mécanisme (Fig. 2). Pour chacun des sujets, le changement de mode de contact entre les cordes vocales s'effectue à peu près sur 2 périodes, soit 2 à 5 ms. Celui-ci semble toujours être précédé par la diminution d'amplitude de quelques ondes. Il peut être suivi par des oscillations d'amplitude et de forme instables (Fig. 2A). Cette instabilité n'apparait pas constante chez un sujet donné au cours de différents essais.

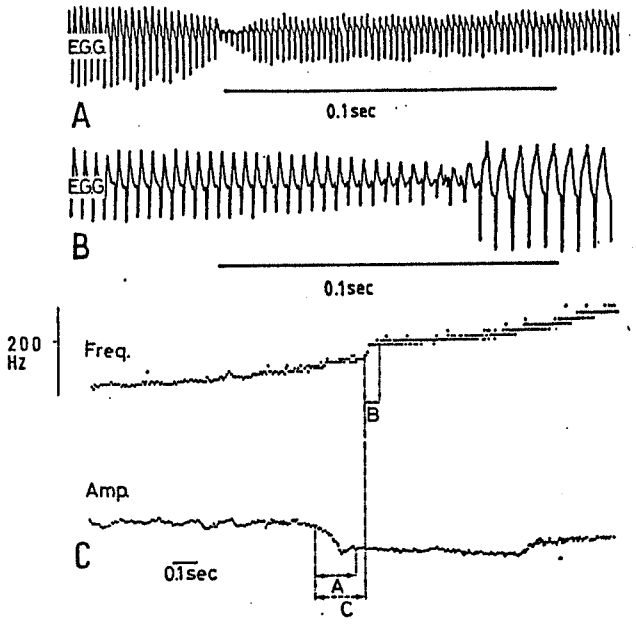


Figure 2 : Glissando sur la voyelle "O".

Figure 2 : A) EGG digitalisé revisualisé. Glissando ascendant, sujet féminin, chanteur professionnel, mezzo-soprano.

B) EGG digitalisé revisualisé. Glissando descendant, sujet féminin, chanteur professionnel, soprano.

C) Evolution de l'amplitude et de la fréquence instantanées de chaque onde du signal EGG. Glissando ascendant, sujet masculin, chanteur professionnel, contre-ténor.

Les mesure informatisées de la fréquence et de l'amplitude du signal EGG (Fig. 2C) montrent que les perturbations concernant ces deux paramètres lors du "passage" n'apparaissent pas simultanément et ne sont pas aussi brèves que le changement de forme de l'onde. Afin d'analyser plus en détail ce phénomène, nous avons mesuré la durée de ces évènements sur dix enregistrements correspondant à des glissandos ascendants et descendants (Tableau I).

Nous avons déterminé de manière constante (Fig. 2C) :

(A) : la durée de la phase de réajustement de l'amplitude des ondes EGG entre deux périodes stables.

(B) : la durée de la phase de réajustement de la fréquence correspondant au moment pendant lequel la pente d'évolution de la fréquence s'accentue.

(C) : le décalage temporel entre le début de chacune de ces phases A et B.

SUJETS	GLISSANDOS	DUREE A	DUREE B	DUREE C
6	ASC.	0.082 s	0.041 s	0.013 s
6	DESC.	0.065 s	0.048 s	0.013 s
4	ASC.	0.282 s	0.155 s	0.093 s
1	ASC.	0.251 s	*	0.027 s
2	ASC.	0.234 s	0.046 s	0.187 s
15	ASC.	0.182 s	0.048 S	0.117 s
15	DESC.	0.072 s	0.024 s	0.072 s
13	ASC.	0.069 s	0.062 s	0.096 s
11	ASC.	0.105 s	0.110 s	0.069 s
11	DESC.	0.072 s	0.048 s	0.062 s
MOYENNE		0.141 s	0.066 s	0.087 s
ECART TYPE		0.09	0.04	0.05

* non mesurée

Tableau I : Durée des phases de réajustement de l'amplitude et de la fréquence lors du changement de mode vibratoire au cours d'un glissando.

Les durées de réajustement sont variables d'un sujet à l'autre (écart type important). Par contre deux phénomènes semblent plus constants :

La durée (B) du réajustement de la fréquence est toujours plus courte que celle de l'amplitude (A) sauf dans un cas (sujet féminin, glissando ascendant).

Le réajustement de l'amplitude débute toujours avant celui de la fréquence (la durée (C) est toujours supérieure à 0).

Changements de mode vibratoire durant l'émission de sons tenus

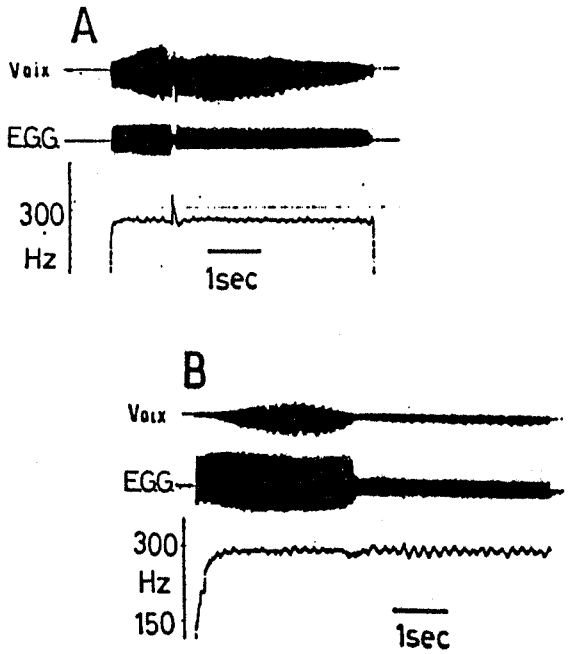


Figure 3 : Note tenue (≃ 293 Hz) sur la voyelle "A". Passage du mécanisme I au mécanisme II. Signal analogique et courbe mélodique.

A) Sujet masculin, chanteur professionnel, ténor.

B) Sujet masculin, chanteur professionnel, ténor.

Changement de mécanisme vibratoire sans saut de fréquence.

La réalisation de sons tenus s'étant avérée très difficile, voir impossible, pour les sujets féminins nous n'avons analysé que les enregistrements obtenus à partir des sujets masculins.

Le son tout d'abord émis suivant le mécanisme I est ensuite produit sans interruption suivant le mécanisme II. Les tracés obtenus (Fig. 3A) présentent les mêmes caractéristiques que celles décrites lors de l'analyse des glissandos : l'amplitude du signal EGG présente une diminution brutale lors du "passage". On constate de plus une perte de contrôle se manifestant par un saut de fréquence vers le haut suivi d'un rétablissement plus lent à la fréquence initiale. Sur la figure 4A on observe un accroissement de la fréquence pendant 0, 0212 seconde et un rétablissement de la fréquence à la

valeur initiale pendant 0,1168 seconde.

L'amplitude de ce saut de fréquence est variable mais a le plus souvent la valeur d'une quarte ou d'une quinte. Elle semble d'autant plus importante que l'intensité du signal acoustique est élevée.

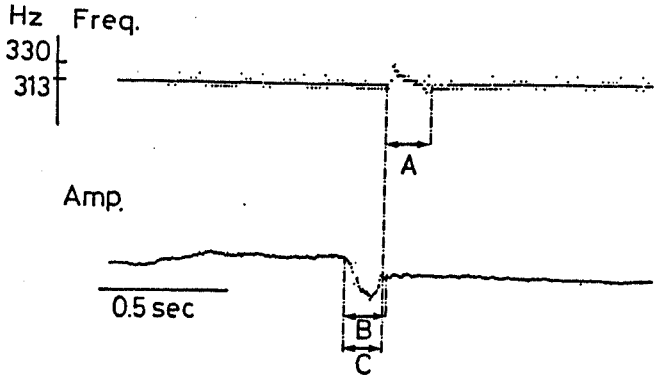


Figure 4 : Note tenue (≃ 313 Hz) sur la voyelle "O". Evolution de la fréquence et de l'amplitude instantanées de chaque onde du signal EGG. Sujet masculin, chanteur professionnel, contre-ténor.

Les représentations séparées de l'amplitude et de la fréquence du signal EGG obtenues après digitalisation ont permis de mesurer de la même manière que précédemment la durée des réajustements de ces deux paramètres lors du changement de mécanisme (Tableau II).

SUJETS	DUREE A	DUREE B	DUREE C
1	0.179 s	0.138 s	0.079 s
4	0.124 s	0.142 s	0.09 s
6	0.096 s	0.151 s	0.117 s
MOYENNE	0.133 s	0.143 s	0.095 s

Tableau II : Durée des phases de réajustement de l'amplitude et de la fréquence lors du changement de mode vibratoire au cours d'un son tenu.

La durée de la phase de réajustement de l'amplitude est la même que pour les glissandos et débute toujours avant la phase de réajustement de la fréquence. Cette dernière par contre apparait plus longue que pendant les glissandos.

DISCUSSION

L'électroglottographie est un mode d'investigation très indirect de l'activité laryngée, fournissant des données qui intègrent l'action de nombreux paramètres.

Tout d'abord la morphologie du cou des sujets étant variable d'un individu à l'autre (importance des tissus graisseux, taille des muscles et des cartilages), il est difficile d'interpréter les différences d'amplitude du signal entre les sujets.

D'autre part, des modifications du signal EGG peuvent être dues aux déplacements verticaux du larynx lors d'importantes variations de fréquence (Van Michel[12]) ce qui fut le cas dans les expériences que nous avons effectuées. Pour éliminer cet artefact possible nous avons construit un système d'électrodes constitué de trois paires d'électrodes de 1 cm2 chacune et montées en parallèle augmentant ainsi la taille du champ électrique afin de capter un signal d'amplitude constante quelque soit la position du larynx.

Cette technique mise au point ultérieurement n'a pu être appliquée à tous les sujets. Les résultats obtenus ont permis d'éliminer de fines variations d'amplitude du signal mais n'ont pas remis en question les principaux évènements observés lors des changements de mode vibratoire.

D'autre part, il a déjà été démontré que l'intensité du signal acoustique dépend principalement de la pression sous-glottique et non de la surface d'accolement des cordes vocales (Fig. 3B) (Koike, [13]). Nous pouvons considérer que les modifications brutales d'amplitude du signal EGG sont uniquement dues aux changements de la surface d'accolement des cordes vocales.

L'EGG permet donc d'objectiver des changements de mode vibratoire bien que le signal acoustique ne laisse transparaitre que peu de signes de cet évènement (Fig. 3B).

Il semble par ailleurs, que l'absence de modalité vibratoire intermédiaire entre le mécanisme I et II soit confirmée sur les tracés EGG tirés des sons tenus, car même chez les sujets ayant masqué le plus possible les perturbations acoustiques dues au "passage" (Fig. 3), le changement brutal de mécanisme apparait très nettement au niveau de l'EGG.

En ce qui concerne la différence de hauteur de "passage" entre les glissandos ascendants et descendants nous pensons que celle-ci est due au fait qu'un mode vibratoire n'est pas spontanément abandonné pour un autre avant d'en avoir atteint les limites. Ainsi le passage du mécanisme I au mécanisme II se produit au cours d'un glissando ascendant lorsque le sujet approche de la limite supérieure du premier. Dans le cas d'un glissando descendant, le "passage" a lieu lorsque le sujet approche de la limite inférieure du deuxième.

Phase de réajustement de la fréquence

Cette phase est plus longue lors du changement de mode vibratoire pendant un son tenu (Tableau I et II). Dans un glissando ascendant le saut de fréquence va dans le sens de l'évolution dynamique de la fréquence, de même au cours d'un glissando descendant. Ceci n'est pas le cas pour les sons tenus car le sujet doit maintenir la fréquence constante. La perte de controle de la hauteur du son doit donc être corrigée par contrôle auditif ce qui explique l'allongement de la durée de réajustement.

L'amplitude du saut de fréquence au cours des sons tenus semble fonction de l'intensité de l'émis-

sion bien que nous n'avons pas effectué de mesure systématique à ce sujet. Un sujet (Fig. 3B) peut arriver à éviter cette perturbation en diminuant considérablement l'intensité sonore avant le changement de mécanisme.

Phase de réajustement de l'amplitude

La durée de cette phase apparait constante dans les deux types de production et semble donc directement liée à la modification de la conformation glottique.

CONCLUSION

Nous avons donc observé au cours de ces expériences l'éxistance de phénomènes au niveau du signal EGG qui apparaissent dans un ordre constant lors du changement de mécanisme.

La diminution d'amplitude du signal EGG traduit l'amincissement des cordes vocales dû au changement d'activité des muscles laryngés, principalement thyro-aryténoidiens et crico-thyroidiens.

Cet évènement déclenché par une modification de la commande nerveuse est le premier à caractériser le changement de mécanisme vibratoire.

La stimultation croissante des propriorécepteurs laryngés (baro et tensorécepteurs) lors de l'élévation de la fréquence atteindrait un seuil qui provoquerait une modification de l'activité des motoneurones des muscles laryngés.

Ainsi, lors du passage se produisait l'inhibition d'une partie de la population des motoneurones bulbaires mis en jeu au cours du mécanisme I tandis que l'autre partie de cette même population serait excitée.

Apparait ensuite la transformation de la forme de l'onde qui traduit la modification de la vibration glottique consécutive à la nouvelle répartition des tensions musculaires à l'intérieur du système vibratoire. Il s'agirait donc dans ce cas d'un phénomène d'adaptation mécanique.

Il pourrait en être de même pour la perte de contrôle de la fréquence, qui serait liée au rétablissement de l'équilibre entre les tensions du vibrateur et la pression sous-glottique. En effet, il semble que plus la puissance de l'émission est importante au moment du changement de mécanisme vibratoire, plus le décalage tensitoire de la fréquence est important.

Enfin, le phénomène le plus tardif est la correction de cette perturbation de la fréquence observable surtout au cours des sons tenus. Cet aspect tardif laisse à penser qu'il s'agit là d'une régulation plus lente de la vibration glottique faisant intervenir le contrôle auditif (boucle cochléo-récurentielle).

Ces différents phénomènes traduiraient l'activation de deux boucles de régulation neuro-motrices. L'une courte ayant pour point de départ les proprio-récepteurs laryngés et l'autre plus longue ayant pour point de départ la cochlée et faisant intervenir des centres d'intégration supérieurs.

Il semble donc que grace à une méthode d'investigation, telle que l'EGG, l'on puisse approcher indirectement des phénomènes physiologiques de régulation motrice et des phénomènes physiques d'adaptation mécanique d'un organe aussi complexe que le larynx.

REFERENCES

1 D. Garnault : Physiologie, hygiène et thérapeutique de la voix parlée et chantée. Maloine ed, Paris, 1895

2 M. Hirano : Morphological structure of the vocal cord as a vibrator and its variations. Folia Phon., 26, 89-94, 1974.

3 M. Hirano : The role of the layer structure of the vocal folds in register control. Vox Humana, University of Iyvaskyta, 1982.

4 A. Askenfelt, J. Gauffin, J. Sundberg and P. Kitzing : A comparison of microphone ane electroglottograph for the measurement of vocal fundamental frequency. J. Speech & Hear, Res. 23, 250-273, 1980.

5 T. Gay, M. Strome, H. Hirose, M. Sawashima : Electromyography of the intrinsic laryngeal muscles during phonation. Ann. Otol. Rhinol. Lar. 81, 401-409, 1972.

6 C. Guidet, C. Chevrie-Muller : Computer analysis of prosodic and electroglottographic parameters in diagnosis of pathologic voice. Quantitative linguistics. In P. Winkler Investigations of the speech process, 233-263, 1983.

7 M. Castellengo, B. Roubeau, C. Valette : Study of the acoustical phenomena charactéristic of the transition between chest voice and falsetto. Proc. Stokholm Music Accustics Conf., Royal Swedish Acad. Music 46, 113-125, 1985.

8 P. Kitzing, B. Calborg, A. Lofquist : Aerodynamic and glottographic studies of the laryngeal vibratory cycle. Folia Phon. 34, 216-224, 1982.

9 P. H. Dejonckere : Comparison of two methods of photoglottography in relation to electroglottography. Folia phon. 33, 338-347, 1981.

10 W. Vennard, M. Hirano, J. Ohala : Chest, head and falsetto. The nats bulletin. 30-37, 1970

11 D. G. Childers, A. M. Smith, G. P. Moore : Relationship between electroglottograph, speech and vocal cord contact. Folia Phon. 36, 105-118, 1984.

12 C. Van Michel : Les artéfacts en électroglottographie. Electrodiagnostic thérapie, n° 4, 1967.

13 Y. Koike, W.H. Perkins : Application of a miniaturized pressure, 1968.

RELATIONS ACTIVITES MUSCULAIRES MOUVEMENTS DE LA MACHOIRE DANS LA PAROLE

M. Gentil & T. Gay

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

ABSTRACT

The activity patterns of the mandibular elevator muscle system were investigated for speech gestures. Intra-muscular wire electrodes were placed into the seven elevator muscles of the mandibular system. The activity of these muscles was recorded along with the jaw displacement in two-dimensional space and the acoustic signal. One adult, native speaker of French, produced different consonant-vowel-consonant syllables at two different speaking rates. Results indicated that intra-individual differences exist in the selection of a given muscle or muscle system for a specific function and/or utterance and in the coordination of muscle activity pattern for a particular system.

INTRODUCTION

Depuis la première application de l'electromyographie à l'étude du fonctionnement des muscles masticateurs (1), un certain nombre d'études EMG ont été rapportées dans la littérature (2, 3, 4, 5, 6). Cependant l'organisation des modèles d'activités musculaires de la mâchoire au cours des mouvements sous-tendant la production de la parole est encore peu connue. Dans une présentation antérieure (7), nous examinions les activités musculaires du système mandibulaire dans des gestes relatifs à la parole et autres, c'est-à-dire gestes fonctionnels et mouvements de mastication, afin de déterminer la possibilité de spécialisation neuromusculaire pour la parole.

Cette étude complétait la précédente et visait à montrer la variabilité intra-individuelle dans le choix d'un muscle donné et dans la coordination des activités musculaires pour un mouvement mandibulaire spécifique. Dans ce but, les activités musculaires et les mouvements de la mâchoire en deux dimensions étaient enregistrés pendant qu'un sujet français produisait des monosyllabes à deux différentes vitesses de parole.

METHODE

Une femme française servait de sujet dans cette expérience.

Le matériau phonétique comportait dix syllabes consonne-voyelle-consonne où les consonnes étaient /p/, /b/ ou /t/ et les voyelles /o/, /i/, /a/, /u/, /ɛ/ /ɔ/. Chacune de ces productions était introduite dans une phrase porteuse "Ce ... encore" et distribuée au hasard dans plusieurs listes. Elle était répétée quinze fois à deux débits de parole (conversationnelle et rapide).

Au moyen d'électrodes à crochets (8), on recueillait simultanément l'activité de huit muscles : ptérygoïdien interne, masseters superficiel et profond, têtes supérieure et inférieure du ptérygoïdien externe, parties antérieure et postérieure du temporal, ventre antérieur du digastrique. Le placement des électrodes, techniques d'insertion et vérification suivaient les indications données par Gross et ses collègues (1, 4, 9).

Les mouvements mandibulaires étaient enregistrés en deux dimensions verticale et antérieure-postérieure, au moyen d'un système de magnétomètres ("mandibular kinesiograph"). Le dispositif se composait de six capteurs fixés sur une structure d'aluminium, elle-même montée sur des lunettes spéciales que portait le sujet. Cet appareillage et l'ensemble électronique qui s'y associait traçait le mouvement d'un aimant permanent fixé à l'incisive centrale inférieure par une pâte dentaire. Les signaux du mouvement étaient enregistrés simultanément avec les signaux EMG et acoustique sur un magnétophone à 14 canaux.

Le traitement des données était effectué sur ordinateur PDP 11/34. Le processus hardware rectifiait le signal et calculait les EMG intégrés sur un intervalle effectif de 5 millisecondes. A partir du disque analogique, on

entrait les données dans l'ordinateur répétition par répétition en utilisant une fenêtre de 2,5 secondes. Les répétitions des productions de parole étaient repérées à un point précis du signal acoustique, c'est-à-dire le début d'explosion de la première consonne. Toutes les données électromyographiques et cinématiques étaient moyennées de manière à minimiser le problème de variabilité.

RESULTATS

L'étude précédente montrait que l'activité de la mâchoire était basée sur des stratégies individuelles plutôt que sur un ensemble de règles universelles (Cf la variabilité des modèles d'activités musculaires entre les trois sujets). Dans cette étude, l'analyse de l'activité musculaire indiquait que pour le sujet donné il n'y a pas une relation causale stéréotypée entre l'EMG dans tout muscle précis et le mouvement qui lui est associé. Nous constaterons la variabilité des modèles d'activités musculaires pour :

- Une même fonction dans différentes productions
- Une même production à différentes vitesses de parole

Variabilité des activités musculaires pour une même fonction dans différentes productions

Considérant les différentes productions de notre corpus, nous examinions les fonctions de fermeture et protrusion au cours de diverses productions afin de montrer que ce ne sont pas nécéssairement les mêmes muscles qui sont actifs pour une même fonction. La variabilité dans les activités musculaires concernant la fermeture est illustrée dans le tableau 1 qui appelle les remarques suivantes :

- Dans toutes les productions, quelle que soit la vitesse de parole, le ptérygoïdien interne (MP) intervient pour la fermeture de la mâchoire
- Deux autres muscles, les fibres postérieures du temporal (PT) et le masseter superficiel (SM) ont une activité synergique. Ainsi, soit le temporal postérieur (PT), soit le masseter superficiel (SM) travaillent avec le ptérygoïdien interne qui peut également travailler seul.
- On observe pour une même production des différences de modèles d'activités musculaires entre les deux vitesses de parole.

TABLEAU 1 - FERMETURE MACHOIRE : ACTIVITE MUSCULAIRE

PRODUCTIONS	FERMETURE	
	CONV.	RAPIDE
/BAP/	MP	MP & PT
/PAP/	SM & MP	MP
/BAB/	MP	MP
/KAP/	MP	MP & PT
/STAP/	MP & SM	MP
/PɛP/	MP	MP & PT
/PIP/	MP & SM	MP
/PUP/	MP	MP
/POP/	MP	MP & SM
/PɔP/	MP	

MP = PTERYGOIDIEN INTERNE, SM = MASSETER SUPERFICIEL, PT = TEMPORAL POSTERIEUR

TABLEAU 2 - PROTRUSION MACHOIRE : ACTIVITE MUSCULAIRE

PRODUCTIONS	MOUVEMENT EN AVANT	
	CONV.	RAPIDE
/BAP/		MP
/PAP/	SM	MP
/BAB/		MP
/KAP/	MP	MP & SM
/STAP/	MP & SM	MP
/PɛP/	MP & SM	MP
/PIP/	MP & SM	MP
/PUP/	MP	MP
/POP/	MP	MP & SM
/PɔP/	MP	

MP = PTERYGOIDIEN INTERNE, SM = MASSETER SUPERFICIEL

Le tableau 2 présente les modèles d'activités musculaires pour la protrusion de la mâchoire. Les remarques suivantes peuvent être faites :

- Nous constatons l'activité de deux muscles, le masseter superficiel (SM) et le ptérygoïdien interne (MP). Ces muscles travaillent soit séparément, soit ensemble (activité synergique)
- Parfois aucun des muscles enregistrés n'est actif pour le mouvement en avant de la mâchoire (Cf production /bap/, /pap/ en vitesse conversationnelle).

- Des différences d'activités musculaires existent pour une même production réalisée à deux vitesses de parole différentes. Par exemple, pour /pap/, le masseter superficiel (SM) travaille en vitesse conversationnelle et le ptérygoïdien interne (MP) en vitesse rapide.

Variabilité des activités musculaires pour une même production à différentes vitesses de parole

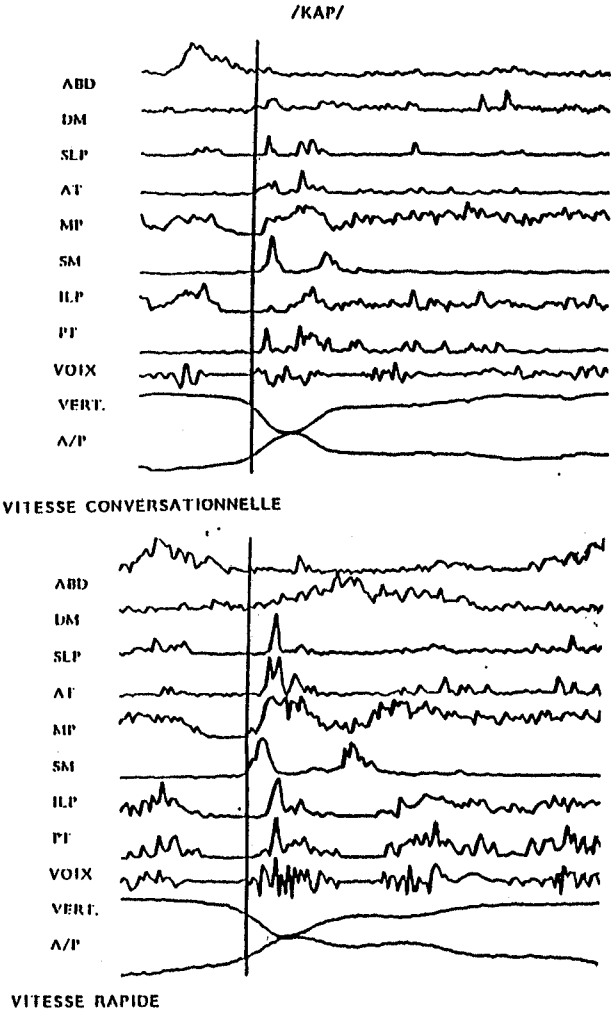
Les deux précédents tableaux ont déjà indiqué une différence d'activités musculaires entre vitesse conversationnelle et rapide. La figure 1 présente pour une même production, /kap/, les modèles d'activités musculaires dans les deux vitesses de parole. En vitesse conversationnelle, nous remarquons principalement l'activité du ptérygoïdien interne (MP). Etant donné sa faible amplitude, 26 µV*, le pic du masseter superficiel* doit être négligé. En vitesse rapide, nous constatons des activités musculaires différentes : celles du ptérygoïdien interne pour la fermeture, du masseter superficiel pour le mouvement en avant et des fibres postérieures du temporal pour les mouvements de rétraction et de fermeture. Comme en vitesse conversationnelle, les pics des SLP, AT et ILP n'ont pas été pris en compte étant donné leur faible amplitude.

En addition à ces différences d'activités musculaires nous observions aussi des différences d'organisation temporelle entre les deux débits de parole d'une même production. Comme on pouvait s'y attendre, la durée d'activité musculaire est plus courte en vitesse rapide quelle que soit la production (en comparaison de vitesse conversationnelle). Par exemple, en ce qui concerne l'activité du ptérygoïdien interne pour /pap/, /bap/ et /kap/, la durée est respectivement 195, 120 et 165 millisecondes en vitesse conversationnelle versus 144, 147 et 162 millisecondes en vitesse rapide.

CONCLUSION

Cette étude prouve que les modèles d'activités musuclaires sont essentiellement variables à l'intérieur des réalisations d'un même individu. Cette variabilité intra-individuelle dans la parole vient renforcer la variabilité interindividuelle déjà constatée dans notre expérience précédente (7), comme dans d'autres expériences ayant trait à des mouvements mandibulaires ne concernant pas

Fig. 1 - MODELES ACTIVITES MUSCULAIRES



la parole (2, 3). Toutes ces constatations nous permettent d'affirmer que les fonctions de la mâchoire en général ne sont pas fondées sur un ensemble de règles neuromusculaires universelles.

Ces résultats mettent une nouvelle fois à l'évidence le principe de "motor equivalence" (5, 10) : les différents muscles de la mâchoire ne contribuent pas stéréotypiquement à un mouvement particulier. Ceci confirme notre constatation antérieure faite à partir d'un autre organe articulatoire, les lèvres (11). Les diverses contributions musculaires à une fonction donnée proscrivent les approches basées sur le fait qu'un muscle ou un mouvement particulier participe toujours de façon identique à un but phonétique particulier.

*l'échelle des amplitudes varie d'un muscle à l'autre

REFERENCES

(1) E.E. Moyers, "An electromyographic analysis of certain muscles involved in temporomandibular movement", Am. J. Orthodontics, 36, pp. 481-515, 1950

(2) J. Ahlgren, "Mechanism of mastication", Acta Odont. Scand., 24 (Suppl. 44), 1966

(3) M. Vitti & J.V. Basmajian, "Integrated actions of masticatory muscles : simultaneous EMG from eight intramuscular electrodes", Anat. Rec., 187, pp 173-189, 1977

(4) T. Gay, B. Gross, D. Lipke & J. Yaeger, "An electromyographic study of both portions of the masseter muscle", J. Dent. Res., 56, Abstract, 1977

(5) J. W. Folkins, "Muscle activity for jaw closing during speech", JSHR, 24, pp 601-614, 1981

(6) B. Tuller, K.S. Harris & B. Gross, "An electromyographic study of the jaw muscles during speech", J. Phonetics, 9, pp 175-188, 1981

(7) M. Gentil & T. Gay, "Jaw muscle activity for speech and non-speech gestures", 106th meeting ASA, 74, S 116, Abstract, 1983

(8) J. V. Basmajian & G.A. Stecko, "A new bipolar indwelling electrode for electromyography", J. Appl. Physiology, 17, p 849, 1962

(9) B. Gross & D. Lipke, "A technique for percutaneous lateral pterygoid electromyography", Clin. Neurophysiology, 19, 47-55, 1979

(10) J.H. Abbs, "Speech motor equivalence : A need for a multilevel control model", 9th Int. Cong. Phonetic Sciences, Copenhagen, 2, pp318-321, 1979

(11) M. Gentil, V. Gracco & J.H. Abbs, "Multiple muscle contribution to labial closure during speech : evidence for intermuscle motor equivalence", 11th Int. Cong. Acoustics, Paris, 4, pp 11-14, 1983

APPROCHE PHYSIOLOGIQUE DES INTONATIONS DE BASE DU FRANCAIS : CRICOTHYROIDIEN ET FREQUENCE FONDAMENTALE

Denis Autesserre, Albert Di Cristo, Daniel Hirst

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

The electromyographic signals of the cricothyroid (CT) of a French speaker were recorded during the pronunciation of isolated syllables and utterances with different intonation patterns (affirmative, interrogative, vocative, imperative, implicative).
In this study, we present the results of an analysis of the various degrees of activity of this muscle which behaves like the first derivative of the basic patterns of FO.
We investigate in particular the effects relative to the register, to the distribution of the patterns in the utterance and to the segmental context.

INTRODUCTION

L'évolution de la fréquence fondamentale (FO) dans la parole dépend essentiellement du double contrôle des muscles du larynx et de la pression sous-glottique (Ps). L'influence de la Ps est surtout **globale** : l'évolution de ce paramètre aérodynamique rend bien compte de l'effet dit de déclinaison, propre à tout un énoncé. Cependant ses variations ne peuvent être à l'origine des **inflexions locales** de FO (montées et chutes), liées à l'activité des muscles du larynx [1, 2, 3].
La plupart des spécialistes, et en particulier [4] et [5], attribuent à trois paires de muscles laryngés le contrôle de la **variation positive de FO** : le cricothyroïdien (CT), le vocalis (V) et le cricoarythénoïdien latéral (CAL). Toutefois le V et le CAL, dont la fonction primaire est d'assurer l'adduction des cordes vocales, semblent moins directement associés aux variations locales de FO que ne l'est le CT. Le contrôle des **inflexions négatives de FO** est moins bien connu [6, 7]. Il n'est pas exclu que le CT joue, cette fois encore, un rôle fondamental, mais dans ce cas, par sa relaxation. Celle-ci s'accompagne certainement d'une participation de certains muscles sous-hyoïdiens tels que le sternohyoïdien [7] et le sternothyroïdien [8, 9].
Nous nous proposons d'analyser les relations entre les diverses phases d'activité du CT et la réalisation des contours intonatifs fondamentaux du français, en fonction de leur distribution dans l'énoncé et en tenant compte du registre atteint et du contexte segmental.

PROCEDURE EXPERIMENTALE ET CRITERES D'INTERPRETATION

1. Pour établir ces relations entre patrons intonatifs et leurs corrélats physiologiques, un corpus a été constitué. Il comprend des logatomes mono- et disyllabiques prononcés isolément, des mots et des logatomes intégrés en position finale d'une phrase porteuse "il a dit ---" (deux versions : l'une interrogative, l'autre affirmative) et un texte conçu afin de permettre la réalisation des principaux patrons intonatifs du français.

2. Technique électromyographique (EMG).
Nous avons recueilli l'activité EMG de 3 muscles : l'orbiculaire, au niveau de la lèvre inférieure, le cricothyroïdien et un muscle sous-hyoïdien. Nous nous limiterons, dans la présente communication, à l'étude des tracés du CT. Alors qu'il est possible d'utiliser une technique EMG de surface pour enregistrer l'activité des muscles des lèvres, celles des muscles du larynx nécessite l'emploi d'électrodes implantées, mises en place dans ce cas par voie transcutanée. La technique choisie est celle des électrodes flexibles bipolaires (2 fils de platine) décrite par [10] et reprise, entre autres, par [11] et [12].
Afin de vérifier si les électrodes ont bien été implantées dans le muscle choisi, on a procédé à plusieurs épreuves de contrôle : respiration calme puis profonde, déglutition, résistance à l'ouverture de la mâchoire. Le déclenchement des potentiels d'action à des moments différents par rapport au muscle sous-hyoïdien, nous permettent de penser qu'il ne peut s'agir que du CT.
Les signaux EMG et la parole sont conservés grâce à un enregistreur magnétique AMPEX (7 pistes) et l'enregistrement sur papier assuré à l'aide d'un enregistreur GOULD ES 1000. Les signaux de parole sont captés, à la fois, à l'aide d'un micro et d'un électroglottographe (celui de FROKJAER-JENSEN, type EG 830), muni d'un détecteur de mélodie. Le sujet et les capteurs sont placés dans une chambre insonorisée et les enregistreurs disposés en dehors de la cabine.

3. Les paramètres enregistrés et leurs mesures.
L'analyse porte sur les tracés du phonogramme, du glottogramme, de la courbe mélodique extraite du signal glottographique et des deux types de représentation des signaux EMG du CT : brut et redressé.

Les divers types de tracés correspondant à des étapes différentes de la contraction musculaire (ébauchée, franche ou forte) ou à des épisodes de relâchement ont été identifiés qualitativement et mis en relation avec le signal acoustique.
Cette comparaison est complétée par une mise en correspondance, à partir du signal EMG redressé, entre les pics d'activité des bouffées EMG les plus importantes et les sommets des contours de FO sur la courbe mélodique.

RESULTATS ET DISCUSSION

1. Rappel préalable du rôle du CT dans la biomécanique des fonctions laryngées.
La contraction du CT (Fig. 1) a pour effet d'étirer longitudinalement le pli vocal (allongement et accroissement du ligament vocal avec en même temps élongation mais tension plus faible dans le muscle vocal), au moyen de deux mouvements possibles de l'articulation crico-thyroïdienne [13].
- un mouvement de rotation, qui a pour conséquence de modifier l'angle crico-thyroïdien antérieur et de pincer l'espace inter-crico-thyroïdien (action des fibres de la pars recta du muscle).
- un mouvement de translation (effet de tiroir), avec subluxation de l'articulation, la corne thyroïdienne glissant vers l'avant (action de la pars obliqua).
Si les effets de la contraction du CT apparaissent bien connus, les signaux EMG enregistrés par notre locuteur nous conduisent à mettre en évidence des faits moins souvent signalés de relâchement de ce muscle (prenant même une allure d'inhibition) qui paraissent, dans ce cas, particulièrement significatifs.

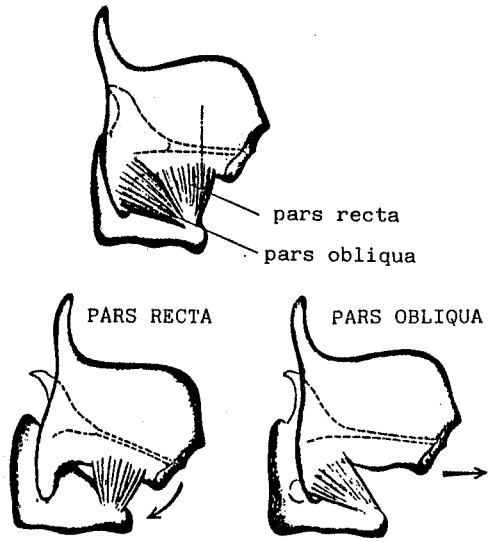


Figure 1 : Illustration des mécanismes d'étirement des cordes vocales par le cricothyroïdien (d'après Fink et Desmaret, 1978).

2. Les phases de relâchement.
Le signal EMG disparaît presque entièrement avec, tout au plus, quelques potentiels isolés. Ce relâchement (de 90 à 140 ms) s'effectue selon un chronométrage influencé par le caractère voisé ou non voisé de l'élément phonique présent à l'initiale absolue. Il précède une voyelle ou une consonne non voisée (ou dévoisée) mais se produit, la plupart du temps, pendant la réalisation d'une consonne non voisée (Fig. 2 et 3). Ce relâchement apparaît au début des phrases énonciatives et des questions totales, mais non à l'initiale des questions partielles, du type "qu'est-ce que tu as dit ?" qui se caractérisent par un registre d'attaque plus élevé (Fig. 4). Il en est de même pour les séquences mono- ou disyllabiques commençant par un ton haut immédiatement suivi d'un ton bas (Fig. 5).
Il y aurait donc là une véritable pré-programmation prenant en compte, à l'attaque de l'énoncé, la présence du voisement et la réalisation d'un registre aigu.
En finale d'énoncé, ce relâchement a bien lieu en relation directe avec la chute de FO pour les finalités, les parenthèses basses, l'implication (Fig. 2) et lorsque la chute suit une montée dans le contour mélodique complexe de l'appel ("Jean Paul !"; Fig. 6).
Dans tous ces cas, le signal EMG reprend de manière anticipée avant que la voyelle ne soit terminée sur le tracé du phonogramme. Il n'y a pas de relâchement après un contour simple de question totale (Fig. 3).

3. Les phases d'activité.
Elles se caractérisent par des bouffées électromyographiques importantes, surmontées de pics intenses. Nous allons envisager les relations entre ces pics du signal EMG et les sommets de FO de la courbe mélodique.
L'activité EMG est systématiquement associée aux sommets des contours intonatifs majeur et mineur (Fig. 7 a et 7 b). En ce qui concerne ces contours, la hiérarchie des pics d'activité du CT reflète bien celle des sommets de FO. Toutefois, cette corrélation peut être partiellement occultée par l'effet intrinsèque des voyelles. Ainsi, dans la phrase "Le fils de Paul est parti" (7 a), on observe la présence d'un pic d'activité du CT de plus forte amplitude sur "fils" que sur "Paul", bien que ces syllabes possèdent des sommets de FO quasiment identiques. Par contre, dans la phrase réitérée (7 b) où l'effet intrinsèque est neutralisé, la hiérarchie des pics du CT est fortement corrélée à celle des sommets de FO.
Dans tous les types de continuation, la phase d'activité la plus importante commence de manière anticipée sur la consonne de la syllabe accentuée porteuse de l'intonème (Fig. 7 a).
Les pics d'activité EMG sont associés à un sommet mélodique correspondant à un contour interrogatif.
Par rapport aux continuations, les questions totales présentent une moindre anticipation de l'activité EMG sur la consonne qui précède. Le pic principal se situe vers le milieu, ou dans le dernier tiers de la voyelle et il est décalé vers la fin de la voyelle et sur la consonne qui suit dans la phrase "Tu es sûr ?" (Fig. 8). L'activité maximale se poursuit après la voyelle quand celle-ci se trouve en finale absolue et continue aussi sur la consonne finale

quelque soit la configuration de F0 de cette consonne.

4. Amplitude des pics d'activité et fréquence intrinsèque des voyelles.
L'amplitude du pic d'activité a été mesurée pour les voyelles finales des phrases interrogatives (phrase porteuse "il a dit --- ?" terminée par un monosyllabe de type CV où C est soit [p], soit [b] et V peut être [i, a, u, ɛ̃], ou [ɛ] pour les mots "pelle" et "peine").
Les valeurs moyennes de l'amplitude de ce pic reflètent la hiérarchie des valeurs de fréquence intrinsèque observée par [15] et [16], car elles décroissent des voyelles hautes aux voyelles basses, les voyelles nasales occupant une position intermédiaire (Tableau I).

Tableau I

Valeurs de l'amplitude du pic d'activité du CT (en -V).

	Voyelles basses [ɛ, a]	Voyelles hautes [i]	[u]	Voyelle nasale [ɛ̃]
Série 1	866	1033	1066	933
Série 2	779	841	957	985
Moyenne (1 & 2)	822	937	1011	959

DISCUSSION

1. La mise en correspondance systématique des pics d'activité EMG et des courbes mélodiques, nous conduit à considérer la courbe de F0 comme l'intégrale de l'activité du (CT) : le maximum d'amplitude de ces pics coïncide avec le point d'inflexion de la courbe ascendante de F0 (Fig. 9), comme s'il s'agissait de la première dérivée de cette courbe.

2. La comparaison du chronométrage de l'activité du CT pour les interrogatives et les continuations justifie la distinction établie par [17] entre courbe concave et courbe convexe, ou encore, après nos analyses, entre montée retardée (interrogatives) et montée anticipée (continuatives).

3. Les premières analyses entreprises sur quelques syllabes tendraient à montrer que la fréquence intrinsèque n'est pas que le résultat d'une contrainte physiologique mais correspond aussi à un processus actif [18].

4. Les résultats de cette étude confirment le rôle majeur du (CT), principal muscle contrôlant les variations de F0, en français, comme cela avait déjà été démontré dans d'autres langues : anglais, japonais, hollandais, thai, etc. Mais la qualité exceptionnelle des documents réalisés nous permet d'affiner l'analyse par rapport à nos prédécesseurs, notamment en ce qui concerne le chronométrage de l'activité ou de la non activité (relâchement) de ce muscle, ou la prise en compte de la nature des unités segmentales dans la programmation nécessaire à la réalisation des cibles intonatives.

CONCLUSION

Pour dépasser ces résultats, pourtant déjà plus que prometteurs, il nous paraîtrait intéressant d'entreprendre une modélisation des rapports entre l'évolution de la fréquence fondamentale et celle du signal électromyographique du (CT) qui lui est associée.
Ces recherches ouvrent la voie à d'autres explorations visant à mettre en relation modèles d'intonation et paramètres physiologiques finement analysés. N'est-ce pas ce que Fujimura [8] annonçait déjà en 1977 lorsqu'il écrivait : "L'activité du cricothyroïdien reflète la structure tonale sous-jacente".

REMERCIEMENTS

Cette expérience a été réalisée dans le service de Neurophysiologie de l'Unité INSERM 3 (Hôpital de la Salpêtrière à Paris). Nous tenons à exprimer nos plus vifs remerciements à Mme le Docteur Chevrie-Muller, au Docteur Lacau-Saint-Gilly, à M. Maton ainsi qu'à Mme Arabia et M. Roubeau.

BIBLIOGRAPHIE

[1] Collier, R. "Laryngeal muscle activity, subglottal air pressure and the control of pitch in speech", **Haskins Lab. Status Rep. Speech Res.**, SR - 39/40 : 137-170, 1974.
[2] Collier, R., "Physiological correlates of Intonation patterns", **J. Acoust. Soc. Am.**, vol. 58, n° 1 : 249-255, 1975.
[3] Ohala, J., **Production of Tone**, in V. Fromkin (ed.) **Tone : a Linguistic Survey**, New York, Academic Press, 1978.
[4] Hirano, M., Ohala, J. & Vennard, W., "The function of laryngeal muscles in regulation of fundamental frequency and intensity of phonation", **J. Speech Hear. Res.**, 12 : 616-628, 1969.
[5] Sawashima, N., "Research on some basic aspects of the larynx", **Status Report on Speech Research** (Haskins Labs.), 23 : 179-208, 1970.
[6] Lieberman, P., "A study of prosodic features", **Status Report on Speech Research** (Haskins Labs.), 23 : 179-208, 1970.
[7] Ohala, J., "How pitch is lowered ?", **Mim. Phonology Lab. Un. of Berkeley** : 25-30, 1972.
[8] Fujimura, O., "Phonological functions of the larynx in phonetic control", **Invited Paper at the International Congress of Phonetic Sciences, Miami**, Dec. 1977.
[9] Maeda, S., "On the F0 control mechanisms of the larynx", **Séminaire Larynx et Parole**, Institut de Phonétique de Grenoble, GALF; 243-257, 1979.
[10] Basmajian, J.V. & Stecko, G.A., "A new bipolar indwelling electrode for electromyography", **Journal of Applied Physiology**, 17, 849, 1962.
[11] Hirose, H., "Electromyography of the articulatory muscles : Current instrumentation and techniques", **Haskins Lab. Status Rep. Speech Res.**, SR 25/26 : 73-86, 1971.
[12] Bonnot, J.F.P., Chevrie-Muller, C., Greiner, G.F. & Maton, B., "A propos de l'activité électromyographique labiale et vélaire durant la production de consonnes et de voyelles nasales en français", **Travaux de l'Institut de Phonétique de Strasbourg**, n° 12; 177-224, 1980.

[13] Fink, B.R. & Demarest, R.J., **Laryngeal biomechanics**, Harvard University Press, Cambridge, Mass. and London, England, 1978.
[14] Collier, R., "Phonological explanations of F0 declination", **Proc. 10th Int. Cong. Phon. Sci.**, (Utrecht), 354-360, 1983.
[15] Di Cristo, A., **De la microprosodie à l'intonosyntaxe**, thèse d'Etat, Univ. de Provence, 1985.
[16] Rossi, M. & Autesserre, D., "Movements of the hyoid bone and the larynx and the intrinsic frequency of vowels", **J. of Phonetics**, 9 : 233-249, 1981.
[17] Delattre, P., "Les dix intonations de base du français", **French Review**, 40 : 1-14, 1966.
[18] Di Cristo, A. & Hirst, D., "Modelling French micromelody" (to be published in **Phonetica**), 1986.

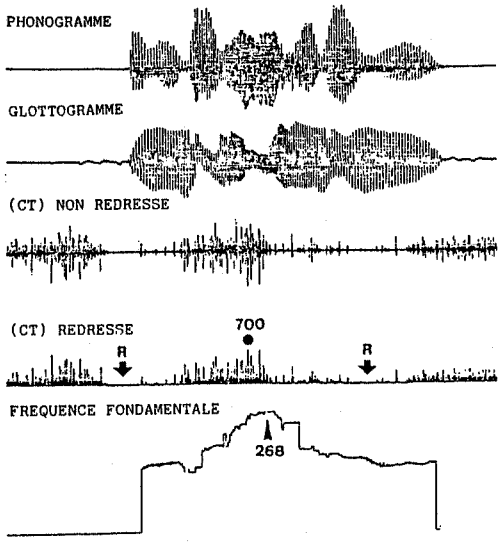


Figure 2 : Phase de relâchement (R) : initiale voisée (Implication : "Evidemment" & parenthèse basse : "je l'ai vu").

Figure 3 : Phase de relâchement : initiale non voisée (Question totale : "Sur le quai ?").

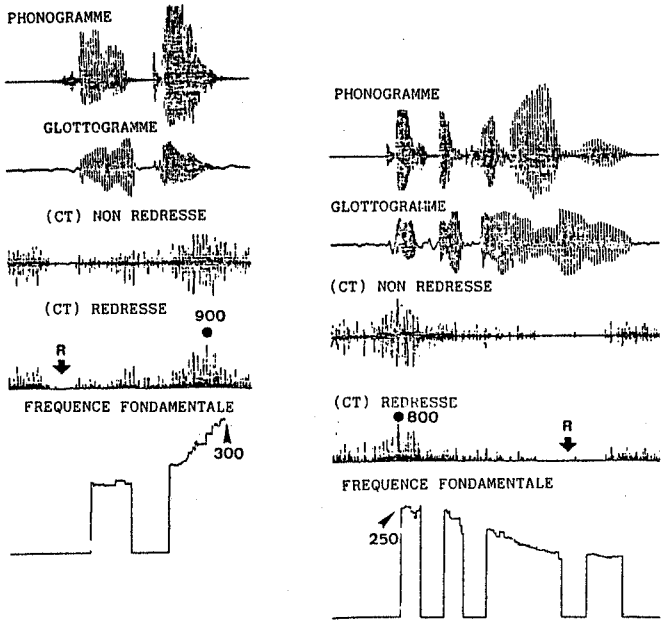


Figure 4 : Attaque dans le registre haut (Question partielle : "Qu'est-ce que tu lui as dit ?").

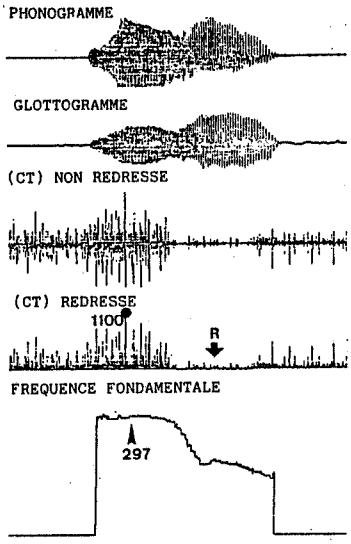


Figure 5 : Attaque dans le registre haut (Ton haut-bas : -mimi-).

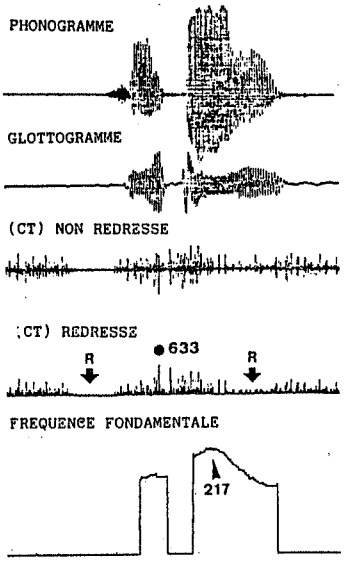


Figure 6 : Phase de relâchement dans un contour mélodique complexe (Appel : "Jean-Paul !").

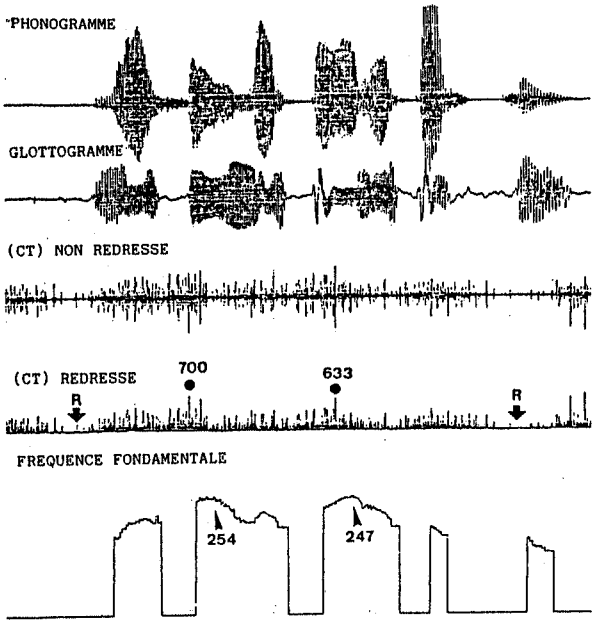


Figure 7 A : Activité du CT accompagnant la réalisation des intonations continuatives ("Le fils de Paul est parti ").

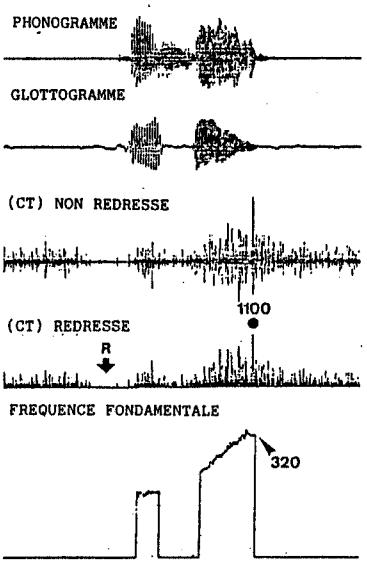


Figure 8 : Activité du CT lors de la réalisation d'une question totale ("Tu es sûr ?").

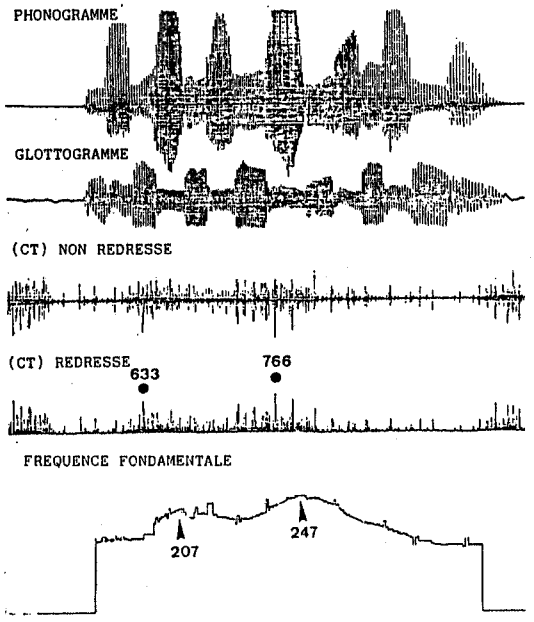


Figure 7 B : Version réitérée de la même phrase.

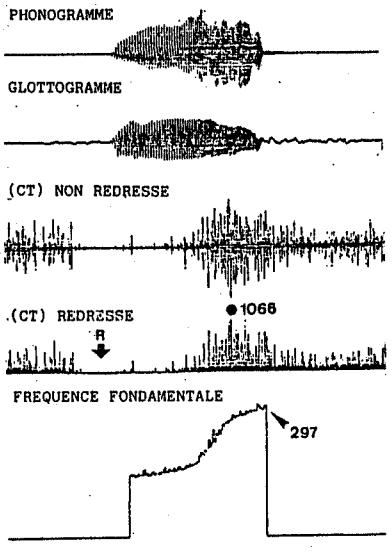


Figure 9 : Correspondance entre le pic d'amplitude du CT et le point d'inflexion de la configuration de la configuration de F0 (Ton bas-haut : -muu-).

ETUDE EMG PRELIMINAIRE SUR LE CONTROLE DE LA RESPIRATION DANS LA PHONATION

A. Marchal (+), Y. Jammes (++), Ch. Grimaud (++)

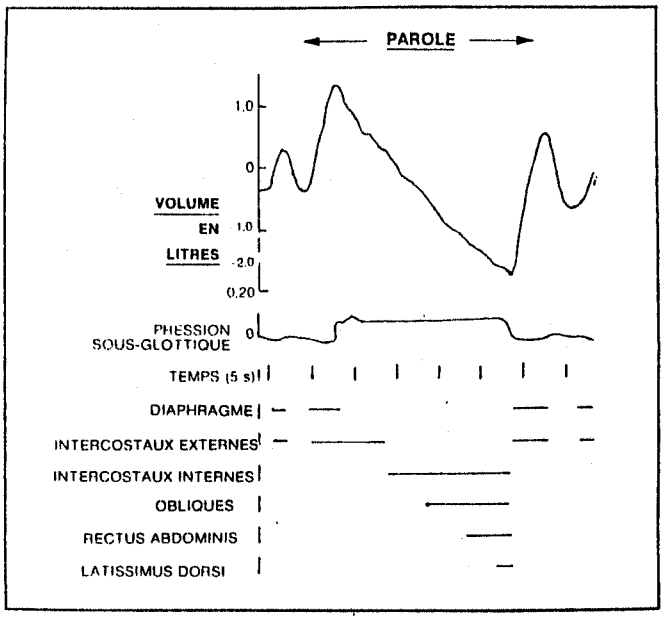
(+) Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'
(++) Département de médecine expérimentale, Faculté de Médecine, Marseille

ABSTRACT

This EMG study is primarily concerned with the activity of the internal and external intercostals, of the diaphragm and abdominal muscles during speech. Contrary to the prevailing theory [1], our data indicate that respiratory muscles act synergically during the entire expiratory phase. This result supports the claim of the existence of a ventilatory coordinative structure for speech.

INTRODUCTION

Les professeurs de chant ou de diction connaissent le rôle important joué par les poumons dans la phonation et font du contrôle de la respiration un exercice préalable à l'entraînement vocal. Le locuteur doit pouvoir faire varier la production d'air phonateur pour l'ajuster aux besoins du discours, composé de phrases plus ou moins longues dites avec une plus ou moins grande intensité et qui nécessitent un apport d'air différent.
Dans la respiration normale, l'expiration est un phénomène passif. Lorsque la contraction des muscles inspirateurs est terminée, la cage thoracique du fait de sa masse, de l'élasticité pulmonaire et de la pression excercée par les viscères abdominaux, diminue de volume, se rétrécit et, pressant sur les poumons, elle chasse une quantité d'air égale à celle qui était entrée dans le temps de l'inspiration.
Dans la parole, le mouvement expiratoire n'est plus automatique mais contrôlé. La période s'allonge et passe de 2-3 secondes à 15-20 secondes. La durée de l'expiration varie selon la longueur de la phrase. Le contrôle de l'expiration est nécessaire pour que la pression d'air sous la glotte demeure stable [2].
On est redevable à Stetson [3] d'avoir le premier proposé un certain nombre d'hypothèses sur l'activité des muscles de la respiration lors de la production de la parole. Les principaux résultats de ses travaux étaient les suivants :
1. Chaque syllabe est accompagnée par une poussée de la cage thoracique produite par l'action des muscles intercostaux internes.
2. Dans les syllabes ouvertes, l'abaissement des poumons est contrôlé par un effort inspiratoire des muscles intercostaux externes.
3. Dans une syllabe accentuée, l'action des intercostaux internes est renforcée par l'action des muscles abdominaux.
Fonagy [4] confirme les observations de Stetson sur les syllabes accentuées. Ladefoged [1] toutefois rejette catégoriquement toutes ces conclusions et propose le patron suivant de l'activité des muscles de la respiration (Fig. 1).



Activité des muscles de la respiration lors de la production de la parole (d'après Ladefoged, 1967)
FIG. 1

Au début de l'expiration, les muscles de l'inspiration (intercostaux externes) continuent d'être actifs pour ralentir la descente de la cage thoracique qui en raison de sa masse et des forces élastiques du poumon aurait tendance à s'effectuer trop rapidement. L'activité des muscles intercostaux externes décroît progressivement en fonction du volume pulmonaire pour cesser lorsqu'elle empêcherait la fourniture d'air phonateur. Ce n'est qu'à ce moment-là que les muscles intercostaux internes entrent en jeu. La contraction progressive des in-

tercostaux internes permet de comprimer la cage thoracique et de chasser l'air contenu dans les poumons. Leur action est renforcée en fin de phase expiratoire par le travail des muscles abdominaux.
Cette présentation de l'activité des intercostaux internes et externes apparaît suspecte en regard des observations systématiques sur le contrôle du mouvement par des muscles antagonistes.

ETUDE

Méthode

Nous avons enregistré simultanément (Fig. 2) les informations suivantes :
1. volume pulmonaire; 2. son; 3. activité des muscles : Diaphragme, Intercostaux internes (10e espace intercostal gauche), Intercostaux externes (3e espace intercostal gauche), Muscles abdominaux (pubis).
Les enregistrements EMG ont été obtenus à l'aide d'électrodes-aiguilles. Leur emplacement a été contrôlé en pratiquant plusieurs répétitions de différentes tâches incluant des changements de posture, des cycles de respiration normale, d'inspiration forcée, d'expiration forcée et d'apnée.
Les signaux physiologiques étaient d'abord filtrés (filtre passe-bas et soustraction des bruits cardiaques) avant d'être enregistrés sur un magnétophone d'instrumentation FM à 8 canaux. Un canal était réservé au son pour la synchronisation avec l'enregistrement acoustique réalisé simultanément sur une enregistreuse audio (Revox B, vitesse 19,5 cm). Pendant l'expérience et pour s'assurer du bon déroulement des opérations d'acquisition des signaux, on a procédé à l'inscription graphique des données suivantes :
1. diaphragme, brut
2. diaphragme, intégré
3. intercostaux internes, intégré
4. intercostaux externes, brut
5. intercostaux externes, intégré
6. abdominaux, brut
7. phonogramme
8. volume pulmonaire.

Une analyse ultérieure a consisté dans la production des données EMG brutes ou intégrées manquantes et dans l'extraction de la fréquence. L'analyse acoustique a produit les tracés synchronisés suivants :
1. oscillogramme; 2. intensité; 3. fréquence fondamentale.

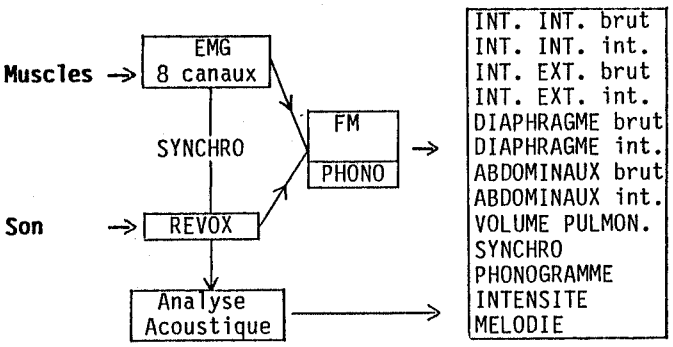


Diagramme expérimental.
FIG. 2

Corpus

Le corpus comporte 4 séries d'items :
1. Six groupes de cinq mots plurisyllabiques simples et composés.
2. Quarante logatomes du type a + c + a.
3. Dix phrases de longueur variable et de structure syntaxique plus ou moins complexe.
4. Des extraits de conversation spontanée sur la situation expérimentale.

Les trois premières séries ont fait l'objet de 15 répétitions dans un ordre aléatoire par 2 locuteurs hommes de 35 et 40 ans connus pour parler un français exempt de traits régionaux marqués et ne présentant aucune pathologie respiratoire.

Résultats

Il convient d'être toujours très prudent dans l'interprétation des tracés EMG. On sait en effet que le même mouvement peut être produit par deux muscles différents et qu'inversement une même activité musculaire ne provoque pas nécessairement le même résultat. Le rôle joué par le contexte et la position de départ est à cet égard critique. Dans le cas de cette étude, plusieurs résultats intermédiaires contradictoires avec les données publiées par Ladefoged nous ont invité à modifier la situation expérimentale pour nous assurer que les signaux recueillis provenaient effectivement des muscles dont on souhaitait étudier l'activité. Pour ce qui concerne en particulier l'activité des intercostaux internes, le risque de pneumothorax nous avait amené à n'enfoncer que trop superficiellement l'électrode. Après plusieurs vérifications, il nous semble que les résultats suivants peuvent être considérés comme des hypothèses valides.
Nos tracés laissent apparaître, pour les phrases et pour les 2 locuteurs, quatre phénomènes généraux :
1. Une activité générale précédant immédiatement la phonation ou dans une phrase après une longue pause.
2. Un patron régulier d'activité croissante du diaphragme et des intercostaux internes jusqu'à la fin de la phonation.
3. La présence de pics asynchrones d'activité du diaphragme et des intercostaux internes se superposant sur cette ligne générale.
4. Une augmentation de l'activité des intercostaux externes à la frontière des groupes de souffle correspondant à des constituants immédiats. Cette action a pour effet de provoquer l'apparition de plateaux sur la ligne de volume pulmonaire.

Pour ce qui concerne la question de la correspondance entre pics d'activité des intercostaux internes et la présence de syllabes, nous n'avons pu relever une telle correspondance que dans les cas de lecture avec un débit lent (liste de mots et logatomes) et de syllabes accentuées (emphase) dans les phrases. Il est toutefois apparu que cette question méritait une plus grande attention et que ce constat ne suffisait pas. Nos données indiquent en effet que les voyelles sont toujours marquées par un pic du diaphragme tandis que les consonnes sont caractérisées par une dépression du diaphragme et par une augmentation de l'activité des intercostaux externes. Dans un débit normal et pour des syllabes ouvertes, on constate qu'il y a une répartition quasi-syllabique des patrons secondaires d'activité EMG. Lorsqu'il y a présence de syllabes fermées ou de

groupes de consonnes, on observe pour de nombreux cas un pic suivant la consonne comme s'il s'agissait d'un "a" pourtant non attesté sur le tracé acoustique. Dans le cas de débit rapide, le groupe de consonnes ne correspond plus qu'à une seule dépression. Faut-il voir dans cette organisation diaphragme - intercostaux externes correspondant à l'organisation syllabique voyelle-consonne une structure de base privilégiée ? La question reste ouverte.
Nos résultats montrent clairement que l'activité des muscles de la respiration n'est pas aussi individualisée dans le temps que le laissait penser le schéma d'organisation de leur activité proposé par Ladefoged. Leur activité est synergique. Nos observations rejoignent donc celles d'Hoshiko [5] et d'Adams et Munro [6].

Conclusion

On peut interpréter ces résultats comme une confirmation des thèses sur le contrôle des mouvements coordonnés [7]. Les théories de l'action postulent en effet que l'activité des muscles antagonistes faisant partie d'une structure coordinative est complémentaire et que leur ajustement fin s'effectue de façon autonome en fonction des conditions périphériques. Dans le cas de la phonation, il semble que la structure coordinative respiratoire initiée au début de la phase (ligne de base correspondant aux groupes de souffle) réagisse localement (pics du diaphragme et des intercostaux externes) aux résistances apportées à l'écoulement de l'air au niveau de la glotte et au niveau de la cavité buccale (distinction sourd-sonore et division voyelle-consonne). Nous poursuivrons nos investigations dans ce domaine en essayant de voir comment des modifications inattendues de la charge résistive affectent le patron d'activité musculaire pendant la phonation.

BIBLIOGRAPHIE

[1] Ladefoged, P. (1967), **Three Areas of Experimental Phonetics**, Oxford University Press, London.
[2] Marchal, A. (1976), "Quelques notions de physiologie pulmonaire appliquées à la description phonétique de l'accent d'insistance en français", in (A. Seguinot, ed.) **L'accent d'insistance**, pp. 93-121, Didier, Montréal.
[3] Stetson, R. (1951), **Motor Phonetics, A Study of Speech Movement in Action**, N.H.C., Amsterdam.
[4] Fonagy, I. (1958), "Elektrophysiologische Beiträge zur Akzentfrage", **Phonetica**, 2 : 12-58.
[5] Hoshiko, M. (1960), "Sequence of action of breathing muscles during speech", **J. Speech & Hearing Res.**, 3 : 291-297.
[6] Adams, C. & Munro, R.R. (1973), "The relationship between internal intercostal muscle activity and pause placement in the connected utterance of native and non-native speakers of English", **Phonetica**, 28 : 227-250.
[7] Fowler, C.A., Rubin, P., Remez, R.E. & Turvey, M.T. (1980), "Implications for speech production of a general theory of action", in (Butterworth, ed.) **Language Production, vol. 1 : Speech and talk**, pp. 373-420, Academic Press, London.

PREMIERS ESSAIS D'UTILISATION DE L'AEROPHONOMETRE III POUR L'ETUDE DU SOUFFLE PHONATOIRE

Denis Autesserre & Bernard Teston

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

The technical characteristics of the latest Phonoaerometer offer new possibilities of research on speech and breathing.
This study illustrates the utilization of this device to examine distributional aspects of inspiratory and expiratory flow rates and volumes during the oral reading of a short text.

INTRODUCTION

L'étude des phénomènes aérodynamiques de la parole à l'Institut de Phonétique d'Aix a surtout porté sur les modifications à court terme de ces paramètres : qu'il s'agisse des variations du débit d'air buccal et nasal lors de la réalisation de voyelles cinétiques oro-nasalisées en français méridional [1] ou des relations entre la pression intra-orale et la fréquence fondamentale [2].
La mise au point d'un nouvel appareil, plus perfectionné, l'aérophonomètre [3] et [4], nous a permis d'entreprendre des recherches ayant trait à des séquences phoniques de plus grande extension temporelle, en relation avec les phénomènes respiratoires, débits et volumes d'air inspirés et expirés, lors de la respiration vitale mais aussi et surtout lors du "souffle phonatoire" (selon l'expression de F. Le Huche) [5].
Quels sont les liens qui peuvent exister entre les phases d'inspiration et les groupes de souffle lors de la production d'un discours suivi ? Les modifications de la morphologie des tracés de débit d'air oral et nasal, et les variations des volumes d'air entretiennent-elles des rapports de dépendance avec les phénomènes linguistiques et lesquels ? Nous n'avons pas la prétention d'apporter une réponse définitive à ces questions pour lesquelles l'accord est loin de se faire entre les divers chercheurs [6]. Nous voulons simplement présenter quelques premiers résultats, illustrant les nouvelles possibilités d'utilisation de l'aérophonomètre et la contribution que l'on peut en attendre pour améliorer notre connaissance des phénomènes aérodynamiques impliqués dans la production de la parole.

PROCEDURE EXPERIMENTALE
CHOIX DES CRITERES D'INTERPRETATION

1. Appareillage et paramètres enregistrés.
Deux enregistrements magnétiques sont conservés sur les deux pistes d'un magnétophone REVOX : le "son buccal" capté par l'intermédiaire d'un micro placé à l'intérieur de l'embouchure buccale et le "son laryngé" à l'aide d'un laryngophone (en vue d'analyses ultérieures de la courbe mélodique).
Les signaux aérodynamiques buccaux sont recueillis à l'aide d'une embouchure buccale spécialement adaptée à la morphologie de chaque locuteur : il s'agit du débit d'air buccal (DAB) et du volume d'air buccal expiré (VABE) ou inspiré (VABI). Deux embouts introduits dans chaque narine pour le débit d'air nasal (DAN) et les volumes d'air nasal expiré (VANE) et inspiré (VANI) complètent l'analyse. Le phonogramme buccal permet la délimitation temporelle des différentes séquences. Tous ces paramètres sont enregistrés sur un oscillographe SIEMENS (à la vitesse de défilement de 100 mm/s; Fig. 1 et 2).

2. Sujets et corpus.
Dix sujets (5 hommes et 5 femmes) se sont prêtés à plusieurs expériences successives et toujours dans le même ordre : il leur est demandé d'effectuer, à l'intérieur de l'embouchure buccale, une respiration buccale profonde, une respiration nasale profonde, et une respiration naso-buccale (inspiration par le nez et expiration par la bouche). Ces épreuves respiratoires sont suivies d'épreuves phonatoires : d'abord, prononciation d'une voyelle [a] tenue jusqu'à l'extrême limite du souffle expiré (en essayant de maintenir l'intensité de la voix à un niveau constant, contrôlé à l'aide du VU-mètre du magnétophone REVOX : ceci afin de déterminer, pour chaque locuteur, le temps maximum de phonation (TMP) [7]. L'enregistrement est interrompu un moment et le sujet lit alors un texte de Claude Simon, extrait de "Le route des Flandres", Editions de Minuit, 1960, p. 63. Ce passage comporte deux longues phrases : "Ils regardèrent ... retroussées" et "Il n'y avait ... humide". On arrête alors l'enregistrement et le locuteur, lorsqu'il le désire, aborde une deuxième série avec les mêmes épreuves et selon un ordre identique. L'épreuve se termine par la lecture de 5 séries complémentaires de voyelles [a] tenues (afin de comparer nos résultats avec ceux de D.E. Finnegan obtenus sur des enfants) [7].

3. Dépouillement et mesures.
Les mesures sont obtenues, manuellement, à partir des 7 tracés des oscillogrammes. Elles concernent les paramètres suivants : (Fig. 1 et 2)
- durée des phases d'inspiration et d'expiration lors des épreuves respiratoires,
- durée des phases d'inspiration, ou prise de souffle (PS) et d'expiration ou groupe de souffle (GS) pour le souffle phonatoire.
Par segmentation conjointe du signal acoustique, on délimite les séquences phoniques associées à PS et GS et l'on isole aussi les pauses non inspiratoires.
- durée de la tenue de la voyelle [a] (TMP) (en ms),
- durée totale de la séquence parlée,
- mesure des volumes d'air expiré et inspiré aux frontières précédemment délimitées (en c.c.),
- mesure du volume d'air expiré pour la voyelle [a] (en c.c.),
- localisation et mesure de la déflexion maximale lors des prises de souffle buccales et nasales (en l/mn),
- analyse morphologique des tracés de VABI et VANI.
Une prise de souffle associée à un groupe de souffle subséquent constitue un cycle de souffle phonatoire (CSP).

RESULTATS

Il ne sera pas possible de présenter un grand nombre de résultats, le dépouillement des données n'étant pas encore terminé. Nous avons choisi quelques résultats préliminaires concernant l'analyse du texte lu.

1. Une première constatation s'impose : pour un même locuteur, il n'y a pas deux séries qui soient réalisées avec une même distribution des prises de souffle et des groupes de souffle. Pour le locuteur A.D.C. (Fig. 3), le nombre des prises de souffle, à l'intérieur du texte, varie entre 13 et 18. En ajoutant la totalité des pauses présentes (qu'il y ait ou non activité inspiratoire), on remarque que, toutes séries confondues, 22 virtualités de pauses existent et elles peuvent toutes comporter une prise de souffle lors d'une des 5 lectures du texte. D'autre part, 9 prises de souffle se situent toujours au même endroit, sur un total variant de 13 à 18.

2. La durée des groupes de souffle va de 505 ms ("osseuse") jusqu'à 3532 ms ("ils regardèrent ... écurie"), pour la locutrice N.B.
La prise de souffle la plus brève (180 ms) se situe après la séquence "se détachant sur le fond clair de la porte" (11 syllabes), elle-même suivie de la séquence "comme une sorte de brouillard légèrement bleuté" (12 syllabes).
La prise de souffle la plus longue (505 ms) après "embuer le doux regard de cyclone", ne contient que 9 syllabes et se trouve tout de suite avant le dernier groupe de souffle "accusateur et humide". Ces quelques coups de sonde montrent combien il sera difficile d'établir une relation directe entre la durée des prises de souffle et la durée des groupes de souffle adjacents.

3. L'amplitude des déflexions de DAB est maximale pour la locutrice C.B. après "soulever" et avant "osseuse" (129 l/mn) et non pas après "retroussées" qui termine la première phrase (55 l/mn - mais une durée de 770 ms -). Il ne paraît pas y avoir pour les deux locutrices N.B. et C.B., de relation étroite entre le nombre de syllabes qui précède (ou qui suit) la prise de souffle et la durée (ou l'amplitude négative) de DAB contrairement à ce que laissaient attendre d'autres travaux [6], [8].
Il y a bien, entre ces divers paramètres, des liens qui font intervenir une préprogrammation syntaxique, ou dans des conditions plus naturelles de communication, un temps d'élaboration du contenu sémantique, mais les recherches psycholinguistiques ne distinguent pas souvent les pauses avec ou sans activité inspirative.

4. Le volume d'air inspiré total (VABI + VANI) atteint sa valeur la plus forte (653 c.c. chez la locutrice N.B.) après "comme une sorte de brouillard légèrement bleuté" et avant "comme un voile". La valeur la plus faible se situe entre "se détachant ... porte" et "comme ... bleuté" (180 c.c.).

DISCUSSION

1. La durée des prises de souffle ou des maxima d'amplitude négative ne sont pas forcément corrélés à la durée des groupes de souffle qui précédent ou qui suivent.

2. Cependant, certains faits, plus locaux, tels que le "plateau" qui existe entre la fin de la remontée du DAB et la voyelle en position initiale, laissent prévoir un ajustage précis en début de groupe de souffle.

3. De même, avant une prise de souffle qui marque une frontière syntaxique majeure, la morphologie du DAB se modifie et l'on voit apparaître un "plateau", de durée supérieure à celle que l'on avait précédemment, avant que la courbe ne devienne franchement négative. D'autre part, les très rares fois où la fin du texte a été suivie d'une phase inspiratrice, celle-ci est en partie isolée là encore par un palier avant la chute brutale du débit. Des faits de ce type devraient être analysés de près et mis en relation avec des analyses de la pression sous-glottique et aussi des signaux électromyographiques sur le modèle de [9]. Il arrive même que dans un nombre non négligeable de cas l'on assiste à une augmentation conjointe du volume d'air buccal et nasal. N'y aurait-il pas aussi un "élan" expiratoire en fin d'énoncé ? [5]

CONCLUSION

Cette première prise de contact avec l'étude des phénomènes respiratoires est assez déconcertante : grande plasticité dans certains cas (durées relatives, groupes et prises de souffle), chronométrage précis pour des faits locaux qui touchent de plus près aux coordinations pneumo-phoniques.
On pourrait parler de contrôle souple. Nous laisserons d'ailleurs le mot de la fin au Dr J. Tarneaud, qui dans la préface de la thèse de G. Cornut [10], utilise une expression qui préfigure l'un des "thèmes obsessionnels" chers à Claude Simon, celui

du "cheval" ! "Je ne mets pas en doute la difficulté d'assimiler la discipline de l'expiration phonique, car pour bien des praticiens et la presque totalité des maîtres enseigneurs en chant ou en déclamation, l'attelage [!] pneumo-phonique est une fonction naturelle, innée ou acquise, que l'exercice développe sans obliger à technique et contrôle [...] La respiration en vue de la phonation est toute différente de la respiration normale habituelle. En effet, prédominent dans la mécanique vocale la pression expiratoire et le débit d'air, variant tous deux à la demande de la qualité sonore émise".

BIBLIOGRAPHIE

-[1] Teston, B. & Autesserre, D., "Réalisation d'une unité d'analyse polyphonométrique", **C.L.O.S.**, 5-6, Hommage à Georges Mounin, 415-437, 1975.

-[2] Di Cristo, A. & Teston, B., "Fréquence fondamentale et pression intraorale", **Larynx et Parole**, GALF, Institut de Phonétique de Grenoble, 329-366, 1979.

-[3] Teston, B., "A system for the analysis of the aerodynamic parameters of speech : the polyphonometer model III", **Abstracts of the Tenth International Congress of Phonetic Sciences**, Foris Publications, Dordrecht, Holland, Cinnaminson, U.S.A., 457, 1983.

[4] Teston, B., "Un système de mesure des paramètres aérodynamiques de la parole : le polyphonomètre modèle III", **Travaux de l'Institut de Phonétique d'Aix**, vol. 9, 373-383, 1984.

[5] Le Huche, F. & Allali, A., **La voix**, vol. 1, Masson, Paris, 1984.

[6] Horii, Y. & Cooke, P., "Some airflow, volume and duration characteristics of oral reading", **Journal of Speech and Hearing Research**, 21, 470-481, 1978.

[7] Finnegan, D.E., "Maximum phonation time for children with normal voices", **Folia Phoniatrica**, 37, 209-215, 1985.

[8] Webb, R., Williams, F., Minifie, F.D., "Effects of verbal decision behavior upon respiration during speech production", **Journal of Speech and Hearing Research**, 10, 49-56, 1967.

[9] Draper, Ladefoged, P. & Whitteridge, D., "Expiratory pressures and airflow during speech", **British Medical Journal**, 264, 1837-1843, 1960.

[10] Cornut, G., **La mécanique respiratoire dans la parole et dans le chant**, P.U.F., Paris, 1959.

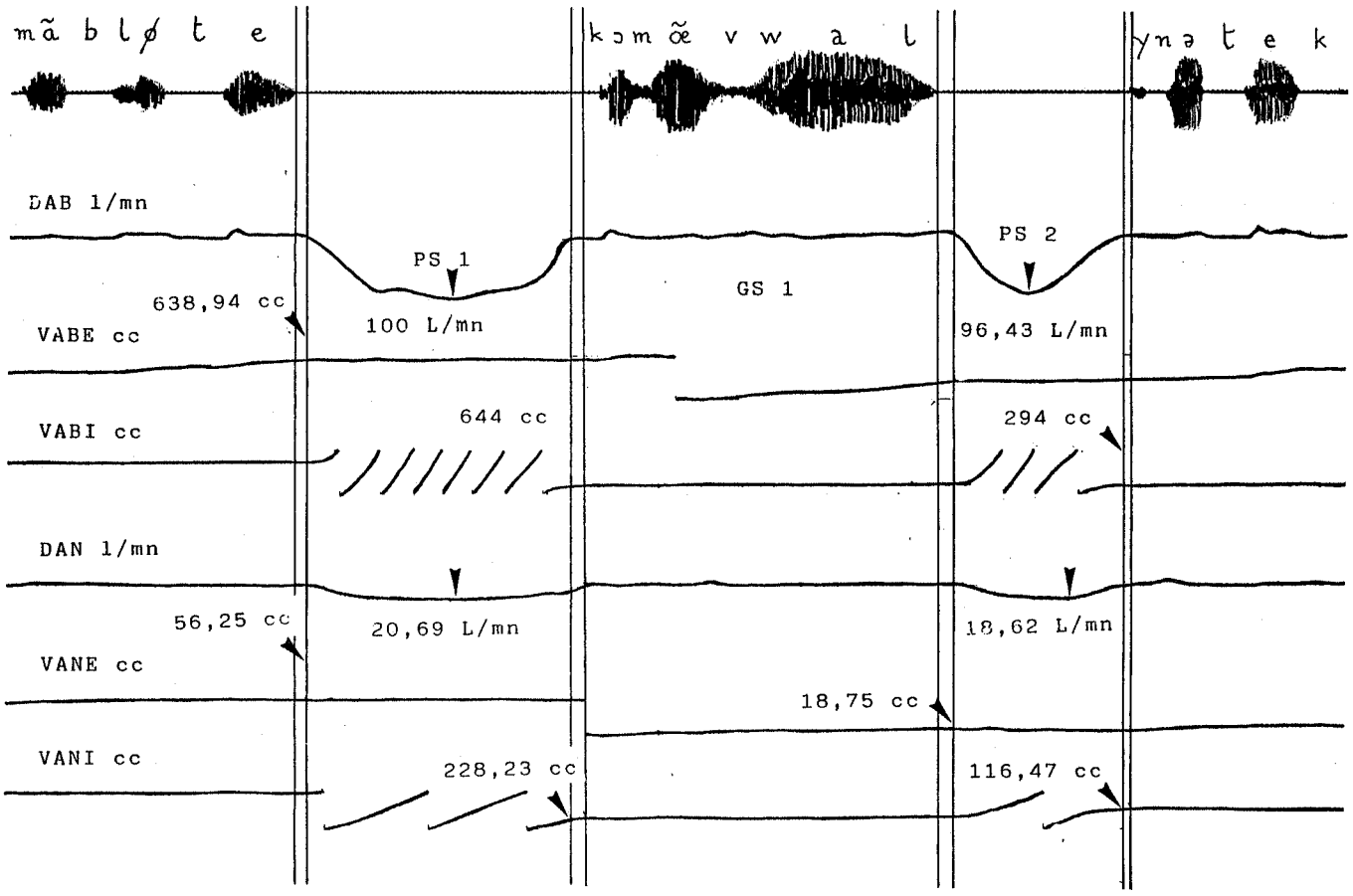


Figure 1 : Représentation des sept paramètres physiologiques étudiés (locutrice C.B.).

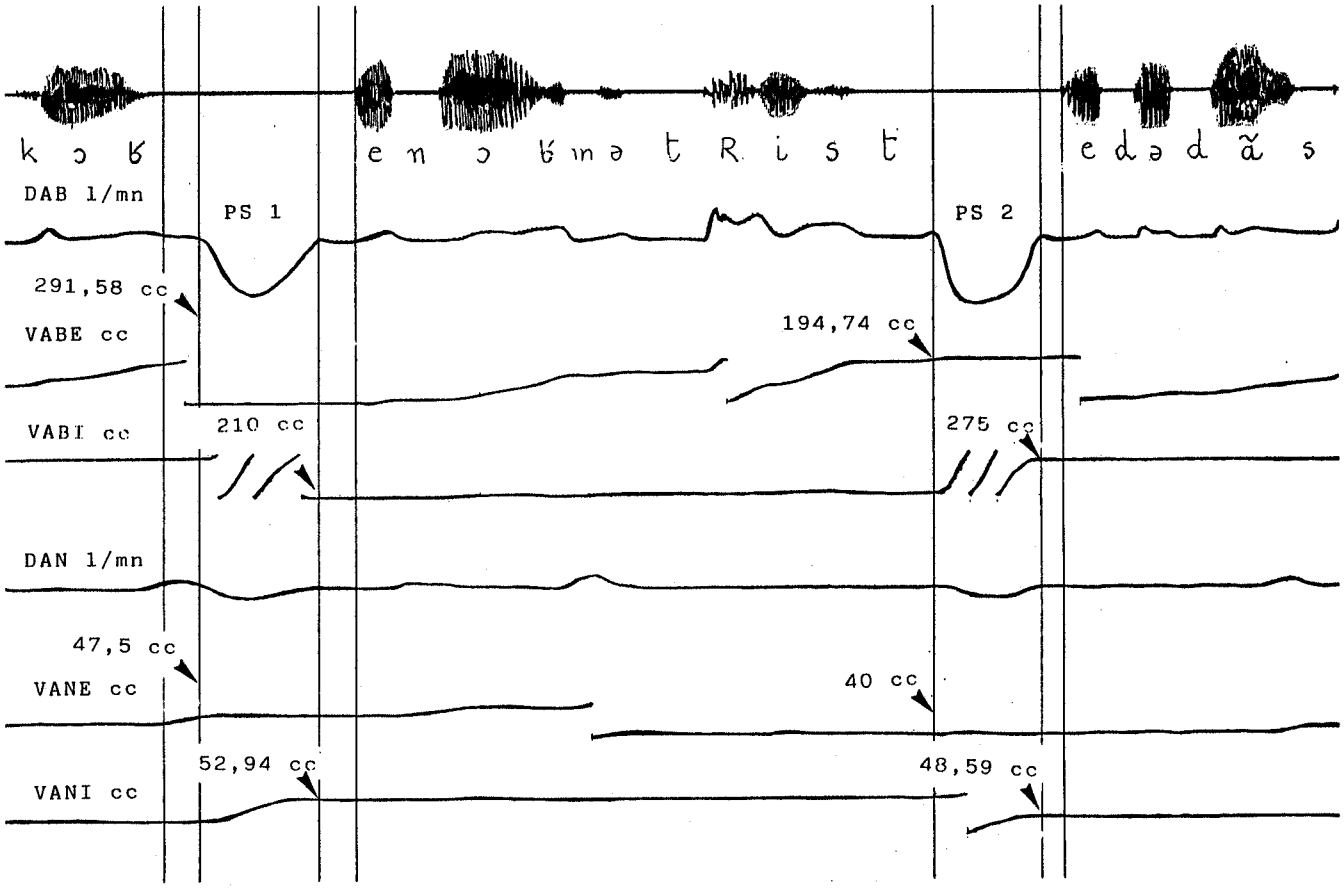


Figure 2 : Illustration d'un groupe de souffle GS encadré par deux prises de souffle PS1 et PS2 de morphologie différente (locutrice N.B.).

Ils regardèrent le cheval toujours étendu sur le flanc au fond de l'écurie /; on avait jeté une couverture dessus et seuls dépassaient ses membres raides /, son cou terriblement long / au bout duquel pendait la tête / qu'il n'avait même plus la force de soulever /, osseuse /, trop grosse avec ses méplats, son poil mouillé /, ses longues dents jaunes / que découvraient les lèvres retroussées /. Il n'y avait que l'œil qui semblait vivre encore / /, énorme, triste /, et dedans, sur la surface luisante et bombée /, ils pouvaient se voir /, leurs silhouettes déformées comme des parenthèses / se détachant sur le fond clair de la porte / comme une sorte de brouillard légèrement bleuté /, comme un voile / /, une taie / qui déjà semblait se former /, embuer le doux regard de cyclone /, accusateur et humide.

A.D.C. 1

Ils regárdèrent le cheval toujours étendu sur le flanc au fond de l'écurie /; on avait jeté une couverture dessus et seuls dépassaient ses membres raides /, son cou terriblement long / / au bout duquel pendait la tête qu'il n'avait même plus la force de soulever /, osseuse, trop grosse avec ses méplats /, son poil mouillé, ses longues dents jaunes que découvraient les lèvres retroussées /. Il n'y avait que l'œil qui semblait vivre encore / /, énorme /, triste /, et dedans, sur la surface luisante et bombée /, ils pouvaient se voir /, leurs silhouettes déformées comme des parenthèses se détachant sur le fond clair de la porte / comme une sorte de brouillard légèrement bleuté /, comme un voile /, une taie qui déjà semblait se former /, embuer le doux regard de cyclone / /, accusateur et humide.

A.D.C 3

Figure 3 : Localisation des pauses non inspiratoires / /, et des prises de souffle / pour deux versions différentes du texte de Claude Simon (lu par le locuteur D.C.A.).

PRODUCTION

Président

Christian ABRY

Institut de la Communication parlée de Grenoble

DISTANCE ET DUREE DES MOUVEMENTS DU DOS DE LA LANGUE EN PRODUCTION DE LA PAROLE: RESULTATS D'UNE ANALYSE FACTORIELLE

Eric Keller

Centre de recherche, Centre Hospitalier Côte-des-Neiges, Montréal et
Département de linguistique, Université du Québec à Montréal

ABSTRACT

A factor analysis of 1353 observations of eleven measures of lingual activity involved in the production of the syllable [ka] showed that three stable factors explain about 75% of the variance. Measures of tongue displacement and velocity loaded strongly on the first factor, durational measures loaded strongly on the second factor, and mid-syllable durational and distance measures loaded on the third factor. Canonical correlations showed that the three best predictors were (1) descending movement displacement, (2) linguo-laryngeal movement onset delay (comparable to VOT) and (3) ascending movement onset duration. These variables probably represent control over phonemic distinction, inter-articulatory coordination and syllable length modulation. The first two factors correspond to central variables of breakdown in Broca's aphasia.

INTRODUCTION

L'objet de cette étude est de présenter une approche empirique au problème de la sélection de variables articulatoires. Idéalement, ces variables devraient être reproduisibles autant pour un sujet que pour plusieurs; elles devraient également prédire le comportement de variables associées et capter les aspects essentiels de la performance articulatoire aussi bien chez les sujets normaux que chez les patients atteints d'une pathologie de la parole. Au moyen d'un système informatisé de mesures ultrasoniques [5], nous avons recueilli 1353 observations des mouvements du dos de la langue dans la syllabe [ka]. Des analyses factorielles ont montré que trois facteurs expliquent environ 75% de la variance de ces données. Ces facteurs prédisent certains aspects importants de la production normale ainsi que des perturbations qui caractérisent l'aphasie de Broca.

METHODE

Sujets et stimuli: Onze locuteurs natifs francophones, âgés de 20 à 65 ans, 6 femmes et 5 hommes, ont prononcé au moins 25 exemplaires de la syllabe [ka] dans quatre conditions: (1) [ka] long répété sans contexte, (2) [ka] court répété sans contexte, (3) [ka] long dans le contexte "le macaque assommé" et (4) [ka] court dans le contexte "le lac à canards".

Enregistrement: Pour chaque enregistrement de 4,5s, un transducteur ultrasonique placé en position verticale au-dessous du menton a suivi le déplacement vertical du dos de la langue. Cette information et le signal acoustique simultané ont été transféré à un ordinateur au rythme de 1 kHz. Les tracés ultrasoniques ont été lissés par des fonctions cubiques spline.

Analyse: Pour chaque observation, onze mesures ont été prises en considération: (1) quatre mesures de distance verticale dans les mouvements linguaux descendants et ascendants pour la syllabe [ka], (2) quatre mesures de durées associées à ces mouvements, (3) le délai entre l'initiation des mouvements lingual et laryngaux (comparable au DEV ou VOT), et (4) les vélocités maximales descendante et ascendante. Avant de combiner les données des onze sujets, les observations ont été standardisées en cote z et les valeurs excédant $\pm$ 3,0 é.t. ont été supprimées.

RESULTATS

Analyse factorielle: Dans chaque condition, l'analyse a fourni trois facteurs qui expliquent entre 68,6 et 77,4% de la variance des données (limitation du nombre de facteurs par le critère de valeur Eigen >1,0 après une rotation Varimax). Les variables qui contribuaient à un niveau >0,5 sur le premier facteur étaient exclusivement les distances de déplacement ($p<0,01$). Les variables qui contribuaient à un niveau >0,5 sur le deuxième facteur étaient exclusivement les durées associées aux distances du facteur 1 et le DEV ($p<0,01$) (sauf pour la durée de l'initiation du mouvement ascendant). Les deux variables qui contribuaient à un niveau >0,5 sur le troisième facteur étaient la durée de l'initiation du mouvement ascendant et la distance correspondante.

Nous interprétons ces résultats en ce sens que les aspects de distance du premier facteur sont probablement reliés à la distinction phonémique de la syllabe, car des distances moindres dans la descente de [k] à [a] (la distance dominante) produiraient des syllabes telles que [kɔ] ou [kɛ]. Les aspects de durée, eux, sont avant tout reliées à la coordination interarticulatoire, le DEV étant la variable dominante pour ce facteur (facteur 2). La durée de l'initiation du mouvement ascendant (facteur 3)

est probablement reliée à la modulation de la mi-syllabe servant à distinguer les syllabes longues de syllabes courtes, car dans des syllabes longues, nous avons souvent observé des "plate-formes" de prolongation à ces endroits qui n'étaient pas visibles dans les syllabes courtes.

Corrélations canoniques: Une analyse de quatre modèles linéaires additifs du type "$\text{Variable}_{1...8}$ = variable x + variable y + variable z" a été effectuée. Un modèle joignant la distance du mouvement descendant (variable associée au facteur 1) au DEV (variable associée au facteur 2) et à la durée de l'initiation du mouvement ascendant (variable associée au facteur 3) était capable d'expliquer en moyenne 45% de la variance dans les huit variables restantes. Ces trois variables clés sont illustrées dans la figure 1. Parmi quatre modèles théoriquement et empiriquement possibles, celui-ci constituait le modèle le plus explicite. Ce résultat appuie donc les conclusions que nous avons déduites de l'analyse factorielle.

DISCUSSION

Les deux premiers des trois facteurs qui ont été extraits dans cette étude semblent directement reliés à des perturbations connues de l'aphasie de Broca. Les spécifications de distance qui caractérisent le premier facteur, sont probablement perturbées quand ces aphasiques substituent une voyelle pour une autre [4]. Or les substitutions phonémiques représentent un des signes les mieux connus de l'aphasie de Broca [1, 4].

La coordination interarticulatoire, captée dans le deuxième facteur, se trouve également perturbée régulièrement en aphasie de Broca, car ces patients se distinguent d'autres types d'aphasiques par leurs perturbations du DEV [2, 3]. Nous trouvons donc que ces résultats, provenant de sujets normaux, nous fournissent des indications fort utiles non seulement pour les facteurs sous-jacents à l'articulation normale, mais aussi à la perturbation de la parole en aphasie.

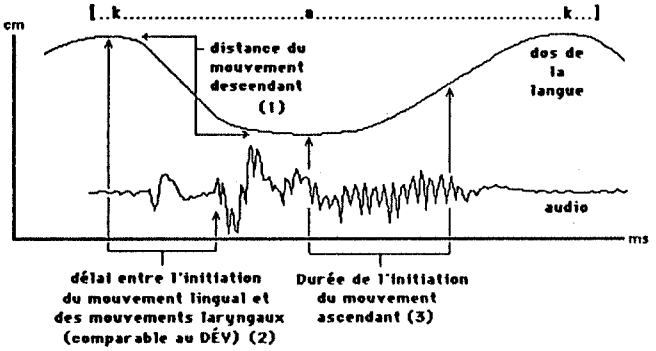


Figure 1. Illustration des trois meilleurs prédicteurs du mouvement vertical du dos de la langue dans la production de la syllabe [ka]. Les facteurs associés à ces prédicteurs sont indiqués entre parenthèses.

REFERENCES

[1] Blumstein, S. (1973). A phonological investigation of aphasic speech. La Haye: Mouton.

[2] Blumstein, S.E., Cooper, W.E., Goodglass, H., Statlender, S., & Gottlieb, J. (1980). Production deficits in aphasia: A voice-onset time analysis. Brain and Language, 9, 153-170.

[3] Itoh, M., Sasanuma, S., Tatsumi, I.F., Murakami, S., Fukusako, Y., & Suzuki, T. (1982). Voice onset time characteristics in apraxia of speech. Brain and Language, 17, 193-210.

[4] Keller, E. (1978). Parameters for vowel substitutions in Broca's aphasia. Brain and Language, 5, 265-285.

[5] Keller, E., & Ostry, D.J. (1983). Computerized pulsed echo ultrasound measurements of tongue dorsum movements. Journal of the Acoustical Society of America, 73, 1309-1315.

EXTRACTION AUTOMATIQUE DE LA FONCTION D'AIRE DES COUPES RADIOGRAPHIQUES

O. Hassan, B. Guérin, P. Perrier

INSTITUT DE LA COMMUNICATION PARLEE (UA CNRS 368)
I.N.P. Grenoble, 46 Avenue Félix Viallet 38031 GRENOBLE Cedex

ABSTRACT

The purpose of this paper is to introduce an algorithm to calculate the vocal tract area function automatically using radiographic data.

The mid-line of the vocal tract is defined as the locus of centers of the circles interiorly tangent to its walls in the mid-sagittal plane. Meanwhile, the transversal distance is considered to be the distance between each two points of tangence of this circle.

Finally, the results obtained are compared with another set of results determined by a semi-automatic method.

INTRODUCTION

Le point de départ de la modélisation acoustique du processus de production de la parole est, bien souvent, constitué par un ensemble de tracés de coupes sagittales du conduit vocal déduits de radiographies. C'est à partir de ces données qu'il faut déterminer les fonctions d'aire, puis les fonctions de transfert correspondantes.

La première étape consiste alors à extraire de ces données les valeurs des distances sagittales en différents points du conduit vocal. Cette opération a jusqu'à présent souvent été manuelle, et notre propos est , ici, d'exposer une technique permettant de l'automatiser. Pour cela, nous recherchons tous les cercles intérieurs et tangents au contour, puis, nous déterminons la ligne médiane du conduit vocal, comme étant le lieu des centres des cercles. La distance sagittale est enfin mesurée comme étant la distance entre les deux points de tangence d'un même cercle avec le contour. La fonction d'aire est ensuite calculée à l'aide d'un jeu de coefficients prenant en compte la position de la section calculée le long du conduit vocal.

Dans la suite nous décrirons plus en détail la méthode utilisée pour déterminer la ligne médiane, nous indiquerons le principe retenu pour calculer la valeur de la distance sagittale et nous préciserons les hypothèses faites pour le passage à la fonction d'aire. Nous comparerons pour conclure les résultats ainsi obtenus avec ceux qui ont été donnés par une méthode semi automatique.

LES DONNEES DE DEPART

Les coupes sagittales ont été acquises par cinéradiographie (à la cadence de 60 images/s) pour des transitions maximales comme (api), (apu), (ati), (ipa) prononcées par un locuteur de langue anglaise (les enregistrements ont été effectués par Thomas Gay du département de biologie orale de Farmington(1)).

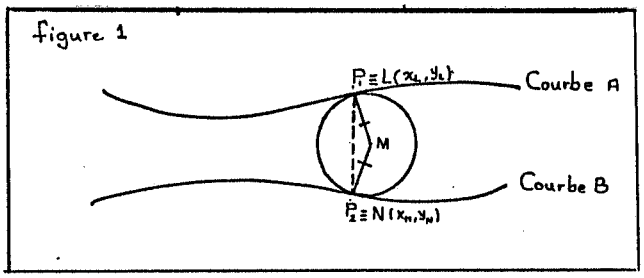
L'acquisition du contour à partir de la radiographie et son stockage sous forme numérique dans un fichier ont été effectués de manière manuelle. Par la suite, jusqu'à la détermination de la fonction d'aire , tout sera automatique.

EXTRACTION DE LA LIGNE MEDIANE

Principe de l'analyse .

La contour ainsi acquis ne présente aucun point de réference permettant de déterminer simplement et automatiquement la ligne médiane. Il nous faut donc nous fixer une stratégie, et pour cela nous avons utilisé la définition suivante :
"La ligne médiane est le lieu de tous les points qui sont à distance égale des deux parois du conduit vocal." La distance entre un point et une courbe étant déterminée par la longueur de la perpendiculaire abaissée de ce point à la courbe, nous pouvons affiner la définition :
"La ligne médiane est le lieu des centres des cercles tangents et intérieurs aux parois du conduit vocal "(2)(figure 1).

figure 1



Plaçons nous alors dans le plan de la coupe sagittale et imaginons un cercle qui roule à l'intérieur du conduit vocal et dont les dimensions sont régulièrement adaptées aux

dimensions internes du conduit vocal de façon à ce qu'il touche toujours les deux parois. On mesure alors la distance sagittale du conduit vocal comme la distance séparant les deux points de tangence, et la ligne médiane sera déterminée par le lieu des centres des cercles.

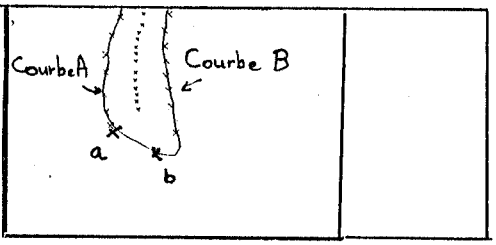
On pourrait aussi définir la ligne médiane selon un autre critère : courbe pour laquelle n'importe quelle ligne lui étant perpendiculaire coupe les deux parois du conduit vocal en des points qui lui sont équidistants. Pour vérifier la validité de ce critére, on considére un tel point, et on étudie la relation entre ce point seul et les parois du conduit vocal. Plusieurs courbes répondent alors à ce critère et des contraintes supplémentaires sont nécessaires pour choisir celle qui sera retenue. La première méthode nous parait devoir déboucher sur une solution plus unique, aussi l'avons nous retenue.

Algorithme retenu

L'algorithme utilisé pour rechercher la ligne médiane est le suivant(2):

1) On divise la courbe représentant le conduit vocal en deux parties: une partie haute (la courbe A) et un partie basse (la courbe B). En effet, à l'acquisition, on repère la position du bas larynx par deux points a et b (figure 2), déterminant respectivement le début du contour A et le début du contour B.

figure 2



Sachant que l'acquisition des points de chacun des contours est séquentielle, par la recherche du point le plus proche de a (respectivement de b), on caractérisera le contour A (respectivement le contour B).

Pour chaque point M sur la courbe A on calcule :

a) La pente de la perpendiculaire LM=m, la valeur de 'm' est :

$$m = -1/(dy/dx) \qquad (1)$$

où dy/dx : pente de la tangente de courbe A au point 'L'. La pente de la tangente est obtenue par la différentiation de la fonction décrivant la courbe A. Puisque la courbe A est exprimée sous forme d'un tableau de valeurs à deux dimensions (3), une différentiation numérique est utilisée.

Alors, on peut représenter la ligne LM sous la forme :

$$Y_L = mX_L \qquad (2)$$

où $c = Y_L - mX_L$ (3)

où Y_L et X_L sont les coordonnées du point 'L'.

3) On cherchera Mn telle que :

a) N se trouve sur la courbe B.

b) Le pente de MN est donnée par m' où

$$m' = -1/(dz/dx) \qquad (4)$$

dz/dx : pente de la tangente du courbe B au point 'N'.

c) La longueur de LM est égale à celle de MN.

donc:

$$Y_M = m'X_M + c' \qquad (5)$$

On aura alors:

$$Y_M = m'X_M + c' = mX_L + c \qquad (6)$$

où X_M et Y_M sont les coordonnées du point 'M', centre du cercle tangent aux courbes A et B.

Il vient:

$$c' = (m-m')X_M + c \qquad (7)$$

où X_M est l'abcisse du point 'M'

On a aussi :

$$Y_N = mX_N + c' \qquad (8)$$

$$Y_N = m'X_N + (m-m')X_M + c \qquad (9)$$

On pourra donc exprimer les coordonnées du point 'M' comme :

$$X_M = (Y_N - c - m'X_N)/(m-m') \qquad (10)$$

$$Y_M = mX_M + c \qquad (11)$$

La condition c) est satisfaite si on vérifie:

$$(X_M-X_N)^2 + (Y_M-Y_N)^2 = (X_L-X_M)^2 + (Y_L-Y_M)^2 \qquad (12)$$

Par substitution des équations (10) et (11) dans (12) et par la recherche de la minimisation de la valeur de err donnée par:

$$err = (X_M-X_N)^2+(Y_M-Y_N)^2-(X_L-X_M)^2-(Y_L-Y_M)^2 ,$$

on obtiendra les valeurs des coordonnées des centres des cercles tangents tout le long du conduit vocal.

Détermination de la coupe sagittale et de la fonction d'aire

Pour définir les points qui constituent la ligne médiane, on compare chacune des points de la partie haute de la courbe, avec un groupe des points de la partie basse de la courbe. Une fenêtre d'analyse détermine ce groupe de points. D'autre part les distances sagittales sont calculées le long du conduit vocal comme étant la distance entre les points de tangence d'un même cercle avec le contour A et le contour B (2) (figure 1).

La dernière étape de notre acquisition consiste à passer de la coupe sagittale à la fonction d'aire. Pour cela nous utilisons les résultats de Sanchez et al.(4-7)

L'étude d'un moulage du conduit vocal a permis d'y déterminer cinq grandes zones (larynx, bas pharynx, haut pharynx, bouche et zone

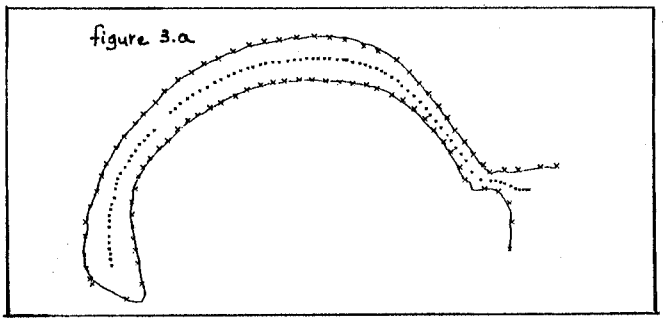
alvéolaire), et dans la zone 'j' l'aire A_{ij} correspondant à la distance sagittale d_{ij} est donnée par :

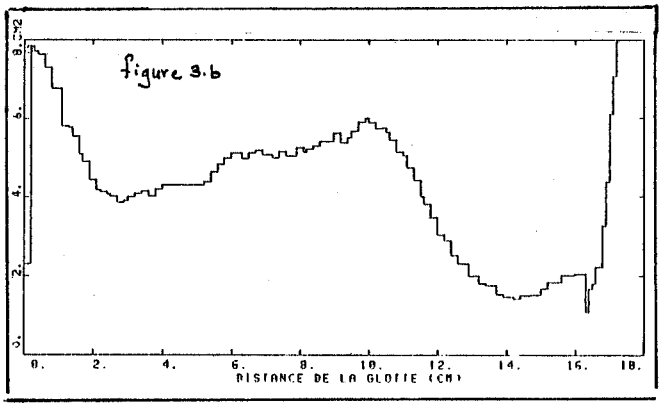
$$A_{ij} = C_j \, (d_{ij})^{1.5}$$

PREMIERS RESULTATS

La programme ainsi réalisé permet de passer du tracé sagittal brut à la fonction d'aire en moins de quatre minutes (sur Digital LSI 11.73). Cela reste certes encore long, mais cela représente d'ores et déja un progrès considérable par rapport à la technique manuelle faisant appel au crayon, à la gomme, au double décimètre et... à beaucoup de patience.

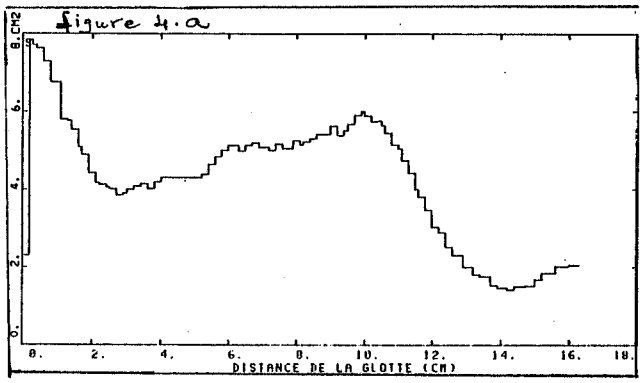
Examinons maintenant les résultats que nous obtenons. Un exemple en est donné figure 3 avec le tracé de la ligne médiane (figure 3.a) et celui de la fonction d'aire (figure3.b).

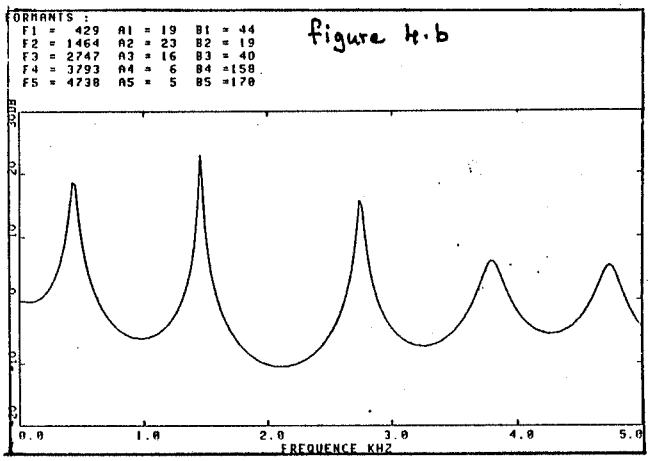




La première constatation que l'on peut faire est qu'il n'est pas possible à partir d'une telle fonction d'aire d'effectuer, sans autre modification, un calcul de la fonction de transfert ; en effet l'indétermination sur la section aux lèvres est grande, et elle implique l'intervention de l'opérateur, qui connaissant l'allophone étudié pourra utiliser son expertise pour lever les ambiguïtés. C'est ce qui est fait figure 4.a et la figure 4.b montre la fonction de transfert associée.

Comparons ces resultats à ceux qu' a obtenus, en 1984, à l'ICP, Hugo Sanchez à partir du même tracé radiographique, mais à l'aide d'une méthode semi automatique. Après la correction manuelle de l'aire aux lèvres, on obtient les résultats suivants (figure 5).





On observe lors de la comparaison que les deux méthodes donnent des allures de fonction d'aire analogues, mais on note que la longueur totale du conduit vocal que nous obtenons par calcul automatique est plus faible que dans l'autre cas, C'est là est un avantage non négligeable de notre technique, car les longueurs ainsi obtenues sont plus conformes aux mesures couramment faites sur le conduit vocal. On observe en effet que la méthode manuelle aboutit à une fonction de transfert à 6 formants entre 0 et 5kHz ce qui n'est pas courant et qui est bien lié à une longueur trop grande du conduit vocal. La fonction de transfert que nous obtenons automatiquement est, elle, parfaitement acceptable.

CONCLUSION

Nous avons proposé une méthode automatique d'extraction de la fonction d'aire à partir de tracés radiologiques. Cette méthode semble donner des résultats parfaitement corrects, en ce qui concerne le conduit vocal jusqu'à la zone alvéolaire. Ils sont même meilleurs, dans la détermination de la longueur, que ceux que donnait une méthode semiautomatique utilisée antérieurement à l'ICP.

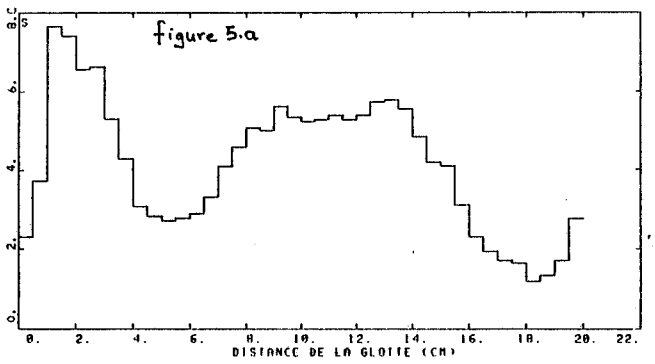
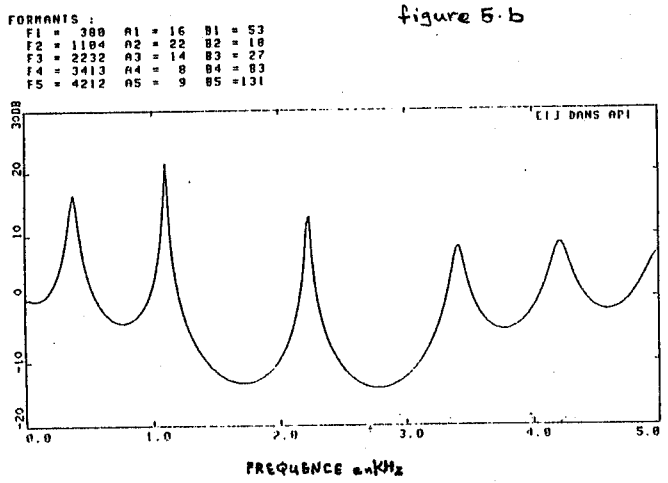


figure 5.b



Cependant, notre méthode ne lève la fameuse indétermination sur l'aire aux lévres, mais cela n'était il pas parfaitement prévisible ?

BIBLIOGRAPHIE

(1) T.Gay, "Articulatory Mouvements in VCV Sequences", J. Acoust. Am. 62,1, 183-193, 1977.

(2) O.Hassan, "Détermination automatique de la fonction d'aire du conduit vocal", Rapport de DEA, ENSERG, 1985.

(3) L.J. Boë & R. Descout, "Un logiciel d'acquisition graphique de données articulatoires", Bulletin de l'Institut de Phonétique de Grenoble, Vol. 14, 25-32, 1985.

(4) H. Sanchez, "De L'articulatoire à l'acoustique une simulation de La production de la parole", Thèse de Doctorat de L'I.N.P. Grenoble, 1985.

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

(6) H. Sanchez & L.J. Boë, "De la coupe sagittale à La fonction d'aire du conduit vocal", 13-ième J.E.P. du GALF, 23625, 1984.

(7) H. Sanchez, T. Gay & L.J. Boë, "De l'articulatoire à l'acoustique : étude de quelques transitions (VCV)", 14-ième J.E.P. du GALF, 30-33, 1985.

Nous remercions L. J. Boë pour ses précieuses suggestions.

FONCTIONS DE SENSIBILITE, MODELE ARTICULATOIRE ET VOYELLES DU FRANCAIS

R. Majid, L.J. Boë & P. Perrier

Institut de la Communication Parlée
Grenoble

ABSTRACT

Sensitivities of mode frequencies to small perturbations of an arbitrary area function A(x) presented by FANT & PAULI (1974) and calculated by MRAYATI (1976) for French vowels are not exploitable :

- they are only valid for less than 10% variations;
- A(x) is not an articulatory parameter and does not integrate articulatory constraints.

Using MAEDA's model we present sensitivity functions corresponding to variations of lip rounding, jaw, tongue body, dorsal shape and tongue tip and for [i, y, a, u] French vowels. Influence of articulatory commands on stability of vowels are discussed in relation to the quantal theory of STEVENS.

INTRODUCTION

La mise à l'épreuve puis l'exploitation d'un modèle va consister, pour une large part, à tester, interpréter, comparer les variations Δ_S des caractéristiques de sortie dues aux variations Δ_C des paramètres de commande.

Dans le cas de la modélisation du conduit vocal, les valeurs de la fonction d'aire discrétisée (Ai) ont été trop souvent les seuls paramètres d'entrée; quant aux formants (Fi) ils restent les caractéristiques "vedette" de la sortie.

A la suite des travaux de SCHROEDER [1] et HEINZ [2], FANT & PAULI proposent en 1974 [3] la solution analytique permettant de calculer les variations ΔF_i entraînées par les modifications d'un élément de la fonction d'aire ΔA_j, à partir de l'évaluation des énergies cinétique et potentielle le long du conduit vocal. MRAYATI [4] applique cette méthode à l'ensemble des voyelles du français. Enfin CHARPENTIER [5] en 1984, en adoptant une démarche lagrangienne, généralise la démonstration à l'ensemble de la fonction de transfert et pour l'ensemble des paramètres de commande. Il montre au passage que la méthode de FANT & PAULI n'est pas tout à fait adaptée dans le cas du conduit vocal avec pertes, conduit vocal plus proche de la réalité.

Le concept de sensibilité est véritablement opératoire. Si l'on s'en tient à la seule production (nous n'aborderons pas ici l'inversion du conduit vocal) il débouche sur les notions d'affiliation entre un paramètre d'entrée et une caractéristique de la sortie, et de stabilité d'une configuration donnée.

Le calcul direct des fonctions de sensibilité est apparu d'emblée comme séduisant puisqu'il permet l'économie de tout un jeu de simulations. Pourtant les travaux de FANT & PAULI et MRAYATI se sont vite révélés inexploitables :

1. Le calcul théorique des fonctions de sensibilité n'a de validité que pour de très faibles variations de Ai, (<10%) ce qui est bien trop peu par rapport aux réalités articulatoires. Au niveau des constrictions il n'est pas rare de vouloir simuler des variations de plus de 300%, et nettement plus pour l'ouverture aux lèvres (4 cm² pour un [i] , 0.16 cm² pour un [y]).
2. Et surtout la fonction d'aire n'est pas un paramètre d'entrée explicatoire. Les connaissances du processus de production nous conduisent à utiliser des commandes de type articulateur, beaucoup plus proche de la physiologie et des descriptions phonétiques : lèvres, mâchoires, langue, voile du palais ... Il est important de souligner que les modifications de type ΔA_j ignorent tout des contraintes articulatoires du conduit vocal.

METHODE

Pour cette étude sur les stabilités vocaliques nous avons utilisé le modèle de MAEDA [6] déjà testé globalement dans une étude précédente [7]. Les 5 paramètres de commande sont, rappelons-le, les lèvres, la mâchoire, la pointe, le corps et le dos de la langue. La fonction d'aire (calculée à partir de nos coefficients [8]) n'est plus qu'une donnée intermédiaire, interface entre la coupe sagittale délivrée par le modèle de MAEDA et un analogue électrique de type FLANAGAN & al. [9] implanté par CHARPENTIER [10] (il inclut les pertes par viscosité, chaleur, vibration des parois et l'effet de

rayonnement). Après un choix des valeurs "type" des paramètres de commande pour les 11 voyelles du français nous présentons les stabilités autour de ces cibles, paramètre par paramètre, voyelle par voyelle.

SELECTION DES PARAMETRES DE COMMANDE POUR LES VOYELLES DU FRANCAIS

Nous disposons, constitué lors de l'étude précédente [7],d'un dictionnaire de 371 293 configurations correspondant à un maillage (13 combinaisons possibles pour les 5 paramètres de commande) relativement fin de l'espace articulatoire. De cette masse de données nous avons extrait plus d'un millier de réalisations possibles par voyelle, pour des valeurs formantiques "standard" du français (et avec des dispersions de l'ordre de celles que nous avions mesurées auparavant [11]). Pour retenir des jeux de paramètres de commande type pour chaque voyelle nous nous sommes fixés une cohérence structurelle à 4 niveaux en intégrant des connaissances a priori tant articulatoires qu'acoustiques et nous avons procédé par approximations successives.

Niveau 1 : les paramètres articulatoires

Pour les lèvres nous avons utilisé les résultats de ABRY & al. [12], pour la mâchoire ceux de BOGNAR [13] et pour la langue de multiples données. Les choix retenus sont présentés à la figure 1.

Niveau 2 : les coupes sagittales

La aussi les données sont abondantes. Puisqu'il s'agit du français nous avons puisé dans les publications de l'Institut de Phonétique de Strasbourg (STRAKA, SIMON et BRICHLER-LABEYE [14]). Les coupes sagittales des 11 voyelles du français ainsi obtenues apparaissent à la figure 2.

Niveau 3 : les fonctions d'aires

Nous avons utilisé les données de BOË [15], MRAYATI [4] et FENG [16] en les comparant aux valeurs proposés par FANT [17] et des données discutées par WOOD [18] . Nous avons tout particulièrement porté notre attention sur le lieu d'articulation, l'aire correspondante et l'aperture aux lèvres (pour laquelle nous disposions de données très précises [12]). Ce sont évidemment les paramètres pertinents à ce niveau (FANT [17]). (Voir figure 3 pour les fonctions d'aire).

Niveau 4 : les formants

Les valeurs calculées ont été comparées aux résultats [4, 15, 16]. Le tableau I présente les 5 premiers formants et les bandes passantes correspondantes.

Contrôle

Enfin les voyelles ont été synthétisées à partir des pôles et des bandes passantes de la fonction de transfert, selon une méthode développée et implantée par FENG [19], pour un contrôle auditif.

LES SENSIBILITES

Pour chaque voyelle nous avons fait varier une par une les valeurs articulatoires de commande par pas de 0.02 pour obtenir 98 valeurs par paramètre (dans le modèle de MAEDA les commandes ont une plage qui évolue en + 3 et - 3 fois l'écart type autour de la moyenne : le conduit "neutre"). Quand cette exploration symétrique n'a pas été possible (il existe des contraintes dans les cas limites) nous avons conservé le même nombre de simulations (99 avec la cible) mais en décalant l'exploration dans une plage possible (cas du [i] par exemple). Dans ce cas la caleur cible n'est plus au centre de la gamme explorée.

Nous présentons ici les résultats relatifs aux voyelles [i, y, a, u] quadruplet dont nous montrons l'importance dans ces mêmes actes [20] .

On se reportera à la figure 3 pour les sensibilités dans le plan F1, F2 et à la figure 5 pour les rapports de variation entre valeurs maxima et minima.

1. Les lèvres

La sensibilité de [u] à ce paramètre est spectaculaire, puisque rétraction et aperture labiales entraînent une véritable centralisation via [o]. C'est pour [i] que les lèvres ont le moins d'influence (en dynamique : très faible sur F1, moyenne sur F2) bien qu'elles permettent de passer directement à [y]. A noter un effet non symétrique, puisque de [y] on ne retourne pas à [i] en ouvrant et étirant les lèvres; il faut faire varier un autre paramètre, le dos de la langue par exemple. Avec [a] l'effet porte essentiellement sur F1 sensible à la fermeture.

2. La mâchoire

Dans tous les cas son abaissement provoque une augmentation de F1. C'est avec [i] qu'apparaissent les variations maximales, plus faibles avec [y] mais la trajectoire F1, F2 reste parallèle. L'effet sur [a] est à double détente : d'abord sur F1 puis sur F2 et ceci à cause de l'interaction avec le corps de la langue; [u] reste stable. Ces résultats confirment ceux de LINDBLOM & SUNDBERG [21] et les précisent.

3. Le corps de la langue

Il agit sur [i] et [y] et dans une moindre mesure sur [a] : F1 croît et F2 diminue si le corps se déplace à l'arrière. [u] affiche sa stabilité.

4. Le dos de la langue

C'est [i] le plus instable (F1 encore plus que F2 et en sens inverse) suivi de [y] (F2 plus que F1). Pour [a] et [u] le dos n'a de l'influence que sur F2 et encore en sens contraire. Bien noter que lorsqu'il s'abaisse, [i] tend vers [e] alors que [u] évolue dans la direction du [y] : les déplacements dans F1, F2 sont presque orthogonaux beaucoup plus nettement que les simples tendances

mises en évidence par GOLDSTEIN [22].

5. L'apex

Il n'influe que faiblement sur [u] et [a] alors que F1 de [i, y]est très sensible.

Sur les trajets F1, F2 on peut remarquer des influences tout à fait similaires : de la mâchoire, du corps et du dos de la langue pour [i], des lèvres et de l'apex pour [a],du dos et du corps de la langue pour [u]; et du corps et de l'apex pour [y].

En se référant à la dynamique des variations (figure 5) on notera que [i, y] sont les voyelles les plus instables (grosses variations de F1) et [a, u] les plus stables. Globalement les effets observés sont nets sur F1 et plus répartis sur F2.

CONCLUSION

Intégrant des contraintes articulatoires ces résultats de sensibilité révèlent toute leur richesse et par là même la validité du modèle. Ils confirment de grandes tendances mises déjà en évidence dans des travaux précédents [21, 22, 23]. Certains points nous semblent affinés : par rapport aux lèvres, relations non symétriques entre [i] et [y], comportements très distincts de [i] et [u] pour des variations du corps et du dos de la langue [i → e] et [u → y].

Dans sa relation articulatori-acoustique [i] se révèle très instable, comparé à [a] et surtout à [u]. Ces résultats corroborent les inductions développées à partir de nomogrammes [20]. Ils incitent à regarder de plus près l'hypothèse quantique de STEVENS [24]. La stabilité la plus remarquable, celle du [u] ne permettrait qu'une imprécision sur la cible linguale (en tout cas pas sur les lèvres). Le [a] n'est véritablement stable que sur le premier formant (il est vrai qu'il n'entre que rarement en concurrence dans les systèmes vocaliques avec un adversaire postérieur, sur F2). Quant à [i] son instabilité ne lui permet guère d'imprécision articulatoire. Heureusement cette voyelle universelle doit pouvoir bénéficier des riches contrôles perceptuo-moteurs dans cette partie du conduit vocal.

Remerciements

Christian ABRY nous a permis de choisir les voyelles, il a assuré la coordination avec le travail sur les nomogrammes [20].

REFERENCES

[1] M.R. SCHROEDER, Determination of the Geometry of the Human Vocal Tract by Acoustic Measurement. J. Acoust. Soc. Am. 41, 1002-1010, 1967.

[2] J.M. HEINZ, Perturbation Functions for the Determination of Vocal-Tract Area Function from Vocal Tract Eigenvalues. STL QPSR 1, 1-14, 1967.

[3] G. FANT & S. PAULI, Spatial Characteristics of Vocal Tract Resonance Modes. Speech Comm. Seminar 2, 121-132, 1974.

[4] M. MRAYATI, Contribution aux études sur la production de la parole. Doct. es-Sci. Physique, USM Grenoble, 1976.

[5] F. CHARPENTIER, Fonctions de sensibilité d'un modèle dissipatif du conduit vocal. 13èmes JEP, GALF-GCP, 27-28, 1984.

[6] S. MAEDA, Un modèle articulatoire de la langue avec des composantes linéaires. 10èmes JEP, GCP-GALF, 152-162, 1979.

[7] P. PERRIER, L.J. BOË, R. MAJID SHIBAB & B. GUERIN, Modélisation articulatoire du conduit vocal : exploration et exploitation. 14èmes JEP, GALF-GCP, 55-58, 1985.

[8] H. SANCHEZ, L.J. BOË, De la coupe sagittale à la fonction d'aire du conduit vocal. 13èmes JEP, GALF-GCP, 23-25, 1984.

[9] J.L. FLANAGAN, K. ISHIZAKA & K.L. SHIPLEY, Synthesis of Speech from a Dynamic Model of the Vocal Cords and Vocal Tract. B.S.T.J. 54, 485-506, 1975.

[10] F. CHARPENTIER, Un logiciel de simulation électrique du conduit vocal. CNET. Communication Personnelle, 1982.

[11] C. ABRY & L.J. BOË, [i, a, u]? Pas si fou ? Ou les lèvres des consonnes maximisent-elles l'espace des voyelles ? 13èmes JEP GCP GALF, 205-207, 1984.

[12] C. ABRY, L.J. BOË, P. CORSI, M. GENTIL, P. GRAILLOT, Labialité et phonétique. Publications de l'Université des Lnagues et Lettres de Grenoble, 304 p. 1980.

[13] E. BOGNAR, 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 3e Cycle de Phonétique, Université de Grenoble 3, 1982.

[14] C. BRICHLER-LABEYE, Les voyelles françaises. Mouvements et positions articulatoires à la lumière de la radiocinématographie. Klincksieck, Paris, 1970.

[15] L.J. BOË, Etude acoustique du couplage larynx-conduit vocal (fréquence laryngienne des productions vocaliques). Revue d'Acoustique 27, 235-244, 1973.

[16] G. FENG, Apport de la modélisation au traitement du signal de parole, le cas des voyelles nasales et la simulation des pôles et des zéros. Doctorat ès-Sciences, à paraître.

[17] G. FANT, Acoustic Theory of Speech Production. Mouton, The Hague, 1960.

[18] S. WOODS, A Radiographic Examination of Constriction Location for Vowels. J. of Phonetics 7, 25-43, 1979.

[19] G. FENG, Vers une synthèse par la méthode des pôles et des zéros. 13èmes JEP GCP-GALF, 155-157, 1984.

[20] L.J. BOË & C. ABRY, Nomogrammes et systèmes vocaliques. 15èmes JEP GCP-GALF, 1986.

[21] B.E.F. LINDBLOM & J.E.F. SUNDBERG, Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement. J. Acoust. Soc. Am. 50, 1166-1179, 1971.

[22] L. GOLDSTEIN, Vowel Shifts and Articulatory-Acoustic Relations. 10th Int. Congr. Phonetic Sci. IIA, 267-273, 1983.

[23] J.S. PERKELL & W.L. NELSON, Variability in Production of the Vowels /i/ and /a/. J. Acoust. Soc. Am. 77, 1889-1895, 1985.

[24] K.N. STEVENS, The Quantal Nature of Speech : Evidence from Articulatory-Acoustic Data. In Human Communication. Mc Graw Hill, New-York, 1972.

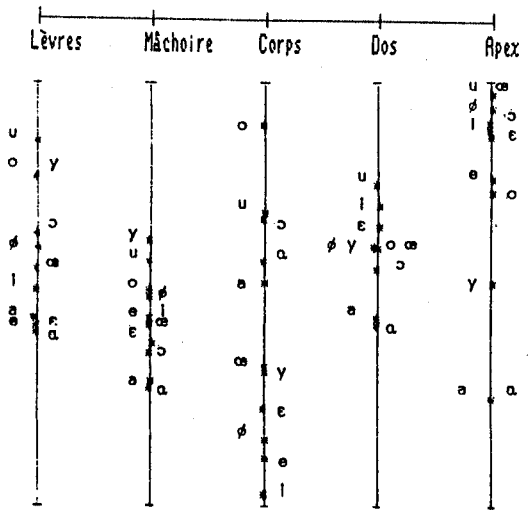


Figure 1 - Répartition des paramètres dans le plan articulatoire pour les 11 voyelles du français.

TABLEAU 1

	F1	F2	F3	F4	F5
[i]	293	2247	2586	3875	4704
	61	43	127	52	145
[e]	371	2145	2554	3845	4670
	54	64	119	47	120
[ɛ]	538	1802	2435	3727	4376
	45	71	82	82	156
[a]	650	1356	2413	3693	4472
	45	38	64	55	88
[y]	285	1777	2313	3806	4268
	63	41	38	47	74
[ø]	401	1691	2358	3689	4262
	44	51	41	46	64
[oe]	531	1485	2353	3646	4245
	41	51	42	51	66
[u]	295	734	2401	3734	4199
	67	37	35	35	37
[o]	419	867	2363	2641	4053
	48	35	46	42	53
[ɔ]	519	1090	2310	3556	4063
	51	36	36	40	51
[ɑ]	658	1244	2371	3707	4404
	51	38	56	57	87

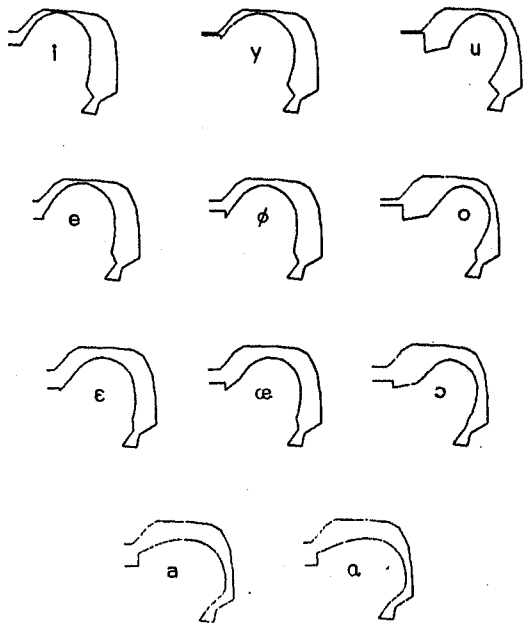


Figure 2 - Coupe sagittale des voyelles modélisées.

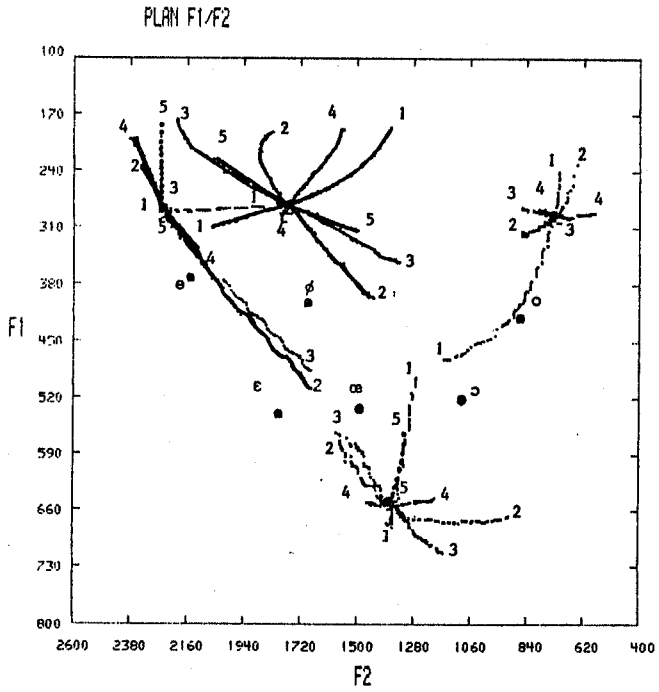


Figure 4 - Fonctions de sensibilité articulatoire pour [i, u, y, a] avec les paramètres suivants :
1 : lèvres 2 : mâchoire 3 : corps de la langue
4 : dos de la langue 5 : pointe de la langue.

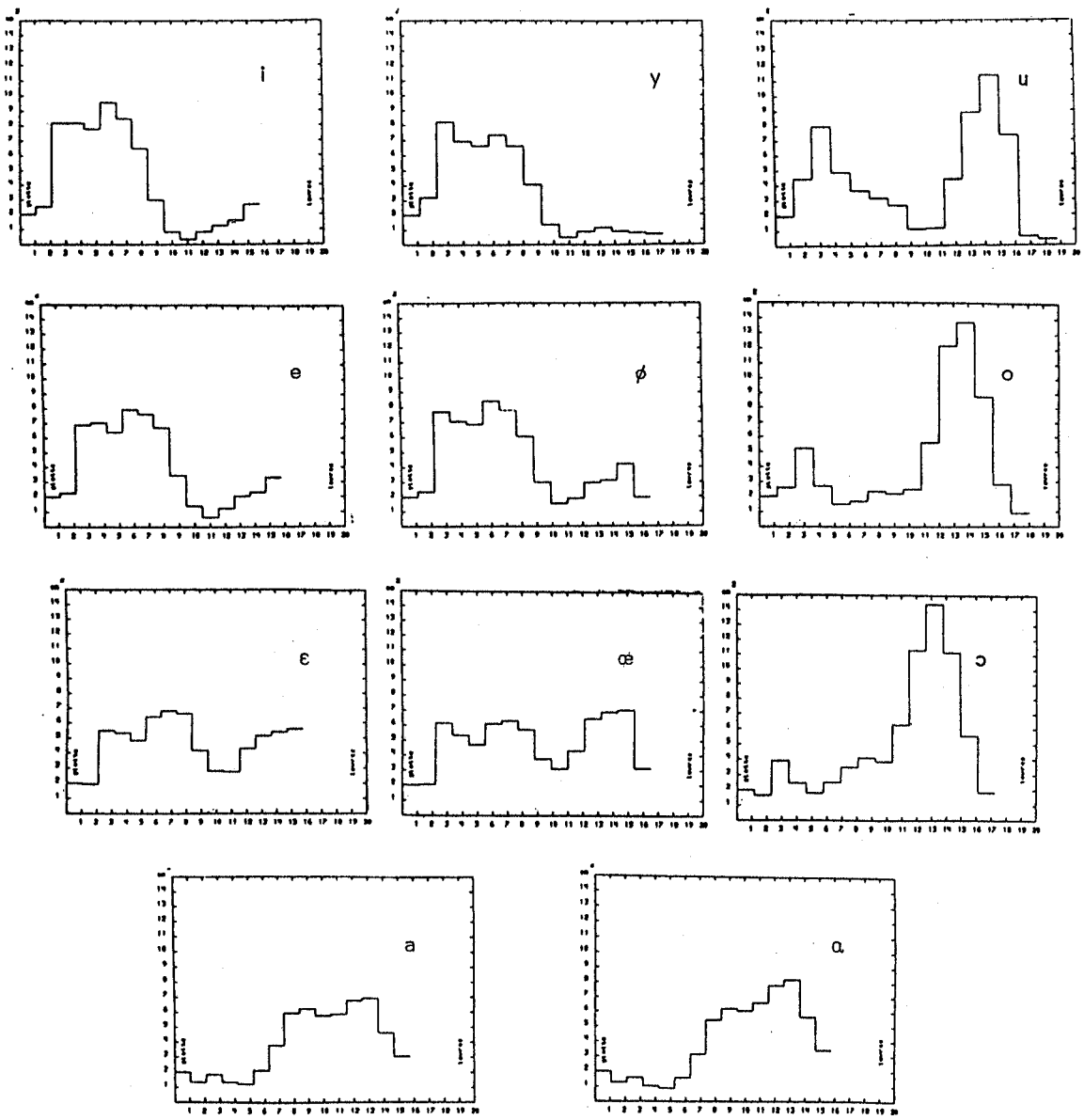


Figure 3 - Fonctions d'aire des 11 voyelles orales.

F1	Lèvres	Mâchoire	Corps	Dos	Pointe
i	○	◯	◯	◯	◯
y	◯	◯	○	◯	○
u	◯	○	○	○	○
a	○	○	○	○	○

F2	Lèvres	Mâchoire	Corps	Dos	Pointe
i	○	○	○	○	○
y	○	○	◯	○	○
u	◯	○	○	◯	○
a	○	◯	○	○	○

Figure 5 - Rapport de variations (min, max) de F1, F2 pour chaque paramètre et pour chaque voyelle.

DESCRIPTION D'UN DISPOSITIF D'ENREGISTREMENT SIMULTANE DES MOUVEMENTS DES ORGANES ARTICULATEURS

Bernard Teston & Denis Autesserre

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

This experimental device allows the simultaneous recording :
- Endoscopic images (velum, pharynx, larynx, oral cavity)
- Labiofilm : frontal, side view
- Complete X-ray film of the vocal tract.

Speech signals and images are synchronized. The various images are recorded simultaneously on video magnetoscopes at 50 frames per second.
As such, this device is particularly valuable for obtaining a perfect correlation between articulatory and acoustical events.

I. INTRODUCTION

L'étude du mouvement des organes articulateurs a toujours présenté un très grand intérêt pour la connaissance des mécanismes de production de la parole. Cependant, ces investigations n'ont jamais fait l'objet d'une utilisation de routine. Elles ont été menées par un petit nombre de chercheurs au sein d'équipes ou d'instituts spécialisés, tant leur mise en œuvre et leur exploitation sont difficiles. L'Institut de Phonétique de Strasbourg en est en France la plus évidente illustration pour la radiocinématographie avec entre autres Georges Straka, puis Pela Simon, André Bothorel, F. Wioland et J.P. Zerling [1, 2], et l'équipe CNET-Grenoble [3] pour les labio-films avec Raymond Descout, Michèle Gentil et Louis-Jean Boé [4, 5].
Les difficultés qui s'opposent à la généralisation des études sur la dynamique de la phonation au moyen d'investigations radiocinématographiques sont de trois ordres : d'abord, celles qui sont propres aux principes mêmes du cinématographe [6]; ensuite, les difficultés du traitement des données qui s'apparente encore à un travail de bénédictin, d'où des tentatives de traitements automatiques des images de films (Gouret et al. 1985) [7].
Enfin, des difficultés générales de mise en œuvre de ce type de manipulations, telles que l'utilisation de matériels et d'installations qui ne sont pas spécifiques à nos investigations, dans des services hospitaliers qui n'ont pas les mêmes vocations que les nôtres, tout cela nécessitant des adaptations physiques et intellectuelles parfois difficiles, toujours délicates, même lorsque existent d'excellentes relations humaines entre les services.
Ceci ne nous permet pas une totale autonomie sur la chaîne instrumentale telle que celle dont peuvent disposer Fujimura et al. [8, 9]. Pour notre part, nous avons réalisé, ou participé à de nombreuses manipulations de ce type sans nous attacher particulièrement à faire évoluer les méthodes d'investigation. A partir de 1983, nous avons commencé à utiliser l'endoscopie par voie nasale pour étudier les mouvements du voile du palais en employant un circuit de télévision, que nous avons rapidement étendu à l'étude des mouvements des lèvres. Nous avons alors défini un système d'investigation des mouvements des organes de la phonation au moyen d'images de télévision qui améliore essentiellement les difficultés propres au principe du cinématographe et doit permettre, à notre avis, de donner un élan nouveau à ce type d'investigations.

II. COMPARAISON ENTRE L'IMAGE CINEMATOGRAPHIQUE ET L'IMAGE DE TELEVISION

Au risque de paraître par trop scolaire et de développer une partie qui n'a pas sa place dans un exposé qui se doit succinct, nous n'exposons pas les principes physiques du cinématographe (cinéma) et de la télévision (T.V). Il nous est cependant nécessaire d'énumérer rapidement les avantages ou les inconvénients de chacun de ces systèmes.
- Le cinéma est une suite d'images photographiques; l'entraînement du film est mécanique, image par image, donc non continu. La T.V est une analyse continue de l'image par un balayage électronique point par point.
Cette différence fondamentale explique la bien meilleure définition de l'image T.V, à l'arrêt, par comparaison avec l'image cinéma (pas de flou, haute luminosité). Ce point est très important, tous les expérimentateurs connaissant ces techniques peuvent en témoigner.
- Les caméras de T.V sont généralement beaucoup plus sensibles que les pellicules photographiques à résolution et nombre d'images comparables.
A luminosité égale de l'image, la T.V exige beaucoup moins d'éclairement que le cinéma. Ce point est très important pour l'endoscopie et surtout la radiocinématographie que l'on peut réaliser en T.V avec des doses de rayons X beaucoup plus basses (140 fois moins dans notre cas) qu'avec le cinéma.
- Le cinéma autorise une prise de vue avec un nombre d'images par seconde variable. La T.V par contre

n'autorise que 25 images par seconde avec un enregistrement au standard V.H.S. et 50 images par seconde au standard U.MATIC (dans ce dernier cas la résolution de l'image est réduite de moitié dans le sens vertical, car l'on utilise qu'une trame sur deux).
- La résolution des images est du même ordre entre le cinéma et la T.V pour 50 images par seconde, au format 16 mm.
- Le contrôle de la prise de vue effective est direct pour la T.V, différé au développement pour le cinéma.
- L'utilisation de la T.V est beaucoup plus souple.
- Le prix de revient du support est très favorable à la T.V.
- La prise de son est parfaitement synchrone pour la T.V.
- Il est possible de synchroniser parfaitement le balayage de nombreuses caméras de T.V, et ainsi d'obtenir des images simultanées de différents mouvements. Ceci est impossible à réaliser en cinéma.
- Enfin, le signal électrique qui contient l'information T.V (le signal vidéo) peut être enregistré et restitué au moyen de magnétoscopes et l'on peut lui faire subir différents traitements qui permettent le mixage, le montage, l'inversion et l'insertion d'images issues de différentes caméras. Il peut également être numérisé directement en vue d'un traitement d'image ultérieur sur ordinateur.

III. LE SYSTEME D'ENREGISTREMENT

Le but que nous poursuivons consiste à obtenir le maximum d'informations simultanées sur les mouvements des organes articulateurs en synchronisation avec le signal de parole.
Dans un premier temps, nous avons associé à un labio-film de face une image endoscopique des mouvements du voile du palais. Le signal de parole apparaît non synchrone avec les images, uniquement dans le but d'avoir des repères de réalisation (consonne, voyelles). Des tops de synchronisation image issus de la synchrotrame du signal vidéo sont enregistrés sur la piste audio n° 2 du magnétoscope, ainsi que des tops représentant les CLAP en début de phrase qui sont déclenchés manuellement par le sujet avant chaque réalisation. Le CLAP apparaît sur l'image sous la forme d'un point lumineux rouge. Les images issues de la caméra de l'endoscope, de la caméra des lèvres, et de la caméra du signal acoustique et du CLAP, sont mixées au moyen d'une console J.V.C KM 2.000 qui réalise également la synchronisation commune de toutes les sources.
L'endoscope est un bronchoscope OLYMPUS; sa caméra est une HITACHI DK 5051. Les caméras aux lèvres et de l'ensemble CLAP-oscillogramme sont des SONY DXC 1820 P. Le dispositif de synchronisation que nous utilisons est celui que nous avons développé pour la radiocinématographie depuis plusieurs années. Il est d'un principe similaire à celui décrit par Simon et al. (1979) [3]. Sur les nombreux enregistrements que nous avons réalisés avec ce système, nous n'avons jamais perdu une image.
Cette manipulation nous a permis de réaliser d'importants travaux sur la nasalité et de faire évoluer le système vers la version actuelle à la lumière des différents problèmes que nous avons rencontrés pendant ces trois dernières années.

Le système actuel est schématisé sur la figure suivante :
Nous disposons de quatre caméras :
1. La caméra de l'installation d'angiographie (C.G.R - 24 cm de champ - 0,5 mA de courant de cathode en radioscopie).
2. La caméra de l'endoscope (HITACHI DK 5051).
3. La caméra du labio-film (SONY DXC 1820 P).
4. La caméra de l'oscillogramme des numéros d'image et de CLAP, et des informations d'enregistrement.

Ces caméras sont toutes synchronisées par une horloge extérieure qui pilote également les magnétoscopes, les images endoscopiques de radiocinéma, et des lèvres, sont toutes mixées avec l'oscillogramme et les informations de synchronisation au moyen des mélangeurs (6), (7) et (8) (SONY CRK 2.000 P).
Elles sont ensuite enregistrées sur les magnétoscopes (9)-(10)-(11) (SONY VO 5800 PS), les informations sur les n° d'image, n° de CLAP et sur les enregistrements (référence sujet, date, etc.) sont envoyées sur un scope qui visualise l'oscillogramme du signal acoustique (18). Pour que ce dernier soit parfaitement synchronisé avec les images, nous l'avons placé verticalement pour tenir compte du balayage vidéo.
Un système particulier (12) permet de compter les CLAP (20), ainsi que les images entre deux CLAP (99 CLAP et 999 images). Il envoie sur toutes les pistes audio n° 2 des magnétoscopes les tops d'image et de CLAP ainsi que sur la piste n° 2 d'un magnétophone (15).
Les images sont comptées à partir de l'horloge de référence (5).
Les mixage général de toutes les images est opéré au moyen d'une console de mixage (13) (JVC KM 2000) pour avoir une image synthétique de toutes les prises de vue. Elle est enregistrée sur le magnétoscope (14) (SONY VO 5 8.000 PS).
L'enregistrement sonore est effectué au moyen d'un microphone à électret (16). Après amplification (17), le signal acoustique est enregistré sur les pistes audio n° 1 de tous les magnétoscopes et sur la piste n° 1 d'un magnétophone (15) (REVOX B 77) les quatre magnétoscopes (9, 10, 11, 14) et le magnétophone (15) sont télécommandés (19).
Quel que soit le système d'exploitation des enregistrements ainsi réalisés, il est possible de repérer exactement les images et le son. Le nombre de caméras n'est pas limité, bien qu'il soit difficilement concevable de filmer autre chose sur le sujet. Cependant nous envisageons d'utiliser une ou plusieurs caméras pour visualiser des mouvements traduits par d'autres capteurs tels qu'électropalatographiques par exemple.

IV. L'INTERFACE SUJET-INSTRUMENT

Nous entendons par interface tout l'appareillage qui, bien que pouvant paraître annexe, permet de réaliser une prise d'information (images et son) pratique, précise et reproductible.
L'ensemble de cet appareillage est constitué par un confortable fauteuil sur lequel prend place le sujet, dont les mouvements de la tête sont contenus au moyen d'un céphalostat efficace et peu contraignant. Le fauteuil est situé au centre d'un parallélipipède rectangle dont le positionnement peut être très précisément ajusté, qui supporte les caméras des lèvres

et endoscopiques, le système d'éclairage, les miroirs, les filtres X, les dispositifs de calibration des images, les microphones, etc. Le volume de cet ensemble facilement transportable est de l'ordre de 2,5 m3. Il est adapté aux installations d'angiographie actuelles.

V. EXPLOITATION DES RESULTATS ET TRAITEMENT DES DONNEES

L'énorme quantité d'informations ainsi récoltées ne peut être traitée que par l'intermédiaire d'un moyen puissant d'aide au dépouillement de ces images en attendant un moyen automatique efficace.
En ce qui concerne les images, leur exploitation est réalisée au moyen d'un convertisseur A/D des signaux vidéo sur microcalculateur MICRAL 30. Les mesures de distance, d'aires et de contour sont réalisées au moyen d'un programme développé par P. Martin. Elles sont effectuées à partir de trames de calibration ou de repères placés sur le sujet. Ce dépouillement est directementnspiré du précédent effectué manuellement sur des calques. Les données sont stockées sur disquettes avec les références des images pour effectuer ultérieurement des traitements statistiques ou des reconstitutions de mouvements. Il est envisagé de réaliser un système de traitement automatique plus élaboré.
L'analyse acoustique du signal de parole associé aux réalisations articulatoires peut être effectuée par les moyens habituels de traitement du signal dont nous disposons. Pour cela, nous utilisons un système de déclenchement automatique de l'acquisition sur la n-ième image du n-ième CLAP.

CONCLUSION

Dans les recherches physiologiques, les résultats qui circulent sont déjà interprétés et, souvent, les chercheurs aimeraient disposer des données brutes qu'il était jusqu'à présent très difficile de diffuser.
Le support de l'image vidéo sur bande magnétique aux standards internationaux (U.MATIC ou V.H.S.) ainsi que la numérisation des images et leur exploitation (mesure) sur calculateur vont pouvoir faciliter la circulation et la diffusion de ces informations [10].

L'exploration endoscopique est effectuée dans le service d'exploration fonctionnelle par le Dr J.F. Dumon, la radiocinématographie est réalisée dans le service d'angiographie du département de Radiologie du Professeur Kasbarian, ces deux services étant situés à l'hôpital Salvator à Marseille.

L'achat du matériel d'exploitation des vidéo-films a été pris en charge dans le cadre d'une convention de recherche avec la Direction de la Musique au Ministère des Affaires Culturelles.
Les frais d'expérimentation sont financés par une subvention de l'A.D.I. dans le cadre de l'action "Décodage acoustico-phonétique" du GRECO "Parole".

BIBLIOGRAPHIE

[1] Brock, G., "Méthode de synchronisation graphique images-son pour l'exploitation des films radiologiques. Présentation de l'appareillage réalisé à l'Institut de Phonétique de Strasbourg", **Travaux de l'Institut de Phonétique de Strasbourg**, n° 9, 221-232, 1977.

[2] Simon, P., Bothorel, A., Wioland, F. & Brock, G., "Méthode de synchronisation image-son pour l'étude radiologique des faits de parole. Application au français", **Actes du 9e Congrès International des Sciences Phonétiques**, Copenhague, vol. 1, 213, 1979.

[3] Simon, P., Wolff, F, Nadjafizadeth, H., Brock, G., "Numérotation automatique et codage synchrone graphique et phonétique pour l'exploitation des films radiologiques", **Travaux de l'Institut de Phonétique de Strasbourg**, n° 16, 181-190, 1984.

[4] Descout, R., Boé, L.J., Abry, C., "Labialité vocalique et consonantique en français. Premiers résultats", **9es Journées d'Etudes sur la Parole**, GALF, 179-189, 1978.

[5] Abry, C., Boé, L.J., Corsi, P., Descout, R., Gentil, M., Graillot, P., "Labialité et phonétique", **Publications de l'Université des Langues et Lettres de Grenoble**, 1980.

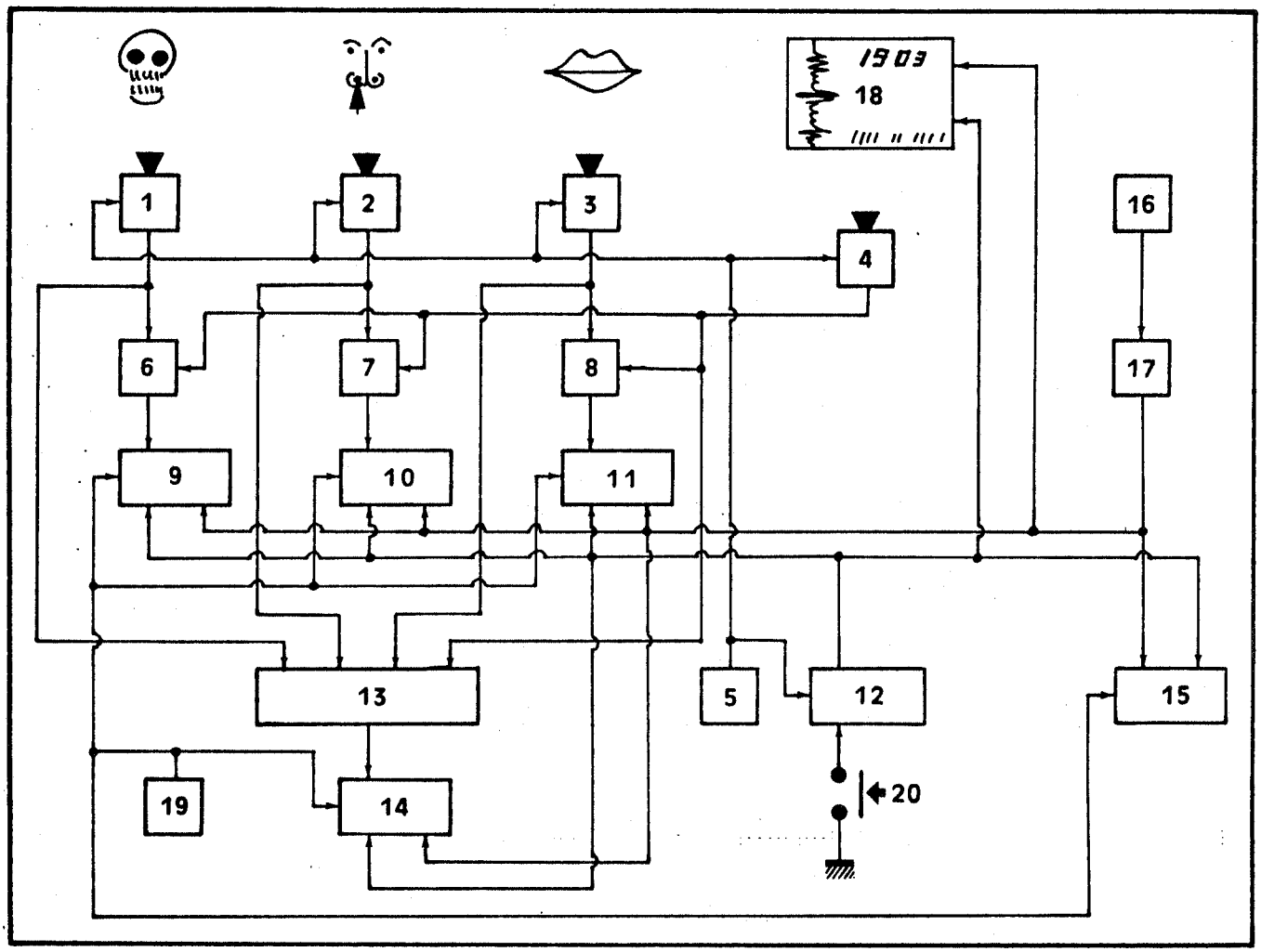
[6] Autesserre, D., Rossi, M., Sarrat, P., Giraud, G., Visquis, R., Demange, R. & Chevrot, L., "Exploration radiologique de l'oropharynx, de l'hypopharynx et du larynx en phonation", dans **Larynx et Parole**, GALF, Institut de Phonétique de Grenoble, 45-73.

[7] Gourret, J.P., Paille, J., Boé, L.J., Descout, R., "Système d'exploitation automatique de labio-films", **14es Journées d'Etudes sur la Parole**, GALF, Paris, 61-64, 1985.

[8] Fujimura, O., Kiritani, S., Ishida, H., "Computer controlled radiography for observations of movements of articulatory and other human organs", **Comput. in Biol. and Med.**, 3, 371-384, 1973.

[9] Kiritani, S., Itoh, K., Fujimura, O., "Tongue pellet tracking by a computer controlled X-ray microbeam system", **J.A.S.A.**, 57, 1516-1520, 1975.

[10] Segura, J., "L'imagerie médicale en réseau", **Sciences et Techniques**, n° 25, 36-43, 1986.



1 - Caméra de l'installation d'angiographie.
2 - Caméra de l'endoscope.
3 - Caméra des lèvres.
4 - Caméra de l'oscillogramme
- le numéro de clap
- le numéro d'image
- des informations de l'enregistrement

5 - Horloge de synchronisation générale
6, 7 et 8 - Unité d'incrustation et de mélange SONY CRK 2.000 P
9, 10, 11, 14 - Magnétoscopes SONY VO5800 PS
12 - Système de comptage des claps et des images
13 - Console d'incrustation et de mélange JVC KM 2.000
15 - Magnétophone REVOX B 77
16 - Microphone d'enregistrement
17 - Unité de traitement du signal acoustique
18 - Ecran de visualisation du signal acoustique, des n° de clap et d'image et des informations d'enregistrement
19 - Télécommande générale des magnétoscopes et du magnétophone
20 - Déclenchement manuel du clap.

PROSODIE

Président

Mario ROSSI

Institut de Phonétique d'Aix

DE LA MELODIE A L' INTONATION (8-24 mois)

Gabrielle Konopczynski

Laboratoire de Phonétique 25030 Besançon-cedex

ABSTRACT

This paper deals with the acquisition of prosodic structures by the child aged 8 to 24 months. The study of this age is vital for the comprehension of the procedures involved in child mother-tongue acquisition. During this period various categories of speech co-exist, going from complete gibber, produced in solitary ludic situations, to language combining a few lexical items, and passing through protolanguage (PL) which is produced in interaction situations. Each category has a specific function.

The analysis shows the following results: gibbering has no defined structure whereas PL shows from the outset the beginnings of language organization. The rich but random melodic variations of gibber are replaced by a few basic contours, which have stable parameters. Melody becomes intonation, with an oppositional function.

INTRODUCTION

Nos recherches portent sur l'acquisition de la structuration prosodique durant la période allant du prélangage (8 mois) au langage articulé référentiel (24 mois) dont l'étude est indispensable pour découvrir les processus de mise en place du tout premier langage. Durant la dernière décennie, un nombre croissant de chercheurs a émis l'hypothèse d'une acquisition précoce des patrons intonatifs, mais très peu ont tenté de mettre cette hypothèse à l'épreuve des faits; ainsi les travaux sont restés, soit très ponctuels, soit subjectifs.

C'est donc à ce domaine quasi inconnu qu' est la prosodie du babillage que nous avons consacré nos travaux. Dans cet exposé, nous étudierons le seul aspect mélodique des émissions enfantines. Afin d'être en mesure de montrer comment la mélodie devient progressivement intonation, avec fonction linguistique, nous rappellerons brièvement les caractéristiques prosodiques de la période capitale que nous appelons période charnière (8-11 mois) puis nous étudierons l'évolution des énoncés phatiques, particulièrement représentatifs de la période 8 à 24 mois.

I. CARACTERISTIQUES DE LA PERIODE-CHARNIERE.

Plusieurs publications antérieures[1,5] nous ont permis de montrer qu'il y a dans les émissions sonores enfantines coexistence de diverses catégories de discours allant du pur jasis au langage référentiel en passant par le Protolangage (PL par la suite) et divers types plus ambigus. Chacune de ces catégories est liée à une situation d'émission spécifique et possède une structuration syllabique, temporelle et mélodique qui lui sont propres. Par ailleurs, dès 8-9 mois l'enfant sait dominer sa voix et la plier aux contraintes sociales lorsque cela s'avère nécessaire[2,3]. Ainsi, le **jasis**, émis en situation de jeu solitaire, apparaît, du point de vue syllabique, dominé par la présence de vocoïdes alors que le PL, qui est produit dans les situations d'interaction avec l'adulte ou un substitut d'adulte(poupée...) est plutôt constitué de structures syllabiques de type CV (fig.1a et b).

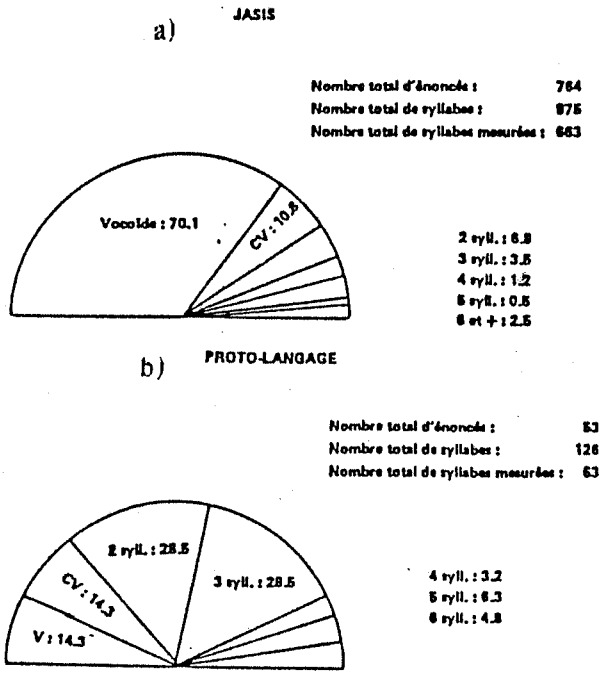


Fig.1:STRUCTURES SYLLABIQUES DU JASIS ET DU PL

Nous n'avons pu dégager aucune structuration temporelle dans les vocoïdes longs et instables du Jasis. En revanche, les structures CV du PL montrent une organisation de type isochronique[1]. Du point de vue mélodique [4,5], le Jasis présente un riche éventail de courbes de Fo (Fig.2). Si certaines sont totalement absentes du langage adulte (courbes ondulées, pas tonaux reitérés), d'autres sont aussi présentes dans le PL et le modèle adulte; mais même lorsque la forme globale de la courbe est identique à celle de l'adulte, la courbe enfantine est très floue, à pente variable, souvent ondoyante. Les schémas complexes présentent une complexité maximale, c'est-à-dire que les courbes changent sans arrêt et brutalement de direction, de pente, de forme. L'enfant semble essayer toutes les combinaisons possibles, au hasard de sa fantaisie, et produit ainsi une succession de mouvements tout à fait aléatoires.

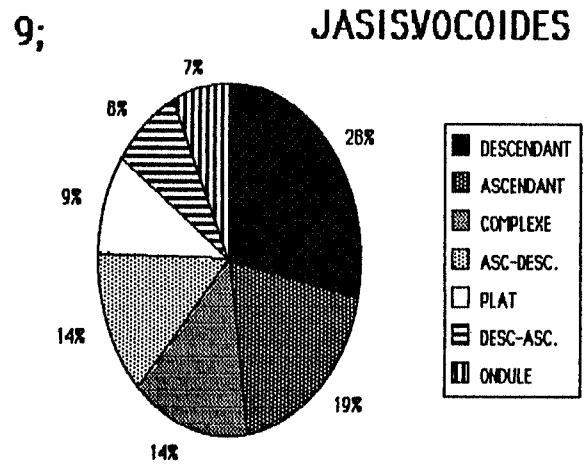


Fig.2 : COURBES MELODIQUES DU JASIS A 9 MOIS

Enfin, l'analyse de la tessiture enfantine montre une utilisation des fréquences extrêmes: la voix peut aller depuis le creak très grave jusqu'aux fréquences sur-suraiguës de l'ordre de 2.000Hz. Le Fo moyen affiche une grande instabilité et une large dispersion, comme le montrent les chiffres du tableau 1 ci-dessous:

	Fo-m	s.d.	Tessiture
sujet1	401Hz	90Hz	280-500Hz:69%,reste >500Hz
sujet2	416Hz	74Hz	240-500Hz:80%,reste >500Hz
sujet3	429Hz	150Hz	250-500Hz:67%,reste >500Hz

L'ensemble des productions sonores témoigne de la part de l'enfant d'une activité de type exploratoire. Tout se passe comme s'il testait ses capacités respiratoires et vocales. Par opposition, dans les énoncés multisyllabiques du PL les riches mais aléatoires variations du Jasis sont remplacées par quelques contours de base, en nombre réduit(Fig.3), peu complexes, stables et récurrents, affectés de traits constants et précis. Ils servent à exprimer quelques grandes modalités langagières, telles que l'interrogation, l'appel, l'énonciation. Un auditeur non averti du contexte situationnel attribue en général une fonction donnée à ces énoncés, malgré l'absence totale de couche verbale et d' informations para- linguistiques.

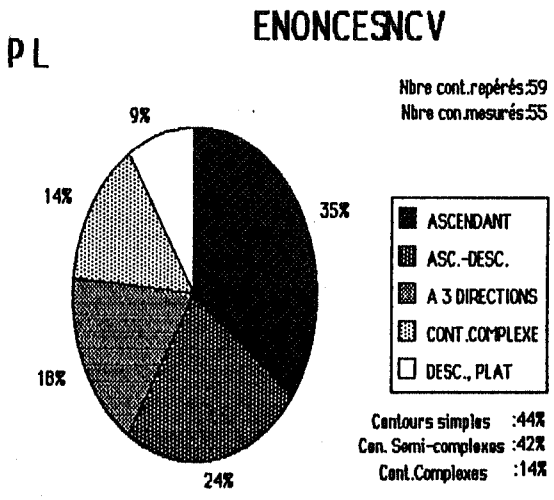


Fig.3 : CONTOURS MELODIQUES DU PL A 9 ET 10 MOIS

En outre, un énoncé apparaît rarement isolé, mais à l'intérieur de séries dont les divers paramètres vont en croissant, alors que ceux du Jasis sont soit isolés, soit en succession aléatoire. Les contours mélodiques du PL s'opposent entre eux à la fois par leur direction, leur pente, le niveau de tessiture employé et sa dispersion, et même par leur intensité (tableau 2).

Tous ces faits manifestent clairement une **utilisation linguistique des variations du Fo. La mélodie est , dès 9 mois, devenue intonation dans les situations d'interaction avec l'adulte.** Certes, ce phénomène est encore peu développé car il y a 11% de PL contre 78% de Jasis à 9 mois, le reste se partageant entre les catégories intermédiaires. Mais l'évolution ultérieure du langage enfantin entre un et deux ans confirme et précise ces premières conclusions.

CATEGORIE	NIVEAU ELEVE (début > Fo-u)	CONTOUR ASCENDANT	Fo ELEVE (>420Hz)	ET ORAND	INTENS FORTE	Durée totale de l'énoncé brève
FONCTION ETABLISSEMENT DE LA COMMUNICATION	+	+	+	+	+	+
° APPELS	+	+	+	+	+	+/-
° PHATIQUES	+	+/-	-	+	+/-	+
° QUESTIONS	+	+	+	-	-	+/-
° CHARME	+	+/-	+	+	-	-
FONCTION MAINTIEN DE LA COMMUNICATION	-	-	-	-	-	-
° ENONCIATIVES	-	-	-	-	-	-
° NON CLASSES	+/-	+/-	-	-	-	-

Tableau 2: Traits oppositifs des diverses catégories d'énoncés du PL

II. STRUCTURATION MELODIQUE ENTRE 1 ET 2 ANS

Avec l'apparition du langage articulé, les modalités langagières présentes antérieurement dans le premier PL continuent à exister et de **nouvelles modalités émergent**, telles que l'exclamation, l'injonction... Ces modalités peuvent se greffer, soit sur du PL qui continue à exister encore longtemps, soit sur des mots isolés, soit sur des combinaisons de plusieurs unités lexicales, soit sur des phrases mixtes contenant à la fois du PL et du langage référentiel. Nous étudierons ici une seule catégorie langagière, celle de l'ensemble des **énoncés de type phatique** (c'est-à- dire servant à établir la communication), car c'est une des premières à apparaître, et à être largement utilisée.

Dans le PL des mois 8 à 11, ces énoncés se subdivisent en deux catégories, des phatiques très généraux , et des appels, qui s'opposent globalement, par leurs points communs, à toutes les autres catégories d'énoncés (tableau 2). Ces points communs sont les suivants :

- apparition toujours liée à la manifestation d'une présence adulte (bruits, voix, téléphone...);
- articulation ferme et stable;
- apparition essentiellement dans les énoncés de structure V ou CV ;
- intensité forte à très forte (> à 37 dB);
- apparition d'une proéminence sur la syllabe finale;
- utilisation majoritaire de contours simples à pente nette
- voix située dans les zones aiguës de la tessiture, entre le haut du niveau 2 et le niveau 3; 77.7 % des énoncés dépassent 420 Hz, dont 24 % dépassent même 510 Hz; cependant dynamique relativement réduite : la voix se déplace uniquement dans une zone bien délimitée; la dispersion est peu importante, car les fréquences extrêmes ne dépassent que très occasionnellement 600 Hz.

En revanche, les deux sous-catégories se distinguent l'une de l'autre par la qualité des divers paramètres employés, comme le montre le tableau ci-dessous.

APPELS	AUTRES PHATIQUES
DUREE :	
- plus longs (limites 18-236 cs)	- plus brefs (limites 19-68 cs)
INTENSITE	
- plus intenses	- un peu moins intenses
Fo (Fig 4)	
-Fo moyen: 452 Hz, ET: 76Hz 75 % entre 380-500 Hz	- Fo moyen: 411 Hz, ET: 55Hz 75 % entre 260-440 Hz
CONTOURS	
- Courbes ascendantes (montée rapide)	- courbes essentiellement descendantes mais aussi plates ou ascendantes.

De façon générale, on constate donc que **les appels sont plus marqués que les autres phatiques** dans tous leurs paramètres (durée plus longue, intensité plus forte, Fo plus haut, pente plus raide) et que pratiquement seul le contour ascendant y est attesté .

Un point est important à noter ici : ces appels, avec leur forme ascendante, sont une **création** de l'enfant. Ils ne peuvent guère en effet être une imitation des appels de l'adulte qui ont une forme circonflexe [6].

Les autres énoncés à fonction phatique présentent des paramètres moins nets que les appels et surtout une plus grande variété dans l'utilisation des niveaux et des contours. La majorité de ceux-ci sont descendants, dans une gamme restreinte (450-400 Hz), avec pente faible; d'autres sont aigus et plats (600-650 Hz), et d'autres enfin sont ascendants, mais la montée est moins forte et moins raide que dans les appels. Ceci peut expliquer qu'ils ne provoquent pas toujours de réponse de la part de leur destinataire. Toutefois, même pour ce type de contour, la conjugaison des divers indices (I-D-Fo et rupture tonale) permet d'opposer, à l'intérieur des énoncés à fonction phatique, les appels aux non-appels.

Entre 12 et 24 mois, les phatiques globaux subsistent, mais diminuent progressivement en nombre, tandis que les appels proprement dits deviennent de plus en plus fréquents. Ils se subdivisent en appels non articulés et articulés. Les phatiques gardent leurs caractéristiques antérieures, alors que les appels émis en PL sont de plus en plus marqués, avec un Fo plus aigu encore, et une dispersion moindre (Fig.5). Ceci montre que l'enfant choisit des stratégies de plus en plus efficaces pour établir la communication. Mais la catégorie la plus intéressante est celle des appels articulés, dont les caractéristiques changent notablement par rapport aux appels non articulés; Le Fo y baisse sensiblement, tout en restant supérieur à 400Hz (Fig.6); l'intensité subit le même sort; toutefois la plus grande nouveauté réside dans la forme du contour: alors qu'en PL tous les appels étaient ascendants, maintenant le contour peut être soit montant, soit circonflexe ou plat-descendant dans la zone aiguë, supérieure à 400Hz. Le fait que l'enfant dispose de mots lui

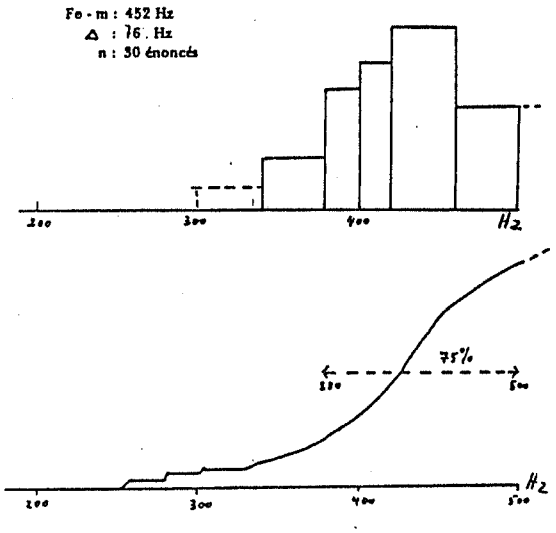
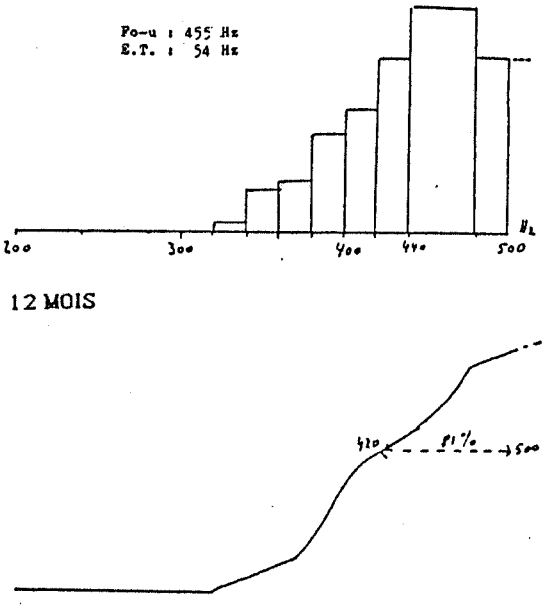


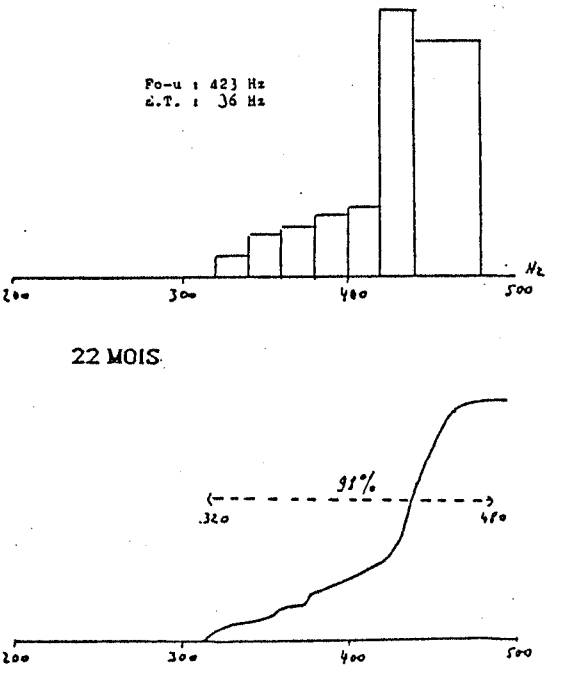
Fig.4 : Fo moyen des appels en PL à 9 mois

permet donc de produire des énoncés moins marqués, car il autorise une redondance moindre. En revanche, lorsque le destinataire de l'appel ne répond pas à la sollicitation enfantine, le locuteur reprend son ancienne technique. Précisons aussi que nous n'avons rencontré ni apostrophes, ni vocatifs avant 24 mois, et que tous les appels attestés ne dépassent pas la longueur de deux syllabes.



12 MOIS

Fig.5 : Fo moyen des appels en PL a 12mois



22 MOIS

Fig.6: Fo moyen des appels articules a 22 mois

L'ensemble de notre étude nous a permis de mettre en relief que la **fonction phatique** (Jakobson) **et la fonction d'appel** (Bühler) **sont, dans l'acquisition du langage, à la fois premières et essentielles**; et c'est précisément pour ces fonctions que l'enfant fait dès le début du PL usage d'une **forme intonative précise, stable et parfaitement contrôlée.**

Nous avons étudié selon la même technique les autres modalités langagières présentes entre un et deux ans. Nous avons ainsi pu montrer qu'il existe une nette opposition, au niveau acoustique, entre les énoncés de type énonciatif, interrogatif, exclamatif, injonctif, dubitatif.., opposition résidant à la fois dans la direction du contour, la hauteur du Fo-moyen, la dispersion de la voix, la pente des énoncés, leur intensité. **Les commutations intonatives sur un même ensemble phonique, qui créent de véritables paires minimales du point de vue prosodique, sont très fréquentes,** car l'enfant ne dispose pas encore des moyens lexicaux (mots outils, tels que particules interrogatives, exclamatives) ou syntactiques qui lui permettraient de situer un énoncé dans une modalité donnée. **La prosodie a donc une fonction linguistique évidente, non seulement en l'absence du lexique (stade du PL) mais aussi une fois que la couche verbale est présente,** car les diverses modalités se séparent plus clairement les unes des autres. Il faut souligner également que l'opposition entre les diverses modalités est dans un premier temps encore plus redondante que dans le langage de l'adulte. Puis, au fur et à mesure que le stock lexico-syntactique de l'enfant s'enrichit, de nouvelles stratégies sont mises en oeuvre, et les divers paramètres prosodiques deviennent généralement moins marqués, comme si la redondance était trop forte. Toutefois, la perte de certains traits ne se fait pas de façon aléatoire; sont évacués ceux qui précisément donnaient au langage ses caractéristiques enfantines, notamment la hauteur trop élévée ou les écarts trop importants. Ainsi, l'enfant rapproche progressivement son système de celui de la langue cible. Toutefois des progrès restent encore à accomplir.

REFERENCES:

[1] G. KONOPCZYNSKI, Acquisition du langage: la période charnière et sa structuration temporelle. Strasbourg : TIPhS 16, 67-131, 1984.
[2] G. KONOPCZYNSKI, La hauteur des voix enfantines, 11° I.C.A., Paris, IV, 259-262, 1983.
[3] G. KONOPCZYNSKI, La voix de l'enfant à la période de babillage, 12° I.C.A., Toronto, 1986 (à paraître).
[4] G. KONOPCZYNSKI, Acquisition du langage: la période charnière et sa structuration mélodique, Bulletin d'Audiophonologie, N.S.1,1-2,63-92,1985.
[5] G. KONOPCZYNSKI, Du prélangage au langage: acquisition de la structuration prosodique, Thèse de doctorat d'Etat, Strasbourg, avril 1986.
[6] A. DI CRISTO, Des traits acoustiques aux indices perceptuels. Application d'un modèle d'analyse prosodique à l'étude du vocatif en français, Aix-en-Provence, TIPA 3, 213-358, 1976.

UN MODELE DE CONGRUENCE RELATIONNEL POUR LA SYNTHESE DE LA PROSODIE DU FRANCAIS

Gérard BAILLY

INRS-Télécommunications
Université du Québec
Ile-des-Soeurs, Québec, Canada, H3E 1H6

Abstract

A system for automatic generation of French prosody from unrestricted text has already been presented [1]. The following article deals with the essential part of this system : the congruency model. The strategy followed by our system for deriving the prosodic structure of the utterance from the deep syntactic structure of the text is described. The system of cues generated by our multiparametric model of French prosody is then presented.

1. Introduction

L'influence de la prosodie sur le naturel et l'intelligibilité de la parole n'est plus à démontrer. Un modèle de prosodie constitue pour ceci une partie essentielle d'un système de synthèse du texte. La conception d'un modèle de prosodie se heurte hélas à la multiplicité des fonctions assumées par l'ensemble des paramètres prosodiques [2] : identification du locuteur et de son état émotionnel, mise en valeur de la structuration syntaxique et sémantique de l'énoncé, désambiguisation lexicale et syntaxique, emphase ... De récentes études sur le rôle de ces paramètres en perception de la parole [3] ont montré cependant que ceux-ci assumaient une fonction linguistique importante.

De nombreux modèles ont été proposés pour rendre compte de la fonction de structuration syntaxique assumée par la prosodie [4,5,6,2,7...]. Ces modèles sont basés sur deux principes fondamentaux :

1) L'aspect segmental de la prosodie : il est possible d'extraire un certain nombre d'indices acoustiques des continuums mélodiques, rythmiques et d'intensité. Ces indices découpent le message en unités minimales de sens, appelées groupes intonatifs, accentuels ou de sens selon l'auteur. Nous adopterons quant à nous la dernière notation.
2) La relation de congruence entre structuration syntaxique (de surface ou profonde) de l'énoncé et la structuration prosodique du message : les groupes de sens cités plus haut peuvent être mis en relation avec les unités syntaxiques de l'énoncé.

Les modèles cités plus haut se distinguent essentiellement sur trois points :

1) La nature des indices distinctifs : articulatoires, acoustiques, perceptifs et le nombre de paramètres modélisés.
2) L'extension des groupes de sens.
3) La nature de la relation de congruence entre structures.

Nous présentons ici un modèle multiparamétrique de la prosodie du français. La relation de congruence entre structure consiste en une projection de la structure profonde de l'énoncé sur l'axe temporel. Cette projection réalise un marquage du message. Ce marquage et la détermination des groupes de sens permettent de structurer la phrase par des indices spécifiques à chaque espace paramétrique.

La génération automatique de cette prosodie à partir du texte en utilisant une analyse lexicale minimale du texte a été réalisée [1] et couplée au sytème de synthèse du texte de l'I.N.R.S. Des tests comparatifs de naturel ont produit des résultats très encourageants pour cette méthode inductive vis à vis d'une méthode déductive [8,9].

2. Modèle relationnel

2.1 Modèle syntaxique

Le modèle syntaxique que nous utilisons est présenté dans [10,11]. Une grammaire de dépendance exprime les relations syntaxiques que les divers mots lexicaux entretiennent entre eux. Ces relations énumèrent les classes lexicales dépendantes qu'une classe lexicale peut gouverneur [11]. On construit ainsi un arbre de dépendance abstrait dont le sommet est en général le verbe principal de la phrase (cf. fig.1).

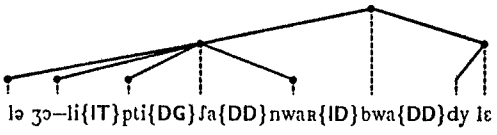


Fig. 1 Exemple d'arbre syntaxique et des marqueurs générés.

2.2 Modèle de congruence

Le modèle de congruence permet de réaliser un marquage du texte par projection sur l'axe syntagmatique. Chaque mot lexical non-clitique et donc accentuable (mots de contenu : noms, adjectifs, verbes...) recoit ainsi une marque corrélative de la relation entretenu par celui-ci et le non-clitique subséquent. Ainsi que le souligne MARTIN [4], ces relations sont quadrivalentes : indépendance (ID : aucun lien de dépendance direct entre les deux mots), interdépendance (IT : les deux mots sont dépendants du même gouverneur), dépendance gauche (DG : le premier mot est dépendant du deuxième) et droite (DD : le deuxième est dépendant du premier) (cf.Fig. 2).

Fig. 2 Relations de dépendance quadrivalentes.

Les marques ainsi générées indiquent des relations plus ou moins faibles entre les mots. La force de la relation syntaxique se traduit par une hiérarchisation de la force des marques : c'est une généralisation de la notion de frontière mineure et majeure [2]. La force des diverses marques peut être renforcée par la présence de ponctuations ou par la nature fonctionnelle des deux constituants. Ceci peut conduire à l'élaboration de systèmes très complets de marqueurs prosodiques spécialisés, qui se révèlent hélas difficiles à générer automatiquement [12].

Nous distinguons deux forces : faibles et fortes. Notre sytème comporte six marques relationnelles classées par force décroissante :

- L'indépendance forte : indépendance faible confirmée par une ponctuation (IF)
- L'indépendance faible (ID)
- L'interdépendance (IT)
- La dépendance gauche forte (GF) : dépendance entre nom sujet et verbe subséquent
- La dépendance gauche faible (DG)
- La dépendance droite (DD)
- Et deux marques d'emphase † : interrogative (EI), placée sur les mots interrogatifs et simple (EM), placée sur les négations et démonstratifs.

Ces marques révèlent les relations locales entre les divers éléments sur l'axe syntagmatique. La structure prosodique qui sera ainsi générée n'inclut donc pas la connaissance a-priori de la structure complète de la phrase. Notre modèle exclut donc la structuration complète de l'énoncé à partir de l'intonème terminal [4].

La génération automatqiue de ces marqueurs ne nécéssite pas l'analyse syntaxique profonde de toute la phrase : il est donc possible de marquer des phrases non-grammaticales et d'envisager la génération automatique de ces marques par des règles contextuelles [1].

2.3 Modèle de structuration

La structure de surface du message est à deux niveaux :

1) le message est d'abord fractionné en groupes de phonation séparés par des pauses.
2) puis chaque groupe de phonation est segmenté en groupes de sens.

Cette structuration est soumise à des contraintes à la fois linguistiques et rythmiques. La présence d'une pause et sa durée éventuelle dépendent de la force de la frontière qu'elle marque et du débit du locuteur [13]. L'effacement de marques et la constitution de groupe de sens de la taille de syntagmes sont influencée par les mêmes facteurs [14]. Les contraintes rythmiques sont exprimées en nombre de syllabes par unité : en français, les valeurs moyennes sont de 8 à 12 pour un groupe de phonation et de 4 pour un groupe de sens, pour un débit normal de 220 syll./mn.

Les groupes de phonation et de sens sont donc déterminés par une procédure récursive appliquée sur chaque phrase (cf.Fig. 3). Elle examine la possibilité d'effacement de marques en commencant par les plus faibles et en tenant compte des contraintes rythmiques données plus haut.

Dans le système de génération prosodique automatique, cette étape consiste donc en l'effacement des certains marqueurs afin de constituer des groupes de sens plus grands que le mot, et en l'insertion de pauses, et l'effacement des liaisons conditionnelles générées par la traduction orthographique-phonétique afin de constituer des groupes de souffle.

3. Modèle multiparamétrique

3.1 Stratégie

Le modèle prosodique permet de générer automatiquement à partir de la structure décrite plus haut les divers paramètres acoustiques. La structure permet d'assurer une synchronisation des indices dans chaque espace paramétrique. Cependant les niveaux de structure pris en compte

† ces marques ne participent pas à la structuration prosodique du discours. Leur fonction est locale et emphatique.

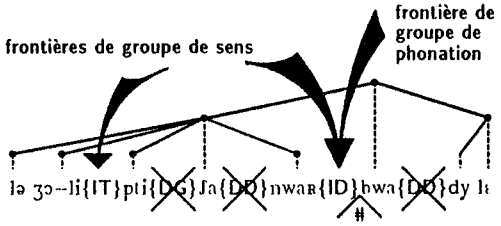


Fig. 3 Génération des groupes de sens (seuil=4 syll.)et des groupes de phonation (seuil=8 syll.). Effacement des marques et insertion des silences.

par chaque modèle paramétrique dépend de la stratégie du locuteur. Celle que nous proposons permet de produire une prosodie référentielle qui se prête parfaitement à la lecture de texte.

3.2 Génération de la durée

3.2.1 Durées phonémiques

Les facteurs d'allongement ou d'accélération de la durée phonémique sont multiples et agissent concurremment. On peut distinguer deux types de stratégies :

- modèle multiplicatif [9] : il génère des pourcentages P_i d'allongement ou d'accélération de la durée phonémique de base selon la formule :

$$\text{durée} = \text{durée\_intrinsèque} \times \prod_i (1 + P_i)$$

- modèle additif [15] : il génère des durées D_i positives ou négatives s'ajoutant à une durée phonémique de base :

$$\text{durée} = \text{durée\_intrinsèque} + \sum_i D_i$$

Bien que les deux modèles ait conduit à des résultats acceptables, le modèle multiplicatif rend compte plus aisément de la relative incompressibilité de certains évènements acoustiques.

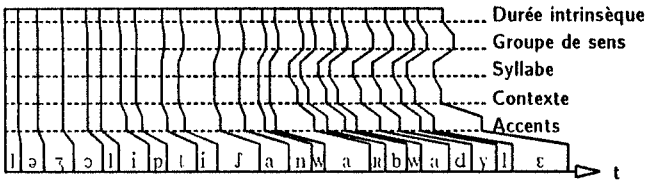


Fig. 4 Structuration rythmique

Notre modèle multiplicatif exploite 4 types d'informations (cf.Fig 4) : 1) structurelle : écart de certains niveaux de la structure par rapport à des unités idéales (4 syll./groupe de sens, 2 syll./mot, 2 phonèmes/syll.), 2) accentuelle : contours de durée spécifique à chaque marqueur (la présence de pause est pris en compte dans ces contours) 3) contextuelle : influence co-intrinsèque des phonèmes précédent et subséquent, 4) débit d'élocution.

Nous tenons compte de la relative incompressibilité des consonnes par deux règles : 1) facteurs P_i réduit de moitié, 2) durées minimales de réalisation pour tous les phonèmes.

3.2.2 Durée des pauses

StBonnet et Boe [13] ont montré que si la distribution des pauses était hiérarchisée par la force des jonctures et la vitesse d'élocution, le rapport *durée de silence / durée de phonation* était pratiquement indépendant de la vitesse d'élocution mais dépendait seulement du style de discours. Ce quotient est de l'ordre de 0.4 en moyenne pour de la parole lue. La durée globale de silence ainsi déterminée sera répartie entre chaque pause selon leurs fonctions respectives.

Chaque pause peut avoir essentiellement trois fonctions : respiratoire syntaxique ou démarcative [6]. Nous ne rendons compte que des deux premières fonctions, la fonction démarcative n'étant pas systématique.

a) Pauses respiratoires

Les pauses respiratoires sont les pauses les plus longues. Elles sont choisies parmi les pauses précédemment déterminées (cf.§ 2.3) par un critère de durée inter-pauses : elles sont situées aux jonctures les plus fortes et sont distantes d'environ 3 s pour un débit normal. Leur position est déterminée par la même procédure récursive qu'au § 2.3 avec un seuil syllabique égal au nombre de syllabes total de la phrase multiplié par le rapport de la durée de la phrase par la durée inter-pauses. Leur durée relative sera fonction de la durée des groupes respiratoires qu'elles délimitent.

b) Pauses syntaxiques

Les pauses restantes sont qualifiées de syntaxiques. Leur durée relative est fonction de la force de la jonction.

3.3 Génération de la mélodie

Bien que la durée soit en français la seule marque corrélative de l'accent, la mélodie a fait l'objet de la majeure partie des études prosodiques du fait de son indépendance vis à vis des éléments segmentaux. Les indices retenus par les divers modèles intonatifs peuvent être classés en trois catégories : 1) commandes [7], 2) contours [6,4], 3) cibles [16]. Ces modèles correspondent aux trois états de la substance prosodique : articulatoire, acoustique ou perceptuelle.

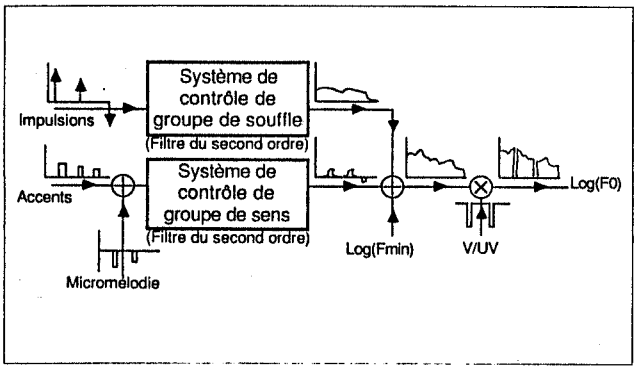


Fig. 5 Modèle de Fujisaki modifié

Notre analyse des faits d'intonation utilise un modèle de commande de source vocale [17]. Cette modélisation nous permet d'avoir accès à des indices indépendants et d'interprétation physiologique immédiate.

Ce modèle de la source vocale considère que le logarithme de la fréquence fondamentale est proportionnel à l'élongation des cordes vocales. Cette élongation est alors égale à la somme de trois valeurs indépendantes (cf.fig. 5) :

- l'élongation minimale fixant le registre de voix du locuteur
- la sortie d'un système de contrôle de commandes de souffle : effet de l'abaissement de la cage thoracique en cours de phonation sur le fonctionnement laryngien [18].
- la sortie d'un système de contrôle de commandes d'accent : commande de tension des cordes vocales.

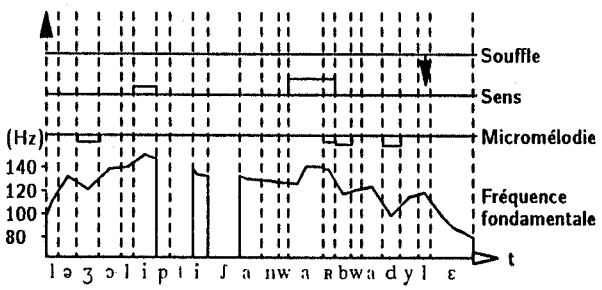


Fig. 6 Structuration mélodique

Les fonctions assumées par les deux principales commandes du modèle permettent de segmenter l'énoncé en deux types de regroupement : groupes de souffle et de sens (cf.fig 6).

3.3.1 Groupes de souffle

Les groupes de souffle sont bornés par les commandes du système de contrôle de commandes de souffle. Les impulsions de commande peuvent être classées en trois catégories :

- Les impulsions d'initialisation : cette commande est placée au début de chaque groupe de souffle précédent une pause respiratoire. Elle modélise la reprise de souffle par inhalation. L'amortissement du système de contrôle permet de rendre compte de l'effet de déclinaison [16].
- Les impulsions de réinitialisation : cette commande est générée lorsque le groupe d'inhalation excède 3s. Elle modélise la présence de réinitialisation du fondamental à des valeurs comparables à celle du début du groupe d'inhalation ceci sans inhalation. Cet effet peut s'expliquer par l'intervention des muscles du diaphragme en cours de phonation par un mécanisme de relai [18].
- Les impulsions finales : ce sont les impulsions de modalité classiques : les phrases déclaratives, exclamatives et interrogatives introduites recoivent une impulsion négative sur la dernière syllabe et les phrases interrogatives non introduites une impulsion positive.

3.3.2 Groupes de sens

Les commandes de groupe de sens assument une fonction essentiellement linguistique. Chaque groupe de sens est affecté d'un accent d'extension et d'amplitude fonction de la marque qui l'affecte. L'amplitude est fonction de la force de la marque. La typologie accentuelle finale du français impose un début d'accent sur la dernière syllabe du groupe, la fin de l'accent est située soit sur la fin de la dernière syllabe du groupe (accent démarcatif) soit sur la fin de la première syllabe du groupe suivant (accent intégratif).

Les marques d'emphase échappent à la typologie accentuelle finale : l'accent emphatique est initial. La première syllabe du mot qu'elles affectent recoit donc un accent mélodique d'amplitude moyenne. Le dernier mot d'un phrase exclamative recoit lui aussi la même emphase, constituant ainsi le "plateau mélodique" caractéristique du vocatif en français [19].

3.3.3 Micromélodie

Une commande micromélodique a été ajoutée au modèle initial de Fujisaki afin de modéliser les "creux" locaux de F0 dus à la présence de consonnes voisées ([b], [d], [g], [v], [z], [ʒ], [m], [ʀ]) dans le discours. Contrairement aux commandes précédentes qui alimentent deux sytèmes aux caractéristiques de filtre constantes et caractéristiques du locuteur, les caractéristiques de filtre associées aux commandes micromélodiques dépendent du mode d'articulation du phonème correspondant.

3.4 Génération de l'intensité

L'intensité est le corrélat acoustique de la pression subglottique et du timbre. Bien que la pression subglottique reste pratiquement constante en cours de phonation, notre analyse de 32 courbes d'intensité [7] ont

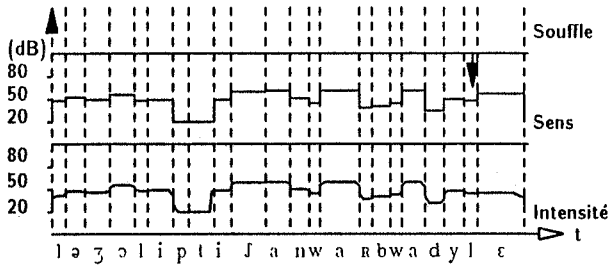


Fig. 7 Structuration de l'intensité

fait apparaître une déclinaison d'environ 2 à 3 dB de l'intensité au cours de chaque groupe de souffle.

Afin de générer la courbe d'intensité, le modèle de fujisaki est à nouveau employé et alimenté par les commandes suivantes (cf.Fig. 7) :

- Chaque groupe de souffle précédent est affecté de commandes d'intensité. Les impulsions d'intensité finales seront cependant toutes négatives. Les amplitudes sont calculées afin de produire une pente de déclinaison de 1 dB/s et une chute finale de -10 dB.
- Le sytème de contrôle de groupe de sens est alimenté par une fonction par plateaux. Chaque plateau correspond à l'énergie intrinsèque de chaque phonème diminuée de 12 dB.

La commande de voisement du modèle permet de générer les occlusions non-voisées du conduit vocal et les silences. L'intensité de référence permet de régler le volume global de la parole de synthèse.

4. Conclusions

Le système de génération de prosodie pour le français décrit plus haut a été couplé au synthétiseur de l'I.N.R.S. Télécommunications LOQUAX. Son fonctionnement automatique a été rendu possible grâce à l'élaboration d'une analyse pré-syntaxique du texte [1], qui permet de générer automatiquement les marqueurs relationnels. L'évaluation subjective du naturel de la parole de synthèse versus l'algorithme de génération prosodique précédemment implanté [20] démontre la faisabilité de l'implantation automatique du modèle et l'amélioration qualitative qui en résulte.

L'apport en qualité de ce modèle peut être justifiée par deux points essentiels : 1) une meilleure connaissance de la nature des frontières prosodiques du discours, 2) une structuration prosodique plus complète par des indices acoustiques indépendants. L'indépendance des indices permet de plus à ce modèle d'augmenter facilement sa compétence sans perturber la prosodie référentielle déjà produite.

Remerciements

Je remercie Mr D.O'Shaughnessy pour avoir suivi et conseillé ce travail et commenté les premières versions de cet article.

Références

1. Bailly G., "Multiparametric generation of French prosody from unrestricted text", *IEEE Int. Conf. on Acous, Speech and Sig. Proc.*, à paraître, 1986.
2. Di Cristo A., *Aspects phonétiques et phonologiques des éléments prosodiques*, Modèles Linguistiques, 3.2, pp.5-24, 1981.
3. Nakatani L.H., Schaffer J.A., "Hearing "words" without words: prosodic cues for word perception", *J. Acoust. Soc. Amer.*, 63, pp.234-245, 1978.
4. Martin P., "Phonetic realisations of prosodic contours in french", *Speech Communication*, 1, p.283-294, 1982.
5. Dell F., Hirst D., Vergnaud J.R., *Les représentations en phonologie*, Hermann, Paris, 1983.
6. Emerard F., *Synthèse par diphones et traitement de la prosodie*, Thèse de 3ème cycle, Grenoble, 1977.
7. Bailly G., *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, Grenoble, 1983.
8. O'Shaughnessy D., "Fundamental frequency by rule for a text-to-speech system", *IEEE Int. Conf. on Acous., Speech and Sig. Proc.*, pp.571-574, 1977.
9. O'Shaughnessy D., "A study of French vowel and consonant durations", *Journal of Phonetics*, 9, pp.385-406, 1981.
10. Hays D.G., "dependency theory: a formalism and some observations", *Language*, 46, pp.511-525, 1964.
11. Courtin J., *Algorithmes pour le traitement interactif des langues naturelles*, Thèse d'état, Grenoble, 1977.
12. Divay M., Guyomard G., "Positionnement d'opérateurs prosodiques dans un texte écrit", *Actes Congrès A.F.C.E.T.*, pp.317-324, 1981.
13. StBonnet M., Boe L.J., "Les pauses et les groupes syntaxiques, leur durée et la distribution en fonction de la vitesse d'élocution", *8èmes Journées d'Etudes sur la Parole*, pp.337-343, 1977.
14. Vaissiere J., "On French prosody", *Quaterly Progress Report*, Res. Lab. of Electronics, MIT, pp.251-261, 1974.
15. O'Shaughnessy D., "Consonant durations in clusters", *IEEE Trans. on Acous., Speech and Sig. Proc.*, 22, pp.282-295, 1974.
16. Pierrehumbert J., "Synthesizing intonation", *J. Acoust. Soc. Amer.*, 70, pp.985-995., 1981.
17. Fujisaki H., Hirose K., "Analysis and synthesis of voice fundamental frequency contours of spoken sentences", *IEEE Int. Conf. on Acous, Speech and Sig. Proc.*, pp.950-953, 1982.
18. Maeda S., *A characterization of American English intonation*, Ph.D.Dissertation, M.I.T., 1976.
19. Di Cristo A., *De la microprosodie à l'intonosyntaxe*, Thèse d'état, Université de Provence, 1978.
20. O'Shaughnessy D., "Design of a real-time french text-to-speech system", *Speech Communication*, 3, pp.233-243, 1984.

PATRONS D'ACCENTUATION DANS LES MOTS COMPOSES ET LES GROUPES DE MOTS EN NEERLANDAIS

A.C.M. Rietveld, A.M. Slootweg

Instituut voor Fonetiek, Katholieke Universiteit Nijmegen

ABSTRACT

In this paper we describe two experiments concerned with primary, secondary and zero accent in phrases and compounds. These accents are investigated as they are perceived by untrained listeners in a setting where the answers obtained are based on acoustic properties of the syllables only. To find out which are the relevant phonetic parameters, we studied acoustic differences between syllables with different degrees of accent, as they were measured in reiterant speech of four speakers. In this way we can describe, in a quantative fashion, the differences between perceived degrees of accent.

INTRODUCTION

Depuis le 'Sound Pattern of English' [1] on a élaboré toutes sortes de règles qui décrivent les patrons d'accentuation sur la base d'information phonologique/morphologique. Il y a une constante dans la littérature sur le nombre de degrés d'accentuation à distinguer dans les domaines perceptifs et acoustiques: on dit qu'il nous manque des données empiriques sur ce sujet. Chomsky et Halle sont d'avis que la capacité d'un auditeur de distinguer des niveaux d'accentuation repose surtout sur la connaissance qu'il a de la structure de l'énoncé en question. Nous nous proposons de rapporter sur les résultats de quelques expériences concernant les patrons d'accentuation dans les mots composés et les phrases en néerlandais. Les patrons d'accentuation sont, d'après la théorie phonologique, les suivants: (exemple)

mot composé : mo^{1}de ki^{2}mono

phrase (AD + N) : mo^{2}ie ki^{1}mono

On observe une permutation des accents qui est analogue, d'ailleurs, à celle de l'anglais. Plusieurs chercheurs, comme Klatt [2] sont d'avis que dans le domaine acoustique on n'a à distinguer que trois types de syllabes: a) les syllabes accentuées, b) les syllabes non-accentuées et c) les syllabes réduites. Cela impliquerait que dans l'exemple ci-dessus les syllabes marquées de '2' se comporteraient comme les syllabes non-accentuées qui ne contiennent pas de ə-muet. Voilà le thème de notre étude.

L'ANALYSE ACOUSTIQUE

Toutes sortes de problèmes se présentent dans l'analyse acoustique qui a pour but de découvrir des régularités prosodiques dans la parole. Ces problèmes proviennent de phénomènes comme a) durée, intensité et FO intrinsèques des segments en question, b) l'influence du nombre de syllabes sur la durée et c) l'influence du nombre et du type des segments dont les syllabes sont composées. Plusieurs chercheurs ont eu recours (cf. Nakatani & Schaffer [3]) à des imitations simples de la parole, nommées 'reiterant speech'; ces imitations consistent en des séquences de syllabes identiques comme mamamama. On sait (cf. Nooteboom [4]) que dans ces imitations les différences temporelles phonémiques se perdent; seules les différences suprasegmentales subsistent. Evidemment, un problème supplémentaire se pose. On introduit une transformation dans deux domaines. Le sujet doit percevoir les régularités rhythmiques et il doit les convertir en mouvements articulatoires. Il est possible que, quelque part au cours de ce processus, les nuances que nous cherchons, se perdent. Il est également possible que des régularités qui n'existent pas dans le signal original, sont créées dans la perception. Dans ces deux cas l'imitation n'est pas une bonne représentation de la parole qui est l'objet de l'analyse.

Corpus

Le corpus à imiter se composait de mots trisyllabiques portant l'accent de mot sur la première, deuxième et troisième syllabe respectivement. En combinant ces mots avec un nom préposé nous avons formé des mots composés, en les combinant avec un adjectif nous avons obtenu des phrases. Exemple:

nom:	nom composé:	phrase:
ba^{1}riton	modebariton	rode bariton
defi^{1}le	modedefile	rode defile
ki^{1}mono	modekimono	rode kimono

(rode = rouge)

Les imitations ont été réalisées par 4 sujets à l'aide de répétitions de la syllabe ma. Les sujets ont été instruits d'imiter la syllabe contenant un ə-muet par la syllabe me; cela facilitait les imitations et donnait aux autres voyelles un caractère plus différent du schwa.
Vu qu'il y avait trois représentants de chaque patron, on disposait, dans la condition de 'phrase' ou de 'mot composé', de douze imitations de chaque type (4 locuteurs x 3 représentants). Le nombre total montait à 12 x 3 (patrons) x 2 (phrases et mots composés) = 72; 288 syllabes en total.
Nous avons mesuré les durées, intensités maximales et les mouvements de la F0 maximaux de chaque syllabe.

Résultats

Le déplacement des accents dans les mots composés et les phrases se manifeste clairement dans les quotients des durées, F0's et intensités des syllabes en question (Figure 1).

	Durée	Intensité	F0
mots composés	1.07	1.10	4.16
phrases	0.83	0.94	0.21

Figure 1 : Quotients des valeurs de 3 paramètres mesurés dans les syllabes qui portent les 2 accents théoriques

Une analyse discriminante avec comme variables prédictrices les quotients des durées et intensités des syllabes portant l'accent primaire et secondaire montre que les phrases et les mots composés ne diffèrent pas seulement dans la fondamentale (c'était prévu), mais aussi dans les autres paramètres. La classification qui résulte de cette analyse discriminante est donnée dans la figure 2.

	mot composé	phrase
mot comp.	88.9%	11.1%
phrase	8.3%	91.7%

Figure 2a: Classification des imitations de phrases et mots composés qui résulte d'une analyse discriminante sur la base de deux paramètres: durée et intensité.

intensité	0.94
durée	0.35

Figure 2b: Corrélations entre les variables prédictrices et la fonction discriminante

Si les deux structures linguistiques 'mot composé' et 'phrase' diffèrent, nous ne savons pas encore si les syllabes portant l'accent secondaire, ont un caractère qui est différent des syllabes qui ne portent pas d'accent primaire ou secondaire.

Afin de vérifier si les syllabes qui portent, en théorie, l'accent secondaire, se distinguent des autres syllabes, nous avons effectué une analyse discriminante sur 3 groupes de syllabes:
a) les syllabes portant l'accent primaire
b) les syllabes portant l'accent secondaire
c) les autres syllabes.
L'analyse a été faite avec les valeurs absolues de trois variables (mouvement maximal de la F0, durée et intensité maximale) et la valeur d'une variable 'relative': le quotient de la durée d'une syllabe et la durée totale d'une imitation.

	F0	Durée	Intensité	Durée/durée totale
accent:				
1	41.3	216	39.3	0.317
2	10.7	193	36.3	0.280
0	16.6	137	35.8	0.201

Figure 3: Valeurs de quelques paramètres acoustiques de 3 types de syllabes: syllabes portant l'accent primaire (1), syllabes portant l'accent secondaire (2) et les autres syllabes (0) dans les imitations de phrases et de mots composés. F0 exprimée en Hz, durée en msec et intensité en dB.

Inspection de la figure 3 nous laisse supposer qu'une classification des trois groupes de syllabes doit être possible. La fondamentale distingue les syllabes avec l'accent primaire des autres syllabes, tandisque les autres paramètres semblent tous contribuer à une classification en trois groupes. L'analyse discriminante confirme cette hypothèse. A l'aide de deux fonctions discriminantes, formées par des combinaisons linéaires de trois variables (quotient des durées, intensité et F0) une classification satisfaisante a été effectuée, comme le montre la figure 4.

	1	2	0
1	93.1%	6.9%	0.0%
2	4.2%	80.6%	15.3%
0	0.0%	8.3%	91.1%

Figure 4a: Résultats de la classification en trois groupes de syllabes à l'aide de deux fonctions discriminantes

	Fonct.I	Fonct. II
quot. durées	0.80	-0.58
intens.	0.39	0.38
F0	0.57	0.70

Figure 4b Corrélations entre les variables prédictrices et les fonctions discriminantes

Les positions des trois types de syllabes dans l'espace formée par les deux fonctions discriminantes est donnée dans la figure 5.

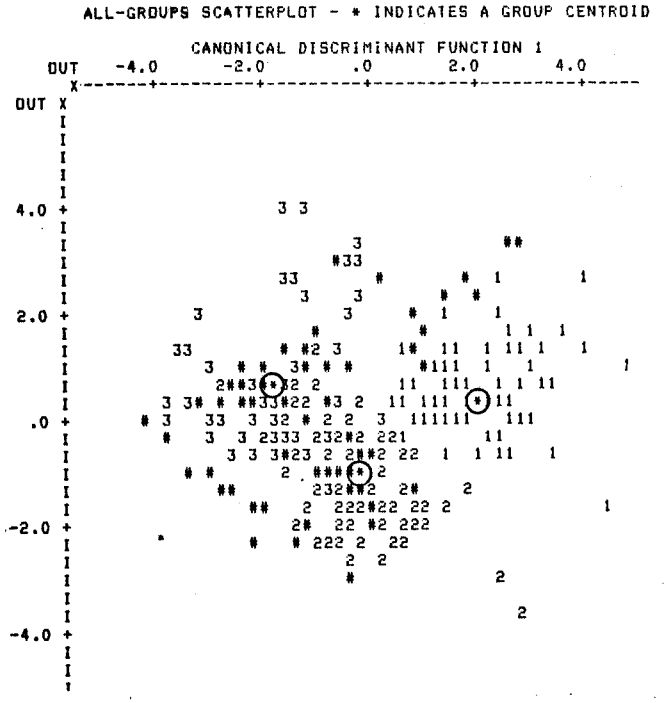


Figure 5: L'espace formée par 2 fonctions discriminantes et les positions des 3 types de syllabes; 1 = syllabe portant l'accent primaire, 2 = syllabe portant l'accent secondaire, 3 = syllabe sans accent. ○indique la centroïde d'un groupe.

La seule conclusion q'on peut tirer de l'analyse acoustique est que plus d'une seule distinction dans les syllabes de mots composés et de phrases a été réalisée par nos sujets: syllabes portant l'accent primaire, syllabes portant l'accent secondaire et autres syllabes.

Notre première expérience a démontré que si des auditeurs entendent des mots composés ou des phrases, ils y perçoivent une structure qui peut être décrite à l'aide de plusieurs degrés d'accentuation. Nous ne savons pas, pourtant, si les imitations ont été réalisées de la sorte parce que les auditeurs étaient conscients de la forme phonologique sous-jacente des énoncés à imiter. Il est toujours possible que les distinctions plus subtiles sont le produit de la perception plutôt qu' une manifestation de la façon dont les mots originaux ont été réalisés.

TEST PERCEPTIF

Dans la deuxième expérience nous avons effectué un test perceptif selon la méthode des comparaisons par paires. Un seul locuteur a prononcé le même type de mots qu'on a employé dans l'analyse acoustique, c.a.d. des mots composés et des phrases avec trois patrons d'accentuation différents. De chaque mot ou phrase nous avons isolé trois syllabes, dont deux portaient l'accent primaire ou secondaire; la troisième syllabe était sans accent. Les mots ont été prononcés dans une phrase complète.

Les paires de syllabes ont été présentées à 25 auditeurs sans connaissances phonétiques. Les membres d'une paire étaient séparées d'un intervalle de 200 ms, les paires avaient des intervalles de 2000 ms. Les sujets ont été instruits d'indiquer quelle syllabe de chaque paire était la plus accentuée.

Résultats

Les matrices de dominance contenant les fréquences des jugements ont été soumises à la méthode d'échelle de Thurstone; cette analyse résulte en des échelles d'intervalles jugés équidistants. Les échelles obtenues pour chaque énoncé, avec les positions relatives des syllabes en question (0 = sans accent, 1 = accent primaire et 2 = accent secondaire) sont données dans la figure 6. Les échelles elles-mêmes représentent le degré d'accentuation des syllabes perçu par les auditeurs.

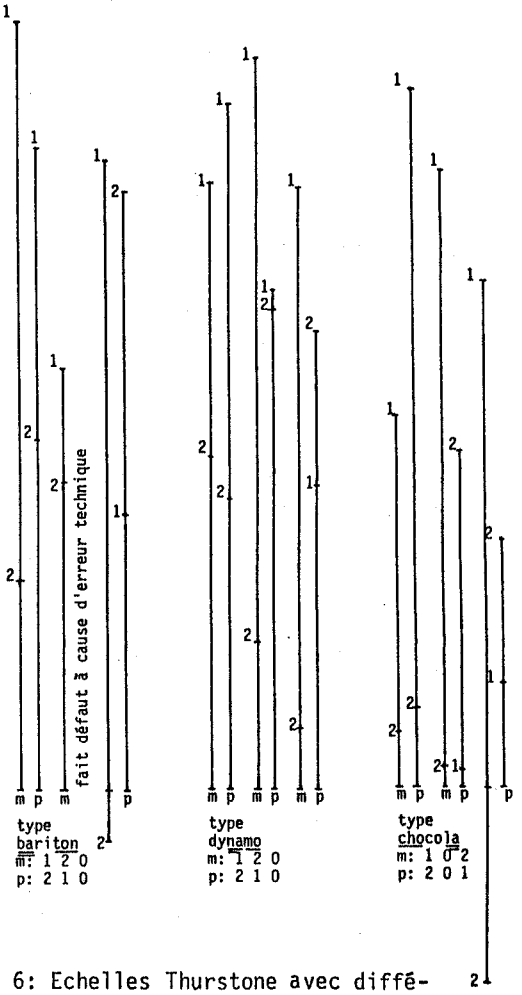


Figure 6: Echelles Thurstone avec différences d'accentuation entre les syllabes présentées en paires. Valeurs ajustées à zéro pour les syllabes sans accent. 1 = accent primaire, 2=acc.second. m = m.comp.,.p = phrase.

Dans la plupart des cas les syllabes portant l'accent primaire se trouvent tout en haut des échelles. Cela veut dire que ces syllabes-là ont été jugées comme plus accentuée que les autres. Les positions des autres syllabes sont moins stables, bienque là aussi, les syllabes avec l'accent secondaire (2) dominent les syllabes sans accent. Il est fort possible, pourtant, que les jugements ne reflètent pas seulement les degrés d'accentuation, mais aussi la tendance descendante de la FO, connue sous le nom de 'déclinaison'. Par exemple, dans les mots composés avec le patron 1 0 2, la syllabe portant l'accent secondaire occupe la dernière position dans le mot. L'effet en est que cette syllabe a été jugée comme très proche de la syllabe sans accent, ou même comme moins accentuée que celle-ci. Si l'on présente des syllabes en (quasi-) isolation, l'influence de la FO semble l'emporter sur l'influence des autres paramètres.

CONCLUSION

L'analyse acoustique des imitations de phrases et de mots composés en néerlandais a confirmé l'existence d'au moins deux degrés d'accentuation. Cependant, il n'est pas sûr si les nuances acoustiques mesurées dans les imitations se manifestent de la même faç on dans la parole 'normale'.
Un test perceptif dans lequel des paires de syllabes ont été jugées par des auditeurs montre qu'une classification perceptive de syllabes quasi-isolées est moins stable. Nous supposons que l'effet de la déclinaison de la FO y joue un rôle important.

REFERENCES

[1] N. Chomsky, M. Halle, "The sound pattern of English", New York, 1968.

[2] D.H. Klatt, "Linguistic uses of segmental duration in English: acoustic and perceptual evidence", JASA, no. 59, pp. 1208-1221, 1976.

[3] L.H. Nakatani, J.A. Schaffer, "Hearing words without words: prosodic cues for word perception", JASA, no. 63, pp. 234-245, 1978.

[4] S.G. Nooteboom, "Temporal patterns in Dutch", Proc. 7th Int. Congress of Phon. Sciences, pp. 984-989, 1972.

DE QUELQUES CARACTERISTIQUES RYTHMIQUES ET MELODIQUES DE LA COMPTINE "UNE POULE SUR UN MUR"

Andy Arleo
Bernard Flament

Université de Nantes
I.U.T. de Saint-Nazaire

ABSTRACT

The findings presented here, obtained through a mingographic analysis of the French counting-out rhyme "Une poule sur un mur", are part of a broader study that is being carried out on the rhythm and melody of children's rhymes. We aim to determine the specific acoustic features of this rhyme by comparing two performances of nine-year old children : a narrative version in which the rhyme was told like a story and a counting-out version in which the rhyme was used to pick "It" ("le loup"). We hope, furthermore, to test the validity of former analyses which have usually been based on subjective musical transcriptions.

INTRODUCTION

Les formulettes enfantines et notamment les formulettes d'élimination que sont les comptines présentent des spécificités rythmiques et mélodiques par rapport à des énonciations courantes ou narratives. Certes, ces spécificités sont perceptibles à la simple audition; il nous est cependant apparu intéressant d'essayer de les appréhender de façon précise afin de les caractériser sur les plans rythmique et mélodique et pour évaluer les divergences de réalisation vis-à-vis d'une diction narrative. Des enregistrements ont été effectués à partir desquels des tracés mingographiques ont été obtenus, comportant les lignes relatives aux trois paramètres acoustiques -durée, intensité, F_o .

Avant de présenter certaines caractéristiques de la réalisation d'une comptine, en l'occurrence celle débutant par "Une poule sur un mur", nous nous proposons de faire brièvement état des recherches déjà menées dans le domaine des formulettes enfantines et de présenter la comptine qui a servi de support à ce travail.

TRAVAUX ANTERIEURS SUR LE RYTHME ET LA MELODIE DES FORMULETTES ENFANTINES

L'ethnomusicologue C. BRAILOIU (1956) a attiré l'attention sur un rythme enfantin "pré-établi" et répandu dans des cultures très diverses, un fait "d'autant plus remarquable qu'à l'intérieur des constructions rythmiques enfantines l'emplacement des accents est immuable, alors que les langues pratiquent des accentuations multiples" (p. 7). De son côté, l'anthropologue R. BURLING (1966), après avoir examiné des formulettes enfantines dans plusieurs langues structurellement différentes, pose l'hypothèse d'un schéma métrique universel qui serait constitué de 16 temps forts ("beats") répartis dans des vers de 4 temps forts chacun.

J. GUERON (1975 : 149), dans le cadre de la métrique générative, propose pour la comptine française, et en particulier "Une poule sur un mur", un schéma abstrait avec 4 syllabes accentuées par vers. B. de CORNULIER (1983 et 1985), partant lui aussi de la même comptine, propose une analyse des structures rythmiques à plusieurs niveaux, ce qui lui permet de constater une concordance entre les unités rythmique et musicale.

En ce qui concerne l'aspect mélodique des comptines, plusieurs chercheurs (GUERIFF, 1946-1948; ARLEO, 1982; CORNULIER, 1985) ont souligné l'importance des premier, deuxième et cinquième degrés de la gamme diatonique, surtout dans les séquences finales, qui sont caractérisées souvent par une montée fréquentielle sur la pénultième suivie d'une chute d'une quinte, e.g. do ré sol do.

Les résultats de ces travaux, si intéressants soient-ils, restent assez théoriques, ou du moins peu quantifiables, puisque ne faisant pas appel aux données de la phonétique. Nous avons voulu, quant à nous, partir des tracés mingographiques pour évaluer les caractéristiques rythmiques et mélodiques des comptines et plus particulièrement de celle qui a donné lieu à de multiples publications, en l'occurrence "Une poule sur un mur".

TEXTE DE LA COMPTINE

Une poule sur un mur
Qui picote du pain dur,
Picoti, picota,
Lève la queue et puis s'en va.

ENREGISTREMENTS ET TRACES MINGOGRAPHIQUES

Les tracés mingographiques ont été réalisés par le Laboratoire de l'Institut de Phonétique de STRASBOURG à partir d'un enregistrement d'une fille de 9 ans (locuteur n° 1) et d'un garçon de 9 ans également (locuteur n° 2), effectué dans le studio d'enregistrement de l'I.U.T. de SAINT-NAZAIRE en juin 1985. Pour la première réalisation, que nous qualifierons de "diction narrative", la consigne donnée était de raconter la comptine comme une histoire. Pour la deuxième réalisation, que nous appellerons "diction pouffée", nous avons demandé aux enfants de se mettre par terre et de "pouffer" comme dans la cour de récréation, chacun à son tour, et en faisant les gestes appropriés. Comme nous disposions de deux versions "pouffées" pour le locuteur n° 1, nous les désignerons respectivement par 1A et 1B dans notre étude.

CARACTERISATION SUR LE PLAN RYTHMIQUE

Nous ne rendons compte ici que de certains aspects qui nous semblent les plus marquants.

- durée phonématique moyenne et débit de la parole

L'analyse des tracés mingographiques nous fournit les données suivantes :

	type de diction	durée phonémat. moyenne (en cs)	débit (en nbre de phonèmes par sec.)
1er loc.	diction narrative	≃ 8,35	≃ 11,98
	diction pouffée 1A	≃ 10,87	≃ 9,20
	diction pouffée 1B	≃ 8,87	≃ 11,27
2ème loc.	diction narrative	≃ 8,36	≃ 11,96
	diction pouffée	≃ 10,85	≃ 9,22

Déjà à ce niveau d'analyse, des divergences sensibles apparaissent, affectant la durée des phonèmes et par voie de conséquence le débit de la parole -sans prise en compte des pauses. Les dictions pouffées présentent des valeurs de durée phonématiques plus importantes et par conséquent un débit phonématique plus lent. La diction narrative est au contraire caractérisée par une fluidité plus grande : le débit peut y être jusqu'à 30 % environ plus rapide que dans la diction pouffée correspondante.

Notons que le traitement des [ə] diverge assez sensiblement suivant les types de réalisations puisque dans les dictions pouffées, ces phonèmes sont très généralement prononcés alors qu'ils sont élidés d'une manière systématique (cas du 1er locuteur) dans la diction narrative. La réalisation ou l'élision des [ə] a une incidence directe sur la composition phonématique des éléments et groupes rythmiques ainsi que sur les groupements syllabiques; par là-même, les schémas rythmiques peuvent en être modifiés.

- réalisations pausales et découpage des énoncés en groupes rythmiques

Les pauses sectionnent l'énoncé et déterminent des groupes rythmiques de durée diverse. Dans les dictions narratives, l'importance numérique des pauses est plus grande que dans les dictions pouffées : on relève 4 pauses dans la diction narrative du 1er locuteur de même que dans celle du 2ème locuteur; dans les dictions pouffées, le nombre de pauses est inférieur : 2 seulement chez le 1er locuteur (1A); 2 également chez le 2ème locuteur; 3 sur les tracés correspondant à la version 1B. Les durées des pauses sont par ailleurs plus faibles dans les dictions pouffées :

	durées pausales dans la diction narrative (en cs)	durées pausales dans la diction pouffée (en cs)
1er loc.	7 - 56 - 9,5 - 35	1A : 5,5 1B : 6,5 - 17,5
2ème loc.	55 - 69 - 44 - 9,5	12,5 - 21,5

Les groupes rythmiques délimités par ces pauses apparaissent sur les plans de leur durée, de leur composition syllabique et phonématique bien différents suivant qu'ils se trouvent réalisés dans une diction narrative ou dans une diction pouffée; prenons par exemple les réalisations relatives au 1er locuteur : dans la diction narrative, les groupes rythmiques ont les durées suivantes (en cs) : 102,5 - 103,5 - 49 - 57,5 - 130; leur composition syllabique est respectivement de 5, 6, 3, 3, 7 syllabes; le nombre de phonèmes qui les constitue est varié : 12, 14, 6, 6, 15; dans la diction pouffée (1A), 2 groupes rythmiques seulement sont relevés, dont les durées sont de 304,5 cs et de 304 cs; leur composition syllabique (14 et 13 syllabes) et phonématique (29 et 27 phonèmes) est quasiment identique.

Ainsi, la *dynamique* des énoncés diverge profondément : aux groupes rythmiques de durée et de composition syllabique, phonématique très similaires dans les dictions pouffées (lesquelles apparaissent peu fractionnées) s'oppose, dans les dictions narratives, une grande diversité des masses phonémiques.

• l'accent : composantes paramétriques et proéminence

L'accent rythmique affecte les mêmes syllabes dans les deux dictions; toutefois, sa nature (= poids respectif des différents paramètres acoustiques) n'est pas identique; en outre, sa proéminence par rapport aux syllabes environnantes ne se fait pas nécessairement suivant le même écart.

Les syllabes accentuées sont les suivantes : poule (ou poul(e)), mur, picote (ou picot(e)), dur, picoti, picota, queue, va.

En ce qui concerne la valeur des différents paramètres dans les syllabes accentuées et l'importance de la proéminence par rapport aux syllabes environnantes, l'environnement étant réduit temporellement ici à la syllabe précédente et à la syllabe subséquente, voici quelles sont les données moyennes suivant le type de diction (nous n'avons reproduit que les données relatives au 1er locuteur : diction pouffée -1A- et diction narrative) :

Paramètres	syllabe concernée	type de diction	
		diction narrative	diction pouffée
durée (en cs)	syll. précéd.	≃ 16,44	≃ 20,63
	syll. accent.	≃ 25,06	≃ 33,88
	syll. subséq.	≃ 13,79	≃ 14,29
intensité (en dB)	syll. précéd.	58,5	55
	syll. accent.	58,75	55
	syll. subséq.	≃ 55,57	≃ 52,71
F_o (en Hz)	syll. précéd.	313,75	322,5
	syll. accent.	≃ 348,13	≃ 299,38
	syll. subséq.	≃ 301,67	275

. le paramètre de la durée joue un rôle beaucoup plus important dans la diction pouffée, à la fois par la valeur qui affecte les syllabes accentuées et par l'ampleur de la proéminence de celles-ci par rapport aux syllabes environnantes;

. pour les deux autres paramètres, on note des valeurs d'intensité et de F_o moindres dans la diction pouffée (la F_o de la syllabe accentuée y est même en moyenne plus faible que celle de la syllabe précédente).

N. BEAUCHEMIN (1970) considère que l'accent ("accent de puissance") est réalisé le plus souvent par la combinaison de deux ou des trois paramètres acoustiques; dans la diction narrative, il apparaît que c'est le cas; dans la diction pouffée, la proéminence est essentiellement effectuée sur le plan temporel. Comme on peut s'en rendre compte, le mode d'accentuation diverge sensiblement.

CARACTERISATION SUR LE PLAN MELODIQUE

. Soulignons tout d'abord la beaucoup plus grande régularité des vibrations laryngiennes accompagnant la réalisation des phonèmes dans les dictions pouffées. Celles-ci, il est vrai, s'apparentent davantage au chant. Dans les dictions narratives, la F_o au niveau des phonèmes est plus fluctuante et les valeurs relevées correspondent en fait à la période de relative stabilisation fréquentielle.

. Un autre élément marquant, selon nous, est la hauteur fréquentielle affectant assez souvent, dans les dictions pouffées, la 1ère syllabe des groupes grammaticaux; on observe en effet ce type de schéma mélodique (1er locuteur -les indications numériques sont données en Hz) :

	U -	ne	pou-	le }	sur	un	mur
1A	[325]	315	290	210	[340]	340	285
1B	[330]	315	260	230	[340]	330	280
	Qui	pi-	co-	te }	du	pain	dur / ...
1A	335	340	290	220	[320]	310	280
1B	[330]	325	290	220	[330]	305	275

ceci est d'autant plus frappant que les termes mis en valeur fréquentiellement sont des éléments peu enclins à être ainsi valorisés (prépositions, articles);

dans la diction narrative, la mélodie est toute différente; une montée a lieu très souvent à la fin des éléments rythmiques :

Un(e)	poul(e) }	sur	un	mur	/	
330	380	320	260	330		
Qui	pi - cot(e) }	du	pain	dur	/	
310	310 320	240	230	420		
Pi -	co - ti /	pi -	co -	ta	/	...
380	390 245	indét.	400	410		

. La séquence finale est intéressante à étudier sur le plan mélodique. Là se révèle une caractéristique essentielle de la formulette d'élimination, qui réside dans l'élévation de fréquence affectant l'avant-dernière syllabe :

	type de diction	et	puis	s'en	va
1er loc.	diction narrative	320	400	340	240
	diction pouffée 1A	280	330	[410]	285
	diction pouffée 1B	270	300	[420]	320
2ème loc.	diction narrative	225	210	240	240
	diction pouffée	260	340	[460]	290

Ce phénomène est très particulier et n'a pas d'équivalent, nous semble-t-il, dans une diction "courante", d'autant que la proéminence fréquentielle sur la pénultième est extrêmement accusée.

Le tableau suivant, établi pour les dictions pouffées à partir des fréquences indiquées précédemment, montre les intervalles musicaux les plus proches selon le système de tempérament égal (voir R. de CANDE, 1961, p. 139 et sq.). A, B, C et D représentent respectivement les F_o de et, puis, s'en, va (le rapport théorique de l'intervalle est indiqué entre parenthèses) :

	type de diction	B/A	C/A	C/D
ler loc.	diction pouffée 1A	1,18 3ce min. (1,19)	1,46 5te (1,5)	1,44 5te dim. (1,41)
	diction pouffée 1B	1,11 2de maj. (1,12)	1,56 5te aug. (1,59)	1,31 4te (1,33)
2ème loc.	diction pouffée	1,31 4te (1,33)	1,77 7me dim. (1,78)	1,59 5te aug. (1,59)

Ces résultats, en montrant la nature très relative des intervalles réalisés "en pouffant", permettent d'affiner les analyses antérieures. Ils suggèrent, en particulier, que la description habituelle de cette séquence finale (do ré sol do) est une idéalisation parfois assez éloignée d'une réalité sonore bien plus complexe. En fait, il faudrait considérer ces quatre dernières notes comme des zones fréquentielles en relation d'opposition qui sont interprétées par l'auditeur (ou le transcripteur) dans le cadre de son propre système musical.

CONCLUSION

Dans la présente étude, nous pensons avoir dégagé quelques traits saillants quant au rythme et à la mélodie de la comptine "Une poule sur un mur". Ceux-ci devraient pouvoir être confirmés ou infirmés lors de travaux approfondis portant sur d'autres formulettes enfantines.

REFERENCES BIBLIOGRAPHIQUES

ARLEO, A., 1982, Etude comparative des comptines françaises et anglaises, thèse de 3ème cycle, Université de Nantes.

BEAUCHEMIN, N., 1970, Recherches sur l'accent d'après des poèmes d'Alain Grandbois -Etude acoustique et statistique, Les Presses de l'Université Laval, Québec.

BRAILOIU, C., 1956, "Le rythme enfantin", art. in Problèmes d'ethnomusicologie, pp. 267-299, Minkoff Reprints, Genève.

BURLING, R., 1966, "The metrics of children's verse : a cross-linguistic study", American Anthropologist 68 : 1418-1441.

CANDE, R. de, 1961, Dictionnaire de Musique, nouvelle éd. revue et corr., Microcosme/Seuil, Paris.

CORNULIER, B. de, 1983, "Musique et vers : sur le rythme des comptines", Recherches Linguistiques 11 : 114-171, Université de Paris VIII.

CORNULIER, B. de, 1985, "De Gallina : l'air et les paroles d'une comptine", Le Français Moderne 53, Paris.

GUERIFF, F., 1946-1948, La poétique enfantine, les formulettes, texte inédit, Saint-Nazaire.

GUERON, J., 1975, "Langue et poésie : mètre et phonologie", in I. Change de forme biologies et prosodies, pp. 136-157, Coll. 10/18, Paris.

PERCEPTION D'UNE PHRASE RYTHMIQUEMENT PERTURBEE

Yukihiro Nishinuma et Danielle Duez

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

We carried out a perceptual study on the time specifications of a French sentence by manipulating syllable durations. Our preliminary results show that stressed or stressable syllables resist duration modifications better than unstressed ones. The duration range that is perceptually acceptable seems to be greater than the variation (standard deviation) observed on acoustic data.

1. INTRODUCTION

Dans une analyse acoustique de phrases simples, nous avons constaté que la structure rythmique du français est caractérisée par le principe d'alternance des unités brèves et longues [1], principe déjà démontré dans d'autres langues [2-6]. Cette alternance peut être modifiée, au niveau de l'énonciation, par la structure syllabique et surtout par l'accentuation du groupe rythmique [7]. La présente expérience a pour objectif d'évaluer l'effet perceptif de nos constatations acoustiques.

2. PROTOCOLE

A l'aide d'un vocoder à phase [8], nous avons généré les stimuli en manipulant la durée des segments de la phrase originale : "Mes beaux bébés dorment tranquillement dans la pièce d'à côté" (voix de femme, 2,9 secondes). La manipulation de durée est limitée pour le moment sur les quatre premières syllabes qui mesurent respectivement : 163, 182, 165 et 286 ms. Le taux de variation est de + ou - 25 % et + ou - 50 %, compte tenu de la largeur du seuil différentiel de durée des stimuli linguistiques longs [9, 10].
Cinq séries de stimuli disposés aléatoirement enregistrés sur une bande magnétique ont été présentés à dix sujets. Leur tâche consistait à juger sous le choix forcé, si chaque phrase entendue était naturelle ou non.

3. ANALYSE

Le dépouillement des résultats du test perceptuel nous montre des tendances intéressantes.

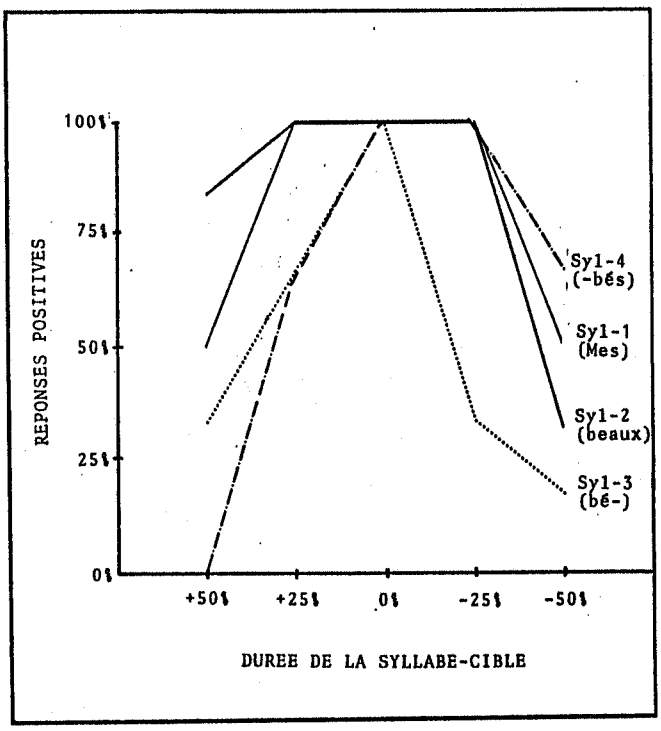


Figure 1 : Réponses positives en fonction de la durée de la syllabe-cible.

Comme c'est souvent le cas dans ce genre d'expérience, les réponses positives se concentrent sur les trois stimuli le moins affectés, c'est-à-dire ceux du milieu (+ ou - 25 % et 0 %). Cette remarque nous indique que l'alternance binaire est d'ordre purement physiologique, car la modification de + ou - 25 % détruit ou exagère la configuration temporelle alternée mais sonne naturel.
Les stimuli des deux extrémités reçoivent des scores médiocres, mais les valeurs extrêmes étant modérées dans cette étude (approximativement 3 écarts type de la durée observée), les réponses ne permettent pas de tracer une courbe en cloche théoriquement attendue.
La première syllabe de la phrase "mes" peut être accentuée dans certain style, de même la deuxième syllabe "beaux" peut recevoir un accent d'insistance. De ce fait, l'allongement sur ces syllabes

est celui qui est le mieux accepté; les réponses positives absolues sont enregistrées sur les trois stimuli centraux.
La troisième syllabe "bé" est intrinsèquement inaccentuable, si bien que la prolongation de sa durée syllabique donne les résultats les plus défavorables. La même dimension de raccourcissement produit également une réaction négative, parce que la syllabe est réalisée initialement en durée minimale, ce qui ne permet pas de la comprimer davantage. En d'autres termes, dès qu'on sort de la durée d'origine, les réponses baissent de façon significative. La syllabe inaccentuée semblerait donc constituer une durée de référence dans la structure rythmique de la phrase.
La quatrième et dernière syllabe du groupe nominal de la phrase est déjà réalisée avec un allongement dans la phrase originale. Un allongement supplémentaire semble être mal accepté sur cette syllabe. Par contre la syllabe supporte mieux le raccourcissement, même à 50 %.
De ce qui précède, on peut dire que la marge de tolérance temporelle ou critère interne, selon les termes de Nooteboom [11], est donc large pour les syllabes accentuées et accentuables; par contre celle-ci est réduite de moitié pour les syllabes inaccentuables. Cette marge de tolérance est nettement supérieure à l'écart type de la durée syllabique observée (environ 15 % de la durée moyenne). A partir de là, il semble y avoir un critère différent pour le codage et décodage du rythme.

4. CONCLUSION

Nous mettons en chantier une série d'expériences perceptives sur l'organisation temporelle de phrases. Nos premiers résultats montrent que les syllabes accentuables supportent mieux la perturbation temporelle que les syllabes inaccentuables. De plus la portée de variation perceptuelle dépasse largement la variation acoustique observée.

BIBLIOGRAPHIE

-[1] Duez, D. & Nishinuma, Y., "Some evidence on rhythmic patterns of spoken French", **PERILUS**, IV, pp. 30-40, 1985.
-[2] Allen, G., "Speech rhythm : its relation to performance universals and articulatory timing", **Journal of Phonetics**, 3, 75-86, 1975.
-[3] Bruce, G., "On rhythmic alternation", **Working papers of Lund**, 25, pp. 75-86, 1983.
-[4] Bruce, G., "On the phonetics of rhythm : evidence from Swedish", **Journal of the Acoustical Society of America**, 75 (suppl. 1), p.s40, 1984.
-[5] Lindblöm, B., Lyberg, B., Holmgren, K., **Durational Patterns of Swedish Phonology : do they Reflect Short Term Motor Memory Processes ?**, Indiana University Linguistics Club, Bloomington, 1981.
-[6] Woodrow, H., "Time perception", **Handbook of Experimental Psychology**, (ed. S.S. Stevens), 1224-1236, 1951.
-[7] Duez, D. & Nishinuma, Y., "Du principe d'alternance des durées du rythme en français", **T.I.P.A.** (à paraître).
-[8] Espesser, R. & Nishinuma, Y., **Traitement du signal sous le système UNIX : Description des commandes**, Institut de Phonétique, Université de Provence, 1984.
-[9] Rossi, M., "Le seuil différentiel de durée", **Papers in Linguistics and Phonetics to the Memory of Pierre Delattre**, (éd. A. Valdman), Mouton, La Haye, pp. 435-450, 1972.
[10] Lehiste, I., "The perception of duration within sequences of four intervals", **Journal of Phonetics**, 7, pp. 313-316, 1979.
[11] Nooteboom, S.G., "The perceptual reality of some prosodic durations", **Journal of Phonetics**, 1 (1), pp. 25-46, 1973.
[12] Rossi, M., "Le français, langue sans accent ?", **Studia Phonetica**, 15, pp. 13-51, 1979.

STRUCTURE PROSODIQUE ET STRUCTURE RYTHMIQUE POUR LA SYNTHESE

Philippe Martin

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

It has been known for some time that, in French, the prosodic structure does not always correspond to the syntactic structure of the sentence, and that rhythmic factors must be considered as well as syntax in the generation of sentence prosody.
It will be shown here that acceptable prosodic structures can be generated independently from the syntactic structure and that the best structure - in terms of eurhythmy - can be selected by using a **disrhythmy index** distributing the number of syllables evenly at all levels of the structure.

INTRODUCTION

Dans la synthèse à partir du texte, la génération de la prosodie est, dans beaucoup de systèmes, basée sur la congruence supposée entre intonation et syntaxe. Dans ce cas, la génération de contours prosodiques adéquats doit être précédée d'une analyse syntaxique, même sommaire, de l'énoncé à synthétiser.
Ainsi, dans Emerard [1], les contours prosodiques sont dérivés de marqueurs placés manuellement de manière à correspondre à certaines catégories syntaxiques, et dans Martin [2], la structure prosodique est déterminée d'après la structure syntaxique de la phrase.
Par contre, dans Choppy et al. [3] la prosodie est générée à partir de critères non syntaxiques tels que la ponctuation, le nombre de syllabes de mots, etc.
Les procédés de synthèse prosodique ne demandant pas d'analyse syntaxique sont évidemment plus simples à mettre en œuvre. En effet, on évite ainsi la recherche de la hiérarchie syntaxique (parenthésage) qui, dans le cas général, requiert des analyseurs complexes fonctionnant à partir de bases de données considérables.
Pour s'affranchir de cette contrainte, d'autres approches pour la génération de la prosodie peuvent être envisagées, en se basant sur des propriétés non syntaxiques des phrases à synthétiser.

EURYTHMIE

Dans son étude sur l'accent de phrase en français, Dell [4] remarque que certaines séquences accentuables sont plus acceptables que d'autres en vertu de leur "eurythmicité", plutôt que de leur adéquation à la structure syntaxique.
Ainsi, si 1, 2, ..., n représentent des degrés décroissants d'accent, la séquence a) est plus accentuable que b) dans

1. Son pantalon rouge (plein (d'grains (d'tabac)))

a) 2 1
b) 2 1

alors que dans 2., qui présente la même structure syntaxique que 1,

2. Son pull (plein (d'grains (d'tabac)))

a) 2 1
b) 2 1

b) est plus acceptable que a) bien que la séquence accentuelle contredise la syntaxe (l'unité **plein** reçoit une proéminence accentuelle plus forte que **pull**).
Ceci s'expliquerait par le nombre de syllabes de chaque groupe terminé par un accent 1 et 2.

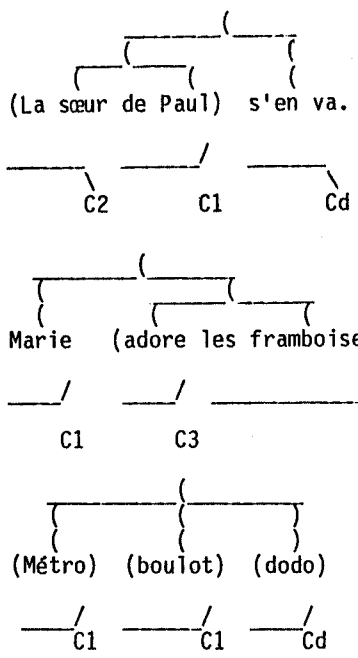
1 a 5 et 6
b 6 et 3

2 a 2 et 4
b 3 et 3

Les phrases 1a et 2b apparaissent donc comme plus équilibrées du point de vue rythmique.
De même, Wenk et Wioland [5], dans leur étude sur l'organisation rythmique en français, remarquent une forte tendance pour les locuteurs à réaliser des groupes rythmiques (i.e. des groupes accentuels) présentant un même nombre de syllabes (de 2 ou 3 syllabes en moyenne).

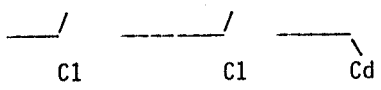
STRUCTURE PROSODIQUE ET STRUCTURE RYTHMIQUE

Il a été souvent admis que la structure prosodique pouvait être dérivée d'une certaine façon de la structure syntaxique de la phrase (cf. par exemple Martin [2]). Ainsi, les structures prosodiques des exemples 3, 4 et 5 correspondent à la syntaxe.

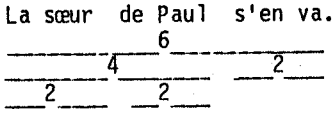


L'exemple 6 par contre présente une organisation syntaxique particulière (non planaire), à laquelle ne peut être associée qu'une structure planaire :

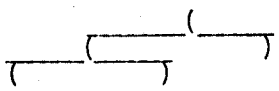
6. Marie, (le matin), déjeune.



L'organisation hiérarchique définie par la structure prosodique peut être considérée également comme une division rythmique de la phrase. En effet, chaque mot prosodique correspond à un certain nombre de syllabes dont la dernière est accentuée, et associée à un contour mélodique spécifique indiquant sa place dans la structure.
Dans 5., par exemple, la phrase est divisée en un premier niveau en deux unités de 4 et 2 syllabes, et en un deuxième niveau en 2 unités de 2 syllabes chacune.



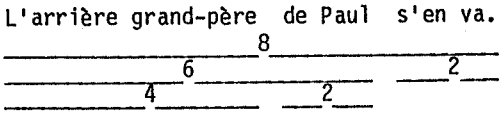
La structure prosodique correspondante est :



et présente sur les syllabes accentuées successives des contours descendant, montant et descendant bas.
La même structure prosodique, associée à la phrase

L'arrière grand-père de Paul s'en va.

présentant la même structure syntaxique que l'exemple précédent, détermine une division rythmique déséquilibrée en terme du nombre de syllabes :



Les 8 syllabes de la phrase sont divisées en un premier niveau en 2 unités prosodiques de 6 et 2 syllabes, dont la première est à son tour divisée en un deuxième niveau en deux mots prosodiques de 4 et 2 syllabes.

INDEX DE DISRYTHMIE

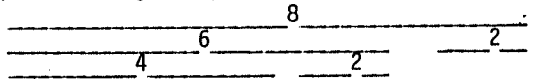
Plutôt que d'utiliser la structure syntaxique de la phrase comme critère de sélection d'une structure prosodique appropriée pour des applications de synthèse, on pourrait donc se servir du caractère plus ou moins eurythmique des structures retenues. La tendance à la réalisation d'unités présentant le même nombre de syllabes, observée par exemple par Wenk et Wiolland [5], ainsi que le concept d'eurythmicité de Dell [4] peuvent être aisément généralisés à l'ensemble de la structure prosodique.
Ainsi les trois structures prosodiques que l'on peut associer à la phrase

L'arrière grand-père de Paul s'en va.

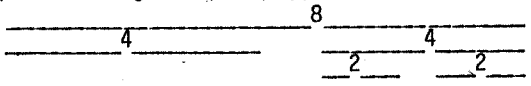
déterminent des divisions rythmiquement différentes :

(les structures prosodiques sont indiquées par le parenthésage)

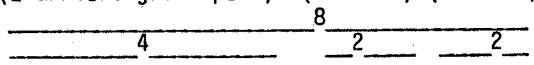
1. ((L'arrière grand-père) (de Paul)) (s'en va)



2. (L'arrière grand-père) ((de Paul) (s'en va))

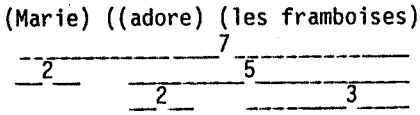


3. (L'arrière grand-père) (de Paul) (s'en va)

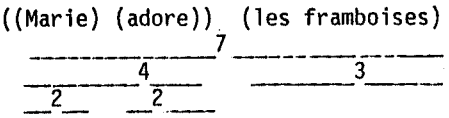


Le degré d'eurythmie correspondant à chacune de ces structures prosodiques est donc différent.
De manière à évaluer le degré de déséquilibre syllabique des différents groupes prosodiques, on peut imaginer un **index de disrythmie** incorporant le degré d'eurythmie existant à chaque niveau de la hiérarchie prosodique, et en le pondérant selon la profondeur des groupes considérés dans la structure. Cet index donnerait donc plus d'importance au déséquilibre rythmique des unités de niveau supérieur.
Si a, b, ..., n représentent le nombre de syllabes des mots prosodiques A, B, ..., N d'une phrase donnée, l'**index de disrythmie** est défini par la somme des valeurs absolues des différences du nombre des syllabes entre les unités prosodiques successives à chaque niveau de la structure, divi-

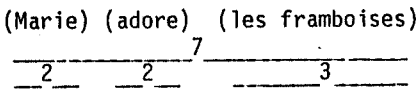
sées par le numéro du niveau considéré.
Ainsi dans



l'index de disrythmie se calcule comme suit :
- disrythmie créée au premier niveau de la structure : [2-5]=3
- disrythmie créée au deuxième niveau : [2-1]/2=0,5 (différence du nombre de syllabes, en valeur absolue, divisée par le numéro d'ordre du niveau).
Disrythmie totale : 3 (1er niveau) + 0,5 (2e niveau) = 3,5.
Les deux autres structures prosodiques que l'on peut associer à cet exemple donnent :

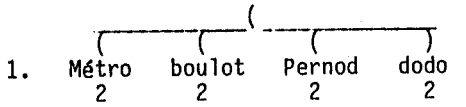


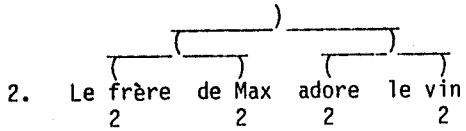
Index de disrythmie : [4-3] + [2-2]/2 = 1

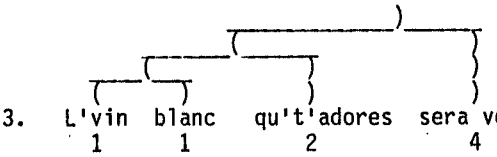


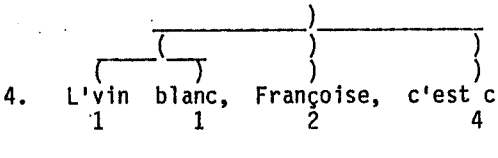
Index de disrythmie : [2-2] + [2-3] = 1

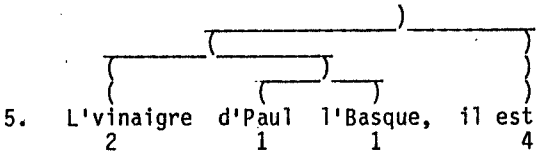
Les deuxièmes et troisièmes structures présentent donc des index de disrythmie égaux à 1, et une eurythmie meilleure que la première structure, pourtant congruente à la syntaxe.
L'index de disrythmie peut être calculé de la même manière pour des structures plus complexes à 4, 5, ..., n mots prosodiques (cf. Martin [6]). Pour chacune des structures, il existe une classe de divisions parfaitement eurythmiques, dont l'index vaut zéro. Il est donc possible d'associer aux phrases qui remplissent ces conditions des structures prosodiques qui soient à la fois congruentes à la syntaxe (les deux structures syntaxiques et prosodiques sont homomorphes) et parfaitement eurythmiques.
Des exemples à 4 mots prosodiques (4 syllabes effectivement accentuées), avec lesquelles 11 structures prosodiques peuvent être associées (en excluant les divisions thème-propos du type (Il est venu, Jean) sont données ci-dessous :

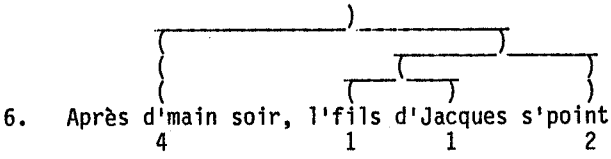


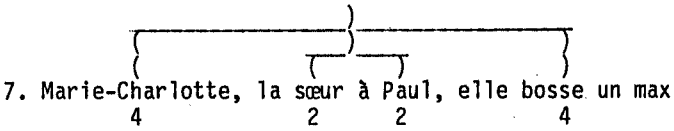












etc.

CONCLUSION

L'emploi d'un critère d'eurythmie plutôt que la congruence syntaxique pour le choix d'une structure prosodique permet de s'affranchir de l'analyse syntaxique de la phrase. L'index de disrythmie, facile à calculer à partir du nombre de syllabes de chaque unité prosodique (c'est-à-dire de chaque unité accentuée), permet de dériver une structure prosodique également satisfaisante même si elle est en contradiction avec la syntaxe dans le cas général.

REFERENCES

[1] Emerard, F, **Synthèse par diphones et traitement de la prosodie,** thèse de 3e cycle, Université de Grenoble III, 1977.
[2] Martin, Ph., "Pour une théorie de l'intonation", in **L'intonation, de l'acoustique à la sémantique,** Rossi ed., 234-271, 1981.
[3] Choppy, C., Liénard, J.-S. & Teil, D., "Un algorithme de prosodie automatique sans analyse syntaxique", **Actes des 6èmes J.E.P.**, Toulouse, 387-395, 1975.
[4] Dell, F., "L'accentuation dans les phrases en français", in **Formes sonores du langage,** Dell, Hirst et Vergnaud eds., Hermann, Paris, 65-122, 1984.
[5] Wenk, B & Wiolland, F., "Is French really syllable timed", **Journal of Phonetics,** (10) 193-216, 1982.
[6] Martin, Ph., "Prosodic and rhythmic structures in French", **Linguistics,** 1986 (à paraître).

UNITES TONALES ET UNITES RYTHMIQUES DANS LA REPRESENTATION DE L'INTONATION

Daniel Hirst et Albert Di Cristo

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

We present a model of a core system for French and English intonation, based on three non-terminal categories : syllable, tonal unit and intonation unit. These categories define a phonological structure which provides a link between two sequences of segments : phonematic segments and tonal segments. We show that according to this model the basic difference between English and French intonation is the choice of a single parameter. This model was developped for the analysis and synthesis of f0 contours. An extension of the model taking into account recent work on metrical grids allows us to account for rhythmic and tonal patterns on the basis of a single phonological representation.

INTRODUCTION

La principale critique que l'on peut formuler à l'encontre de la plupart des travaux sur l'intonation est que ces derniers tentent d'établir un lien trop direct entre l'acoustique et la sémantique (ou la syntaxe) sans se préoccuper de la nature de la **représentation phonologique** de l'intonation qui constitue cependant un niveau d'analyse intermédiaire autonome et indispensable.
Nous avons proposé antérieurement [4, 6] de représenter l'intonation au niveau formel par une structure métrique associée à des segments tonals. Nous ne parlerons pas dans cet exposé de la façon dont cette forme interagit avec la syntaxe, la sémantique et la pragmatique qui relèvent du plan du contenu. Nous nous attacherons plutôt à décrire l'organisation des unités du plan de l'expression. Les rapports entre la forme et la substance qui constituent les domaines d'analyse propres à ce plan peuvent être schématisés de la façon suivante :

FORME	SUBSTANCE
segments tonals	f0
	intensité
structure métrique	durée segmentale

Il va de soi qu'au niveau de la substance, tout énoncé, quelle que soit la langue considérée, doit être pourvu d'une courbe de f0 et d'intensité ainsi que de durées segmentales. Au niveau formel, nous faisons l'hypothèse que nous pouvons également utiliser un même type de représentation pour toutes les langues. La distinction entre langues à tons, langues à accent libre et langues à accent fixe reposerait alors sur une différence entre ce qui doit être spécifié dans le lexique de la langue et ce qui est prévisible par des règles phonologiques. En chinois, par exemple, le lexique doit spécifier, pour chaque mot, sa structure métrique et ses segments tonals. En anglais, par contre, au moins certains mots devront être représentés avec leurs structures métriques, mais il n'est pas nécessaire de spécifier leurs segments tonals. En français, enfin, les mots seront représentés sans structure métrique ni segments tonals. Comme il n'est pas possible de prononcer un énoncé sans structure métrique ni segments tonals, les règles phonologiques spécifiques pour chaque langue transformeront les représentations abstraites sous-jacentes en représentations **prononçables.**

LE MODELE HIERARCHIQUE

Conformément à la plupart des descriptions phonologiques récentes des langues à tons, nous utilisons essentiellement trois segments tonals : H (ton haut); B (ton bas); A (ton abaissé par le phénomène dit de **faille tonale** (downstep)). Le ton A, dans notre modèle, est toujours le résultat de l'application d'une règle d'abaissement du type :

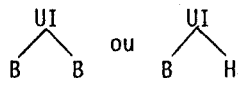
Downstep $H \longrightarrow A \;/\; H\ B \;\_\_$
$B \longrightarrow 0 \;/\; \_\_ \;A$

règle que l'on rencontre très fréquemment dans des descriptions tonologiques des langues africaines (cf. [7] et références citées).
Nous avons montré qu'une représentation de l'intonation comme une séquence de segments tonals peut facilement être mise en relation avec la représentation phonétique d'une courbe de f0 au moyen d'une fonction d'interpolation de type **spline quadratique** qui permet d'obtenir une très bonne approximation d'une courbe réelle ([3], cf. aussi [8] pour une utilisation récente de notre modèle).
Dans des applications de ce modèle à l'intonation de l'anglais [4] et du français [6] nous avons utilisé

une structure métrique constituée par un arbre phonologique comportant deux catégories terminales : les segments phonématiques et les segments tonals, ainsi que trois catégories non-terminales : S (syllabe), UT (unité tonale) et UI (unité intonative). Chaque unité se voit affectée de tons au moyen d'un **gabarit prosodique** spécifique à la langue considérée. Nous définissons ainsi une représentation à trois dimensions car les UTs et les UIs se trouvent à l'intersection de deux plans : le plan phonématique et le plan tonal. Elles constituent de ce fait le domaine de synchronisation des segments correspondants. Comme une représentation de ce type n'est pas **prononçable** (les segments tonals ne se situant pas tous sur la même ligne) nous devons faire appel ensuite à une règle universelle de **linéarisation** qui copie les tons de l'unité intonative sur l'unité tonale appropriée. Nous utilisons enfin une règle de **simplification tonale** (correspondant au "Obligatory contour principle" (OCP) de Leben [9].
La différence essentielle entre l'intonation de l'anglais et celle du français apparaît alors comme une différence dans le gabarit prosodique indiquant les tons qui doivent être insérés dans une unité tonale. Il existe également une différence dans le contexte d'application de la règle de Faille Tonale mais cette différence n'est pas pertinente pour les exemples que nous présentons. (Cf. [6]).

Gabarits prosodiques :

- unités intonatives (anglais et français) :



- unités tonales :

anglais français

Prenons un exemple de dérivation d'une réalisation possible de la phrase suivante :

Ces exemples sont illustrés par les figures 1 et 2.

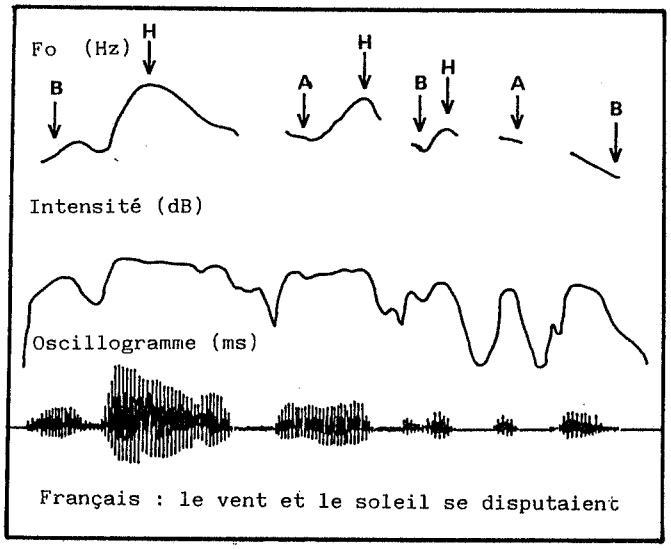


Figure 1

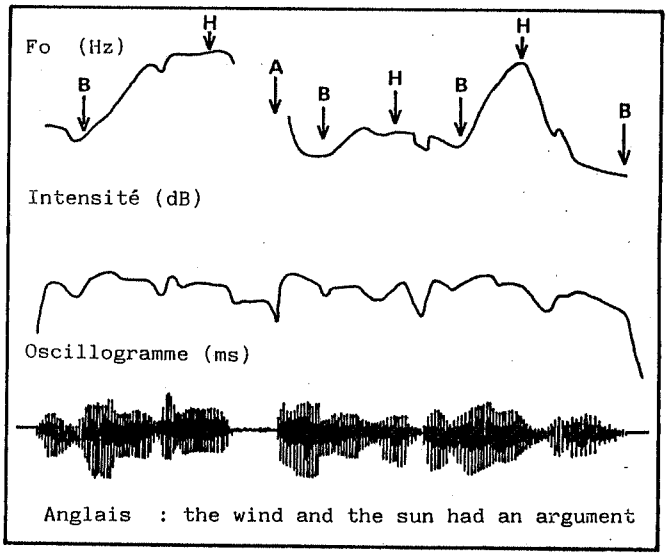


Figure 2

	Le vent et le soleil se disputaient
Affectation UIs	(Le vent et le soleil) (se disputaient)
Affectation UTs	((Le vent)(et le soleil))((se dis-)(putaient))
T-segments UI	(B H)(B B)
T-segments UT	(B H)(B H) (B H)(B H)
Faille Tonale	(B H)(A) (B H)(A)
Linéarisation	((BB H)(A H))((BB H)(A B))
Simplification (OCP)	((B H)(A H))((B H)(A B))

et de la phrase équivalente en anglais :

	The wind and the sun had an argument
Affectation UIs	(The wind and the sun) (had an argument)
Affectation UTs	(The (wind and the)(sun))(had an (argument))
T-segments UI	(B H)(B B)
T-segments UT	(H B)(H B) (H B)
Faille Tonale	(H)(A B) (H B)
Linéarisation	((B H)(ABH))((B H B B))
Simplification (OCP)	((B H)(ABH))((B H B))

Ce formalisme permet de rendre compte d'un système noyau de l'intonation de ces deux langues au moyen d'un matériel formel extrêmement simple et bien motivé par ailleurs.

GRILLES METRIQUES

Le modèle hiérarchique avait été élaboré pour rendre compte essentiellement des configurations de fréquence fondamentale. En ce qui concerne les phénomènes rythmiques la représentation que nous avons proposée présente un défaut majeur par rapport à d'autres formalisations faisant appel à la notion de **grilles métriques** [10, 1, 11]. En effet nous sommes obligés de stipuler que la proéminence rythmique en français porte sur la **dernière** syllabe de l'unité tonale alors qu'en anglais elle porte sur la **première** syllabe de cette unité car cette proéminence n'apparaît pas dans la représentation elle-même. Dans ue représentation utilisant une grille métrique, par contre, cette proéminence serait directement représentée de la manière suivante :

```
                +           +      3
      +         +     +     +      2
. + + + + + + + + + +      1
Le vent et le soleil se disputaient.

                +           +      3
      +         +           +      2
  + + +   + + + + + + +     1
The wind and the sun had an argument.
```

où le nombre de + au-dessus de chaque syllabe indique son niveau relatif de proéminence.

UNE SYNTHESE

Cependant, nous avons montré ailleurs [5] qu'une représentation sous la forme d'une grille métrique n'est pas compatible avec une représentation phonétique de l'intonation au moyen de cibles et de transitions comme dans le modèle que nous préconisons. Cette représentation impliquerait une modification considérable des fondements mêmes des travaux sur la tonologie.
Or, le formalisme très récent proposé par Halle et Vergnaud [2] nous permet d'éviter une telle remise en cause puisque ces auteurs font appel à une structure de grille incorporant des constituants métriques. Selon cette formalisation, chaque constituant métrique contient un élément qui est sa **tête** et qui est représenté par un + sur la ligne supérieure. La tête d'un constituant à un niveau donné est spécifiée pour chaque langue comme étant soit le premier soit le dernier élément du constituant. L'intérêt de cette représentation est de permettre aux règles phonologiques de faire référence aussi bien à la formation de constituants (provoquant la promotion d'un élément comme tête) qu'à l'affectation d'une proéminence (provoquant la formation d'un constituant).
Nos exemples auront maintenant la représentation rythmique :

```
 . . .  . . +  . . . +      3
(. + .  . . +) (. + . +)    2
(+ +) (+  + + +) (+ +)(+ +) 1
Le vent et le soleil se disputaient.

 . . .  . + .  . + . .      3
(. + .  . +) (. . + . .)    2
 + (+   +  +)(+) + + (+ + +) 1
The wind and the sun had an argument.
```

Rythme et mélodie apparaissent alors comme deux projections d'une même représentation car les unités rythmiques définies au niveau 1 correspondent aux unités tonales et les unités définies au niveau 2 sont les unités intonatives. A partir de cette représentation métrique, les unités seront donc projetées sur le plan tonal et la dérivation tonale sera identique à celle présentée dans la première partie de ce travail.

REFERENCES

[1] Dell, F., "L'accentuation dans les phrases en français", in Dell, Hirst & Vergnaud (eds.) 1984, **La Forme Sonore du Langage**, Hermann, Paris, 65-122.

[2] Halle, M. & Vergnaud, J.R., "Stress and the cycle", Communication au **Colloque de Phonologie Plurilinéaire**, Lyon, juin 1985.

[3] Hirst, D.J., "Un modèle de production de l'intonation", **T.I.P.A.** 7, 1980, 297-315.

[4] Hirst, D.J., "Structures and categories in prosodic representations", in Cutler & Ladd (eds.), **Prosody : Models and Measurements**, Springer, Berlin, 1983.

[5] Hirst, D.J. (sous presse), "Tonal units as constituents of prosodic structure : the evidence form French and English intonation", in H. van der Hulst & N. Smith (eds.), **Autosegmental Studies on Pitch Accent**, Foris, Dordrecht.

[6] Hirst, D.J. & Di Cristo, A., "French intonation : a parametric approach", **Die Neueren Sprachen**, 83 (5), 1984, 554-569.

[7] Hyman, L., "Word domains and downstep in Bamileke-Dschang", **Phonology Yearbook**, 2, 1985, 45-82.

[8] Ladd, D.R.; Silverman, K.; Tolkmit, F.; Bergmann, G. & Scherer, K., "Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect", **J.A.S.A.**, 78 (2), 1985, 435-444.

[9] Leben, W.R., **Suprasegmental phonology**, Thèse Ph.D. distribuée par IULC, 1973.

[10] Prince, A., "Relating to the grid", **Linguistic Inquiry**, 14 (1), 1983, 19-100.

[11] Selkirk, E., **Phonology and Syntax : the Relation between Sound and Structure**, MIT Press, Cambridge, Mass., 1984.

INFLUENCE DE LA VITESSE D'ARTICULATION SUR LA DUREE DES SYLLABES ET DES GROUPES CONSONANTIQUES EN FRANCAIS

Danielle Duez et Yukihiro Nishinuma

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

Analysis of the variability of the duration of stressed and unstressed syllables shows that, in French, the duration of stressed syllables is more stable. In contrast to this phenomenon, an analysis of syllabic components in English and Russian reveals that in those languages consonants show a greater resistance to rate variations. This experiment constitutes only a pilot study; more data are needed in order to better understand the process underlying rate variability.

1. INTRODUCTION

Cette étude a pour objectif d'analyser l'influence des variations de la vitesse d'articulation sur la durée des syllabes situées en position accentuée et inaccentuée, en fin de groupe intonatif sujet et en fin de phrase. Ces syllabes ont une structure complexe CCV et l'incidence de la vitesse d'articulation sur la durée des groupes consonantiques et de leurs éléments constitutifs est également examinée.

2. PROCEDURE D'ANALYSE

a. Corpus

Les syllabes cibles /d r a/, /s m a/, /s k a/ sont situées à la finale et à la médiane des mots trisyllabiques suivants : "recadra", "les drapeaux", "l alaska", "la scala", "fantasma", "la smala". Ces mots sont respectivement situés en fin de syntagme sujet et en fin de syntagme verbal dans les phrases "l'expression ... fut relue par chacun" et "chacun répéta l'expression ...".

b. Tâche et locuteurs

Les phrases ont été enregistrées sur Nagra, en chambre sourde par cinq locuteurs ne présentant aucun accent régional. Chacune des deux séries de phrases a été lue à vitesse normale, ensuite à vitesse lente, et enfin à vitesse rapide, sans marquer de pause silencieuse ni d'accent externe. Cette tâche a été répétée cinq fois.

c. Analyse

La durée de chacune des phrases, de chacune des syllabes cibles, et de leurs éléments constitutifs a été mesurée à partir de tracés oscillographiques, de la courbe d'intensité, et de la courbe de fréquence fondamentale. La vitesse d'articulation moyenne a été calculée (vitesse normale, vitesse lente, vitesse rapide) ainsi que la vitesse d'articulation de chacun des sujets.

3. RESULTATS

La comparaison des moyennes obtenues pour la vitesse d'articulation normale, rapide, et lente révèle des variations peu élevées : l'écart entre les différentes vitesses moyennes est de l'ordre d'une syllabe (respectivement 5,5 syll/sec, 6,4 syll/sec et 4,6 syll/sec). La marge de variation de la vitesse d'articulation semble assez limitée : cette variable étant soumise à des contraintes d'ordre articulatoire, rythmique et perceptif.
La comparaison de la durée des syllabes accentuées et inaccentuées en fonction de la vitesse montre une plus grande stabilité de la durée des syllabes inaccentuées : en fin de syntagme sujet, les syllabes accentuées sont soumises à une compression qui est de l'ordre de 17 % et à une expansion de 20 %, les pourcentages correspondants sont de 9 % et 15 % pour les syllabes inaccentuées. En fin de phrase, ces différences tendent à être annulées : il est probable que la décélération constatée en fin de phrase vienne perturber l'influence des variations de la vitesse d'articulation (voir figure 1).
Cette plus grande stabilité de la durée des syllabes inaccentuées est une caractéristique du français. En anglais, ce sont les syllabes accentuées qui sont le moins sensibles aux variations de la vitesse d'articulation. Peterson et Lehiste [1] constatent à ce propos : "the duration of the syllable with main stress is elongated less than the duration of the unstressed syllables" et des résultats similaires sont obtenus par Port [2].
La spécificité de la réalité acoustique de la syllabe inaccentuée en anglais et en français peut expliquer cette différence. Ces résultats suggèrent, par ailleurs, le fait que la perception des temps forts, c'est-à-dire des syllabes accentuées, se fait par référence aux temps faibles (syllabes inaccen-

tuées). Ces résultats sont en accord avec l'hypothèse de Fraisse [3] concernant la perception du rythme.
L'analyse des variations de durée des consonnes et des voyelles confirme les résultats obtenus pour le russe par Kozhevnikov et Chistovich [4] : la durée des consonnes résiste mieux aux variations de la vitesse que les voyelles. L'examen de la durée respective de chacune des consonnes des groupes ne permet pas de tirer de conclusions concernant les variations de durée de la consonne en fonction de sa place dans le groupe, on peut noter cependant que dans le cas des groupes /s+k/ et /s+m/, la durée de /k/ est plus stable que celle de /m/, sans doute cette différence est-elle due au caractère sibilant de la nasale (voir figure 2).

4. CONCLUSIONS

Ces premiers résultats révèlent certaines caractéristiques des variations de durée en relation avec la vitesse d'articulation (plus grande stabilité de la syllabe inaccentuée, de la consonne). Un approfondissement de cette étude devrait permettre de mieux comprendre le phénomène de variabilité de la vitesse d'articulation, ce phénomène longtemps sous-estimé joue un rôle important dans la perception de la parole.

BIBLIOGRAPHIE

[1] Peterson, G & Lehiste, I., "Duration of syllable nuclei", **J.A.S.A.**, 32, 6, pp. 693-703, 1960.
[2] Port, R., "Linguistic timing factors in combination", **J.A.S.A.**, 69, 1, pp. 262-273, 1978.
[3] Fraisse, P., **La psychologie du rythme**, P.U.F., Paris, 1974.
[4] Kozhevnikov, A & Chistovich, L., "Speech : articulation and perception", **U.S. Dept. of Commerce. Joint Publication Research Service**, No. 30 543, 1965.

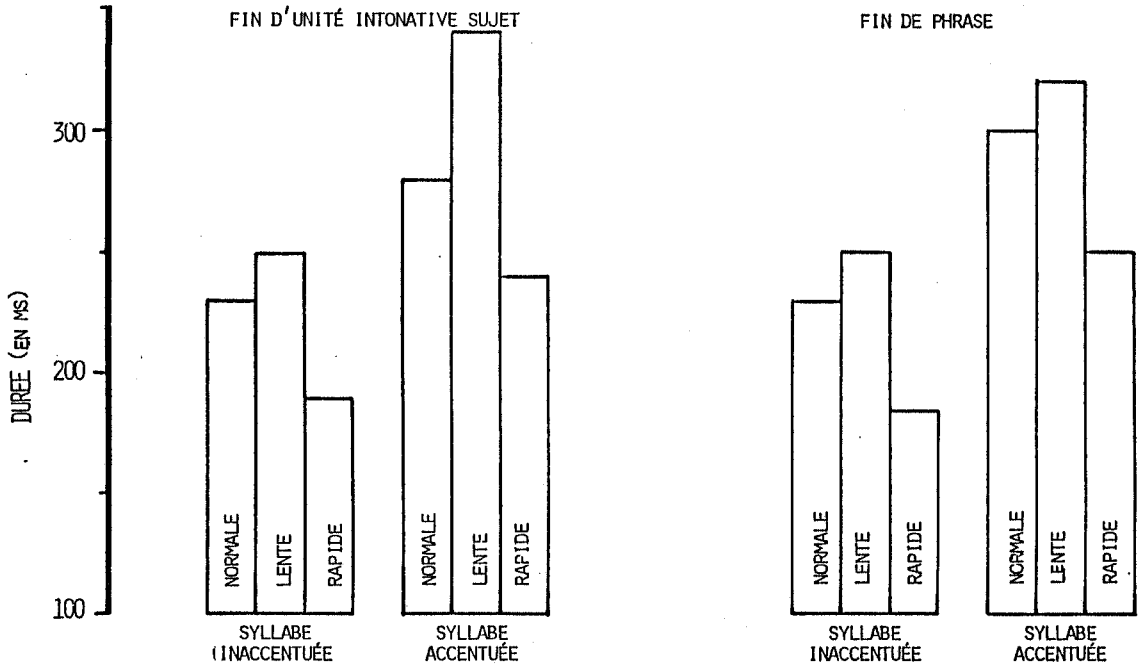


Figure 1 : Durée syllabique en fonction de la vitesse d'articulation.

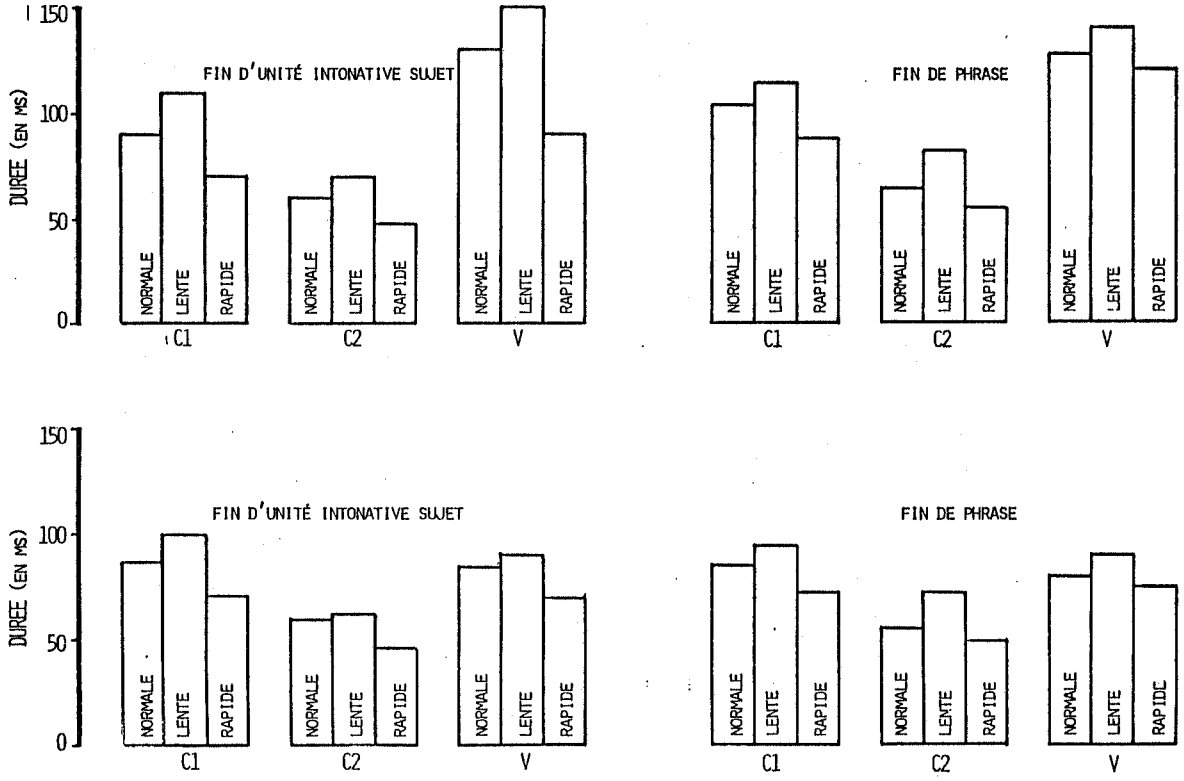


Figure 2 : Durée des éléments syllabiques en fonction de la vitesse d'articulation.

AUDITION

Président

Christel SORIN

C.N.E.T. de Lannion

PRETRAITEMENT DE LA PAROLE AVEC UN DISPOSITIF SIMULANT LA SUPRESSION LATERALE

V.C. Dang, R. Carré

Institut de la Communication Parlée, U.A.CNRS 368. 46, Av. Félix-Viallet, 38031, GRENOBLE CEDEX

RESUME

Dans ce travail, on a testé le phénomène d'inhibition latérale du système auditif en utilisant une FFT suivie par un système d'inhibition latérale et de lissage. Les résultats mettent en évidence une amplification des maxima importants de l'enveloppe spectrale avec leurs évolutions. On a aussi testé la possibilité de segmentation des logatomes en suite de sons voisés et non voisés.

I. EXPERIMENT PRINCIPALE

La notion d'inhibition latérale est une notion classique reconnue dans le fonctionnement du système auditif. Ce phénomène a été utilisé par Karnickaya et al. (1973) et repris, avec succès semble-t-il, dans un système de reconnaissance par Zagoruiko et Lebedev (1974, 1985). Par ailleurs, une étude détaillée récente sur l'inhibition latérale est développée par Shamma (1985).

Dans ce travail, nous testons ce phénomène en vue de l'extraction des maxima importants d'une enveloppe spectrale et leurs évolutions pour application en reconnaissance de la parole.

Le signal original est traité tout d'abord par une transformée de Fourrier. Ensuite, on applique un filtre d'inhibition latérale dont les caractéristiques sont représentées par la figure 1.

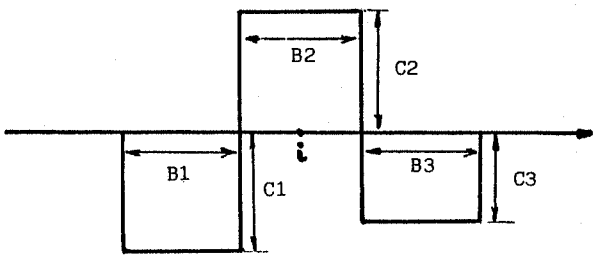


Figure 1. La loi d'inhibition laterale

Le signal de sortie S(i) de ce filtre est la somme pondérée d'un ensemble de composantes. Ce filtre comprend 3 plages : une plage centrale et deux plages d'inhibition.

$$S(i) = - C1\sum_{j\subset B1} e(j) + C2\sum_{j\subset B2} e(j) - C3\sum_{j\subset B3} e(j)$$

On a testé un lissage par intégration spectrale pondérée pour améliorer les résultats de traitement : integration sur 16 points placée avant l'inhibition latérale, intégration sur 1 Bark placée avant ou après l'inhibition latérale, intégration sur 3 Barks placée après l'inhibition latérale.

Pour étudier les possibilités d'extraction des maxima importants, on a utilisé un signal de synthèse pour les voyelles /a/, /i/ et /u/ avec une fenêtre d'analyse de 256 points. Pour étudier l'évolution des maxima on a utilisé des logatomes synthétisés ou bien naturels.

Enfin, on a voulu exploiter des caractéristiques d'aplatissement spectral lors de transition pour tester des capacités de segmentation. La première approche est d'utiliser une inhibition renforcée pour ne recevoir que quelques maxima très importants dans le cas de signal périodique et, dans le cas de bruit, il n'y a pas de détection. La deuxième approche est de chercher un seuil de dynamique de spectre lissé pour pouvoir le classer comme plat ou non plat.

II. RESULTATS ET DISCUSIONS

1. Etude de la détection des maxima.

La figure 2 représente les résultats de test sur la voyelle /i/. Dans tous les cas, la courbe 1 représente la FFT, la courbe 2, la FFT + intégration 1 Bark, la courbe 4, le Cepste.

La courbe 3 de la figure 2a représente la FFT + inhibition.

La courbe 3 de la figure 2b permet de voir le rôle d'une intégration 1 Bark avant le filtre d'inhibition.

La courbe 3 de la figure 2c montre le rôle de l'intégration 1 Bark lorsqu'elle est placée après le filtre d'inhibition.

Figure 2. Les résultats sur voyelle /i/
Courbe 1 : FFT
Courbe 2 : FFT + intégration 1 Bark
Courbe 4 : Cepstrum
Courbe 3 :
Figure 2a : FFT + inhibition
Figure 2b : FFT + 1 Bark + inhibition
Figure 2c : FFT + inhibition + 1 Bark
Figure 2d : FFT + inhibition + 3 Bark

Figure 2a

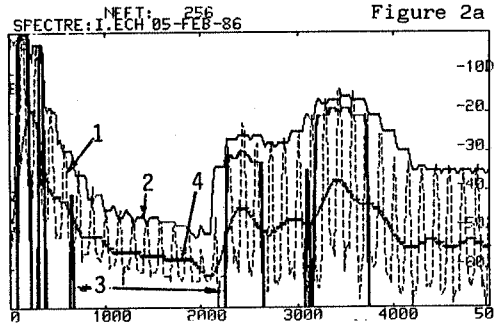


Figure 2b

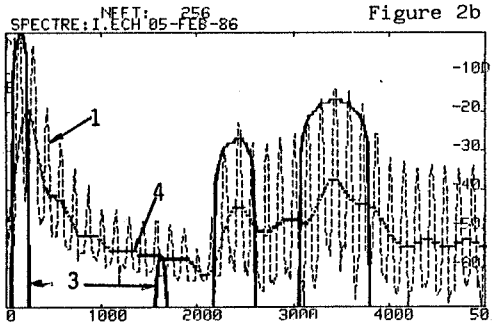


Figure 2c

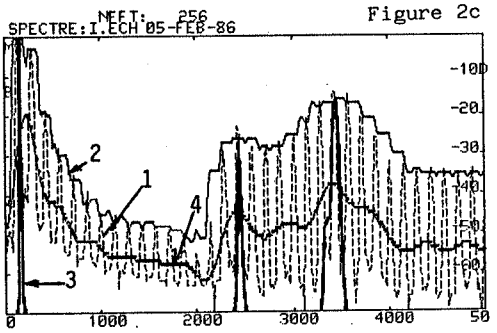
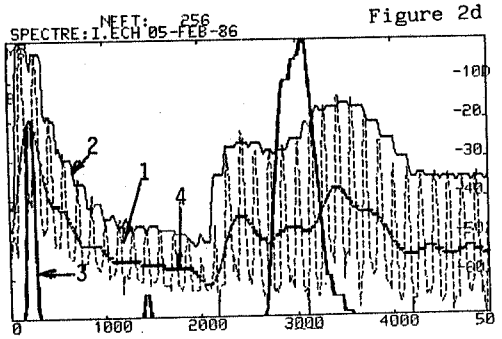


Figure 2d



La courbe 3 de la figure 2d donne une première information sur le rôle de l'integration 3 Barks (théorie de centre de gravité - Chistovich 1979).

On peut faire les constatations suivantes :

- L'inhibition joue un rôle d'amplification de constraste et, en particulier, avec l'intégration 1 Bark, de suppression de petites variations spectrales.
- En basse fréquence, le résultat présente des harmoniques. Ceci est cohérent avec les résultats de Shamma (1985), Zagoruiko et Lebedev (1974), Chistovich (1974) et Evans (1982). Avec une intégration, ce phénomène peut disparaître, mais si la fondamentale et le premier formant sont proches, on ne peut détecter que la fondamentale (pour /i/ par exemple).
- L'intégration 1 Bark placée après l'inhibition donne des pics de formants beaucoup plus pointus.
- Une intégration 3 Barks permet de détecter premier formant et regroupe les formants élevés dans une perspective de centre de gravité.
- La position des formants est correcte par rapport aux données Cepstrales.
- Une augmentation de C1, C3 et B1, B3 provoque une diminution de la détection des petites irrégularités et rend plus pointu les pics détectés.
- La première plage joue un rôle plus important que la troisième plage.

2. Etude de l'évolution des maxima.

La figure 3 représente le résultat sur traitement de type FFT + intégration 1 Bark + inhibition du mot /BAKI/ prononcé par un locuteur masculin.

On constate une augmentation des constrastes par rapport à un sonagramme classique. La fondamentale est clairement détectable. Le premier formant est généralement mal représenté. Ce qui n'est pas le cas pour les formants supérieurs. Les transitions et les sons faibles sont ici fortement accentués de par le type de représentation adopté, la normalisation en amplitude se faisant pour chacune des trames d'analyse.

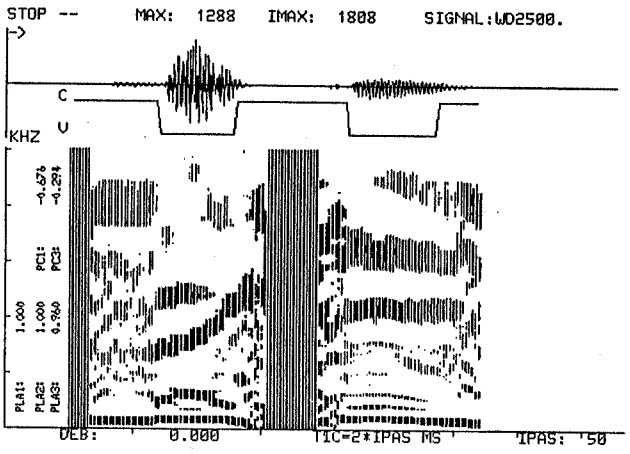


Figure 3. Spectrum de /BAKI/ avec algorithme de détection de voyelle.

3. Tentative de segmentation.

La figure 4 montre le résultat obtenu sur le mot /BABAB/ synthétisé avec un traitement de type FFT + inhibition mais B2 est petit et C1, C3 sont grands. On peut noter toutes les irrégularités temporelles du signal.

Sur la figure 3, on voit aussi le résultat de segmentation en utilisant une deuxième approche. Le seuil du dynamique utilisé est 10. Le signal est normalisé avant FFT par sa valeur maximale. Sur un test sur 110 logatomes CVCV enregistrées on constate qu'il n'y a que 9 erreurs qui parviennent de la consonne /ʃ/. Ces erreurs peuvent être limitées en tenant compte de l'effet de l'adaptation (40 ms - Chistovich 1974).

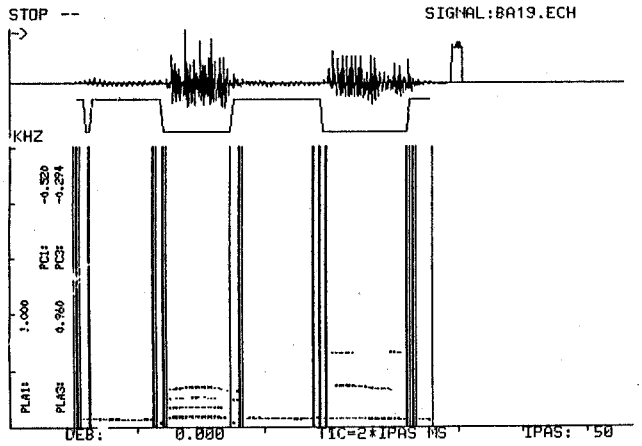


Figure 4. Segmentation obtenue par inhibition latérale renforcée

BIBLIOGRAPHIE

1. M. Blomberg, R. Carlson, K. Elenius and B. Granstrom (1983)
Auditory models and isolated word recognition.
STL-QPSR, **4**, 1-15.

2. I.A. Chistovich & al. (1974)
A functional modele of signal processing in the peripherique system.
Acustica, **31**, 349-353.

3. L.A. Chistovich, R.L. Sheikin and V.V. Lublinskaya (1979)
"Center of gravity" and spectral peaks as the determinants of vowel quality.
In **"Frontiers of speech com. research"**, ed. B. Limdblom and S. Ohman, 143-157.

4. E.F. Evans (1982)
Representation of complex sounds at cochlear nerve and cochlear nucleus levels.
In "The representation of speech in the auditory system" ed. R. Carlson and B. Granstrom, 27-42.

5. N.G. Karnickaya, V.N. Mushnikov, N.A. Slepokurova and S.Ja. Zhukov (1973)
Auditory processing of steady-stats vowel.
Symposium on Auditory Analysis and Perception of speech, 1-10, Leningrad.

6. V.G. Lebedev and N.G. Zagoruiko (1985)
Auditory perception end speech recognition.
Speech communication, **4**, 97-103.

7. Shihah A. Shamma (1985)
Speech processing in the auditory system II: Lateral inhibition and central processing speech evoked activity in the auditory nerve.
J. of Acoust. Soc. of Am., **78**, 1622-1632.

8. N.G. Zagoruiko and V.G. Lebedev (1974)
Models for speech signals analysis taking into account the effect of masking.
Acustica, **31**, 346-348.

ETUDE DE LA RECONNAISSANCE, AUTOMATIQUE ET PAR UN PATIENT IMPLANTE, D'UNE LISTE DE MOTS ARTIFICIELS

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

*Laboratoire G.B.M. Université Lyon I - **C.N.E.T. Grenoble

ABSTRACT

In this work the authors have studied the results of the recognition, automatically performed by a machine and by a patient fitted with a Chorimac cochlear prosthesis, of a list of artificial words.

In each one of these situations the recognition has been performed individually with each one of the prosthesis 12 channels and globally with the seven channels working on the patient who participated at this work.

The results show that :
*the recognition of the vowels used in this work is roughly equivalent for the machine and for the patient,
*the distinction between the plosive and the fricative consonants is better performed by the patient than by the machine,
*the perception of voicing is not very good, mostly with the automatic system,
* the mean performances observed with the patient and with the machine are similar,
* the increase of the number of channels is better used by the machine than by the patient.

It must be kept in mind that these results have been obtained so far on a small number of tests but the methodology seems worthwhile to be considered for the exploration of some high functions in the human brain.

INTRODUCTION

Depuis les expériences de Djourno et Eyries en 1957, la stimulation directe du nerf cochléaire [1] des personnes atteintes de cophose bilatérale profonde a ouvert la voie à un domaine de recherche passionant sur les mécanismes de perception de l'audition [2].

A priori, le principe de la transmission auditive chez ces patients est simple ; il suffit d'envoyer à l'implanté des informations, les plus complètes possibles, pour qu'il décode et reconnaisse le message acoustique. L'introduction d'un grand nombre d'électrodes réparties le long de la cochlée semble être une solution satisfaisante puisqu'elle transmet des indications sur la hauteur et l'intensité des sons. Mais, s'il semble très intuitif que le fait d'augmenter le nombre d'électrodes doit conduire à une meilleure reconnaissance de la parole, certains travaux interdisent d'être aussi sûr de l'exactitude de ce postulat [3,4].

Ce résultat, connu au niveau des vocodeurs, mérite d'être vérifié avec les implantés. Dans une première étape, nous nous sommes attachés à comparer les performances obtenues par un système de reconnaissance automatique de mots et par un patient implanté qui a largement collaboré à ce travail.

Pour l'instant, on gardera en mémoire que le travail présenté ici comporte un grand nombre d'unicités :
- un seul patient,
- une seule méthode de reconnaissance,
- un vocabulaire spécifique,
.... et on pondèrera les conclusions en conséquence.

En partant du principe que les J.E.P. sont un lieu privilégié de discussion sur les travaux qui touchent le langage, nous souhaitons présenter nos premiers résultats, dans un domaine qui approche l'intelligence artificielle et l'intelligence naturelle, et qui peut à terme donner des indications sur des modes de fonctionnement du cerveau humain ; enfin, telle est notre espérance.

Le système qui a été utilisé pour ce travail est la prothèse cochléaire Chorimac [5], fabriquée par les laboratoires Bertin à Aix en Provence, et qui est implantée en France par les équipes des professeurs Chouard (Hôpital St Antoine, Paris), Charachon (Hôpital des Sablons, Grenoble) et Morgon (Hôpital Edouard Herriot, Lyon).

Le travail d'évaluation qui a été entrepris, jusqu'à présent, a surtout eu pour but de quantifier les performances des patients qui étaient implantés par cet appareil [6,7,8]. Mais, comment l'information contenue dans l'analyse spectrale effectuée par la prothèse cochléaire est-elle utilisée par le patient ?

De plus, la présence d'un grand nombre d'électrodes est-elle, dans tous les cas, un avantage certain pour le patient ?

Apporter des éléments de réponse à ces questions est le but principal de ce travail.

Le protocole que nous avons mis sur pied a été de comparer les performances obtenues, par un système automatique et par un patient implanté, lors de la reconnaissance de mots artificiels. Cette étude s'intègre dans le cadre des travaux sur la perception [9,10].

MATERIEL ET METHODES

Fonctionnement de la prothèse

La prothèse Chorimac est essentiellement formée de deux parties (figure 1) :

* une externe (l'émetteur) qui détecte, à l'aide d'un banc de 12 filtres passe-bande, l'énergie contenue dans le message vocal.
L'énergie correspondant à chacun des filtres sert à moduler la longueur d'implusions électriques d'amplitude constante.
Ces impulsions modulent l'amplitude d'une onde électromagnétique porteuse qui va franchir la barrière cutanée.

* une interne (le récepteur) qui détecte la largeur des impulsions et distribue sur des électrodes une énergie électrique correspondante.
Ces électrodes sont distribuées le long de la cochlée pour restituer au patient une sensation de tonalité. Toutes les 3 millisecondes, un train de 12 impulsions est ainsi émis pour l'oreille interne du patient.

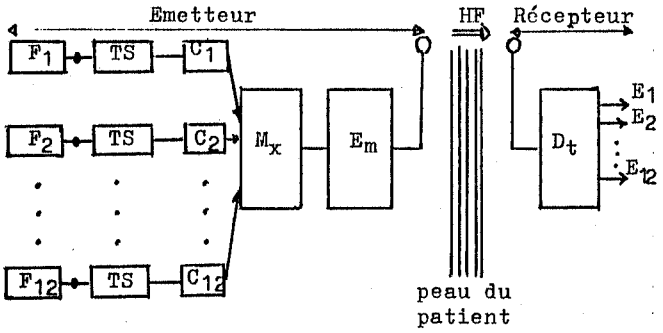


Figure 1 - Représentation schématique du système Chorimac :
- $F_1, F_2, \ldots, F_{12}$ sont les filtres,
- TS est le traitement du signal,
- $C_1, C_2, \ldots, C_{12}$ sont les canaux,
- M_x est un multiplexeur assemblant en série les impulsions issues de $C_1, C_2, \ldots, C_{12}$,
- E_m est un étage d'émission ; les impulsions modulent une porteuse HF,
- D_t est un détecteur ; les impulsions venant de $C_1, C_2, \ldots, C_{12}$ sont distribuées sur les électrodes $E_1, E_2, \ldots, E_{12}$.

Mots du vocabulaire

Le vocabulaire qui a été utilisé est formé à partir de trois oppositions phonétiques simples [11] :
* voyelles (/a/ , /o/ , /i/),
* voisé/non voisé ({/d/,/z/} , {/t/,/s/}),
* plosif/sifflant ({/d/,/t/} , {/z/,/s/}).

Les traits phonétiques ont été associés sous la forme consonne/voyelle ; pour une meilleure reconnaissance, chacune des syllabes élémentaires ainsi formées a été répétée pour former un mot. Les 12 mots ainsi obtenus, de façon analytique, sont les suivants :

DADA , TATA , SASSA , ZAZA ,
DODO , TOTO , SOSSO , ZOZO ,
DIDI , TITI , SISSI , ZIZI.

Ces mots ont été répétés au hasard pour former une liste de 129 mots qui a été présentée à l'implanté, sans lecture labiale, et au système de reconnaissance automatique.

Reconnaissance automatique des mots

A l'aide d'un dispositif électronique rapide, nous avons enregistré l'évolution temps-fréquence de la durée des impulsions électriques correspondant à l'analyse spectrale effectuée par la prothèse Chorimac [12].
Des représentations ont ainsi été obtenues pour chacun des 129 mots de la liste, représentations qui seront appelées "spectrogrammes" par la suite.
Dans chacun de ces spectrogrammes, un algorithme de "segmentation" élémentaire a permis d'isoler un échantillon de 3 millisecondes correspondant à la première voyelle et à la seconde consonne.
Pour des raisons de fiabilité, ce travail a été effectué sur des spectrogrammes lissés ; la fenêtre de lissage est de 24 échantillons.
Le repérage sur les spectrogrammes lissés d'un échantillon de 3 millisecondes pour les voyelles et pour les consonnes a permis de caractériser chaque prononciation des mots de la liste par 24 durées d'impulsion, prononciation qui a été représentée ainsi par un point dans un espace à 24 dimensions.

Chacune des prononciations d'un mot a été comparée aux 128 autres, réunies en classes (chaque classe correspond à un mot). Dans chacune des classes, le plus proche voisin a été déterminé en utilisant une métrique euclidienne :

$$d(X, M^i) = \sum_{j=1}^{24} \left[(x_j - m_j^i)^2 \right]^{1/2} \qquad (1)$$

où :

X_i est le mot à classer,
M^i représente un des 128 autres mots de la liste,
x_j est la j^e coordonnée de X,
m_j^i est la j^e coordonnée de M^i.

Le mot X a ainsi été attribué à la classe $_kC$ qui contenait le plus proche des plus proches voisins de X. Cette politique de reconnaissance a été retenue, compte tenu des bons résultats qu'elle donnait en reconnaissance, relativement à d'autres métriques [3].

Un tableau 12 x 12, indiquant les reconnaissances et les confusions des mots, a été établi; en regroupant des lignes et des colonnes [13] nous avons établi les pourcentages de reconnaissance des voyelles, des consonnes, du voisement et de l'opposition entre les traits plosif et sifflant.

On a défini ensuite un facteur de qualité global pour chaque reconnaissance de la liste, à partir des pourcentages de succès observés pour chacun des traits du langage. Un rééquilibrage a enfin été effectué en considérant le fait que la reconnaissance a priori des voyelles est plus faible (1/3) que celle du voisement (1/2) ou des traits plosif et sifflant (1/2).

Le pourcentage moyen attaché à la reconnaissance d'une liste est alors :

$$\overline{P} = \frac{1}{7} \left(3 P_V + 2 P_{\overline{vv}} + 2 P_{P/S}\right) \qquad (2)$$

où :

* P_V est le pourcentage de reconnaissance des voyelles,
* $P_{\overline{vv}}$ et $P_{P/S}$ sont les pourcentages de détection, respectivement, du voisement et de l'aspect plosif/sifflant pour les consonnes.

Reconnaissance des mots par l'implanté

L'implanté qui a collaboré à ce travail est un homme de 57 ans implanté il y a trois ans à Lyon. Ce patient était atteint d'une cophose bilatérale tardive depuis 10 ans environ au moment de l'implantation.

Il possède une très bonne lecture labiale, mais seulement sept électrodes (ou canaux), 1/7/8/9/10/11/12, sont fonctionnelles sur son appareil.

Sa collaboration à ce travail a été parfaite ; toutes les reconnaissances ont été effectuées sans l'aide de la lecture labiale.

Sélection du meilleur canal

Il est possible de sortir tous les filtres sur l'émetteur de la prothèse Chorimac.

La liste a donc été lue à l'implanté en ne laissant en place qu'un seul filtre. La reconnaissance de chaque lecture de la liste a été quantifiée par un pourcentage moyen $\overline{p}$, conformément à la formule (2).

Le meilleur des canaux a ainsi été défini. Pour améliorer la fiabilité des résultats, chaque lecture de la liste a été précédée d'un petit apprentissage de dix à quinze minutes environ, pour permettre à l'implanté de repérer les sons, dans une structure acoustique totalement nouvelle pour lui.

Performances avec les canaux

La meilleure des électrodes (c'est à dire des canaux) ayant ainsi été déterminée, nous avons enlevé les 12 filtres de la prothèse pour éviter de parasiter la détection par des effets de bord dus aux filtres voisins.

Chacun des 12 filtres a été ensuite successivement introduit sur le meilleur canal déterminé précédemment et les performances ont été analysées conformément à la procédure décrite au paragraphe précédent.

Les 12 bandes passantes des filtres sont les suivantes :

1	:	0	-	240 Hz
2	:	240	-	350 Hz
3	:	350	-	450 Hz
4	:	450	-	600 Hz
5	:	600	-	800 Hz
6	:	800	-	1250 Hz
7	:	1250	-	1490 Hz
8	:	1490	-	1850 Hz
9	:	1850	-	2500 Hz
10	:	2500	-	3500 Hz
11	:	3500	-	4900 Hz
12	:	4900	-	7500 Hz

RESULTATS

Reconnaissance automatique des mots

En utilisant la métrique euclidienne et le plus proche voisin pour archétype, on a reconnu les 129 mots de la liste et déterminé les pourcentages de détection suivants :

* des consonnes,
* des voyelles,
* du voisement,
* du trait plosif/sifflant,

ainsi que la moyenne pondérée correspondant à la formule (2).

La liste a été reconnue dans les circonstances suivantes :

* l'information issue de chacun des filtres est successivement considérée,
* les sept canaux fonctionnels chez le patient sont pris en compte,
* les 12 canaux sont utilisés.

L'ensemble des résultats est indiqué sur le tableau I.

Reconnaissance par l'implanté

Présélection du meilleur canal

La reconnaissance de la liste n'a pu être effectuée valablement que pour les canaux 1,7, 9, 10, 11 isolés. Pour les autres, l'intensité sonore était trop faible, l'implanté ne percevant que des "claquements" dans son appareil.

Les résultats sont indiqués sur le tableau II. Le canal 10, conduisant au meilleur résultat, sera retenu pour la suite de cette étude. On voit que les canaux 9 et 10 donnent nettement les meilleurs pourcentages de reconnaissance (moyenne pondérée).

Etude sur le canal 10

La reconnaissance de la liste a ensuite été proposée à l'implanté en mettant, successivement, sur la voie 10, chacun des filtres.

filtres	voyel-les	voise-ment	opposition plosive sifflante	moyenne pondérée	clas-sement
1	44,2	50,3	54,3	48,8	8
2	62,7	55	51,9	57,4	4
3	62	55,8	52,7	57,6	3
4	64,3	47,3	60,5	58,4	2
5	65,1	52,7	47,3	56,5	5 ex.
6	61,3	61,3	57,4	60,2	1
7	62,8	55,8	46,5	56,1	7
8	58,9	54,3	55	56,5	5 ex.
9	45,7	46,5	51,2	47,5	9
10	34,1	55	55,8	46,3	11
11	41,9	48,8	52,7	47,0	10
12	35,6	50,4	49,6	43,8	12
12 canaux	96,12	69,76	68,26	80,6	
7 canaux du patient	68,21	63,56	67,44	66,7	

Tableau I - Pourcentages de reconnaissance obtenus à l'aide de la reconnaissance automatique.

* * * *

filtre	Mots	cons.	voy.	voise-ment	oppos. plos./ siff.	moy. pondé-rée	clas. sement
1	13,7	34,6	32,6	64,5	52,7	47,6	5
7	15,4	31,5	46,6	51,9	58,6	51,4	4
9	23,3	46,6	50,4	58,6	79,8	60,1	2
10	25,6	48,5	50,4	59,7	76,2	60,6	1
11	23,3	49,6	41,5	53,9	80,6	56,7	3

Tableau II - Pourcentages de reconnaissance obtenus par l'implanté ; chacun des filtres est placé sur sa voie normale.

* * * *

Une reconnaissance globale a aussi été effectuée, le patient disposant alors de ses sept filtres fonctionnels.

filtres	voyel. les	voise-ment	opposition plosive sifflante	moyenne pondérée	clas-sement
1	31,8	62,0	46,5	44,6	12
2	28,7	70,0	45,0	45,2	10
3	64,3	56,6	46,5	57	7
4	39,6	52,7	44,2	44,7	11
5	79,0	51,9	49,6	62,9	3ex
6	72,9	57,3	53,4	62,9	3ex
7	70,6	63,6	52,7	63,5	2
8	60,5	59,7	74,4	64,2	1
9	49,6	58,9	79,8	60,9	5
10	50,4	55,8	77,5	59,7	6
11	28,7	49,6	79,8	49,3	9
12	24,0	59,7	78,3	49,7	8
Total	41,9	57,4	63,6	52,5	

Tableau III - Pourcentages de reconnaissance obtenus avec l'implanté (les filtres sont placés sur la voie 10).

DISCUSSION

En comparant, filtre à filtre et pour chacun des éléments du langage les résultats obtenus par la reconnaissance automatique et par l'implanté, on remarque que :

* pour les voyelles, la reconnaissance automatique est bonne entre les filtres 2 et 9, ce qui correspond aux performances de l'implanté ;

* pour le voisement, l'implanté a des résultats plutôt satisfaisants sur les filtres 1 et 2 ; par contre le système automatique distingue mal ce caractère ;

* pour la discrimination des traits plosif/sifflant, le traitement automatique est peu performant, tandis que l'implanté est très efficace en haute fréquence à partir du canal 8 ; on peut penser que le filtrage temporel introduit pour la localisation des phonèmes n'est pas favorable à la détection des phénomènes rapides et qu'il est responsable de la faiblesse des résultats obtenus avec le traitement automatique [14] ;

* la qualité globale de la reconnaissance est à peu près constante sur l'ensemble du spectre, pour le système automatique et pour l'implanté.

Une interprétation plus fine de ces résultats n'a pas de sens actuellement, à notre avis, compte tenu de la particularité du travail qui a été effectué.

La reconnaissance effectuée avec un seul canal est, dans ce travail, aussi bonne voire même meilleure que celle effectuée avec plusieurs canaux, pour l'implanté.
Par contre, les techniques automatiques conduisent à de meilleurs résultats lorsque le nombre de canaux est plus élevé.
Ce point mérite d'être approfondi.

En comparant les canaux(tableaux II et III) on remarque que les filtres 7 et 11 donnent de meilleurs résultats lorsqu'ils sont placés devant des canaux qui passent bien, plutôt que devant ceux qui leur sont dévolus par la prothèse Chorimac et auxquels le patient est habitué.

CONCLUSIONS

La comparaison entre les performances obtenues par une reconnaissance automatique et par un implanté montre que la correspondance entre les possibilités de ces deux "systèmes de reconnaissance" n'est pas absolue et que dans l'ensemble, le patient n'a pas grand chose à envier aux algorithmes mathématiques que nous avons employés.

Les résultats du patient sont nettement supérieurs pour la séparation des traits plosif et sifflant ; ils sont comparables pour les voyelles et un peu meilleurs pour le voisement.

En outre, pour le patient avec lequel nous avons travaillé et sur le vocabulaire réduit que nous étudions, le bénéfice de l'emploi de plusieurs électrodes n'est pas nettement établi ; par contre, le traitement automatique des spectrogrammes donne, avec l'ensemble des électrodes, des résultats très supérieurs à ceux qui sont obtenus avec une seule électrode. Cette différence dans la discrimination montre que les travaux pour comprendre le mécanisme humain de la reconnaissance phonétique sont loin d'être terminés.

BIBLIOGRAPHIE

[1] A. Djourno, C. Eyries, "Prothèse auditive par excitation électrique à distance du nerf sensoriel à l'aide d'un bobinage inclus à demeure", Presse Méd., Vol 35, pp 1417-1423, 1957.

[2] J.C. Lafon et al, "L'information acoustique à travers la cochlée", Rev. d'Acoustique, Vol 42, pp 258-260, 1977.

[3] R. Mouhssine, "Prothèse cochléaire. Evaluation objective du codage mis en oeuvre par le système Chorimac", Thèse 3e cycle, Lyon, Nº 1654, 1985.

[4] M. Couvrat, C. Gagnoulet, "Optimisation de l'analyse en vue de la reconnaissance globale", Ref 286, LAR - TSF ATP, CNET Lannion (12 p), 1981.

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

[6] C. Fugain, B. Meyer, F. Chabolle, C.H. Chouard, "Clinical Results", 2e Symposium sur les implants cochléaires, Paris, Proc. pp 237-246, 1983.

[7] D. Morel, R. Charachon, J. Genin, "Cochlear Prosthesis ; Clinical Results", 2e Symposium sur les implants cochléaires, Paris, Proc pp 254-256, 1983.

[8] A. Morgon, C. Berger-Vachon, J.M. Chanal, G. Kalfoun, C. Dubreuil, "Cochlear Implants. Experience of the Lyon Team", 2e Symposium sur les implants cochléaires, Paris, Proc pp 195-203, 1983.

[9] C. Sorin, "Spectres psychoacoustiques des consonnes vélaires en français", 14e J.E.P. Paris, Proc pp 6-9, 1985.

[10] J. Genin, "Application du test de rimes à des patients sourds totaux", 14e J.E.P., Paris, Proc pp 10-12, 1985.

[11] A. Morgon, C. Berger-Vachon, "Implantations cochléaires. Premiers résultats de l'Ecole Lyonnaise", Ann. Franç. des Microtech. et de Chronom., Vol 37, pp 79-82, 1983.

[12] R. Mouhssine, "Analysis of the Spectral Information Produced by the Implanted Cochlear Prosthesis Chorimac", 4e séminaire espagnol, marocain, portugais sur le traitement du signal et ses applications. Marrakech, Proc part.B/4, 1984.

[13] P. Benfaï, A. Karczag, SR Petra Lüers, "Clinical Results ; the Rehabilitation", 2e symposium sur les implants cochléaires, Paris, Proc pp 183-194, 1983.

[14] R. Charachon, J. Genin, "Appareillage de prothèse cochléaire", Rapport MRT/GBM 82 MO1033/34 (66 p), 1985.

LE SYSTEME AUDITIF HUMAIN COMPREND-IL UN MECANISME D'INTEGRATION SPECTRALE A LARGE BANDE ?

J.L Schwartz, P.Escudier

Institut de la Communication Parlée - Grenoble

ABSTRACT

The concept of large scale frequency integration ("center of gravity" with a 3.5 Barks "critical distance") has been developed by Chistovich et al. (/4/, /5/) and used to various theoretic discussions about F'2 predictions /9/ and vowel classification (/6/, /7/).

We present further arguments in favour of this concept. First, we discuss two assumptions about the "perceptual reality" of the F'2 parameter. Secondly, we show how the center of gravity effect and the 3.5 Barks critical distance help understanding the whole set of experimental data about the influence of formant amplitudes on the perception of front vowels. Our conclusion is that it is necessary to assume the existence of such a large scale frequency integration mechanism in the human auditory system.

INTRODUCTION

La notion d'intégration spectrale à large bande a été introduite depuis maintenant plus de 30 ans dans les théories sur le traitement perceptif du spectre des voyelles : elle est à la base des résultats de Delattre et al. sur la synthèse des voyelles par des sons à un ou deux formants /1/. Reprise et développée par Carlson et al. (/2/, /3/) pour l'étude du paramètre F'2, elle a été formalisée par Chistovich et al. (/4/, /5/) sous le nom d'effet de "centre de gravité" (CG) avec l'introduction de la notion de distance critique, c'est-à-dire la proposition d'une largeur de bande pour cette hypothétique intégration : 3 à 3.5 Barks.

Cette théorie, et cette valeur critique de 3-3.5 Barks, ont été à l'origine de récents développements, notamment pour la classification des voyelles (/6/, /7/, /8/). Mais les bases de la théorie restent fragiles et souvent critiquées, explicitement ou implicitement, et remises en cause par un certain nombre de contradictions expérimentales apparentes. Nous voulons montrer ici plus précisément ce que ce concept peut aider à comprendre à divers niveaux, lever certaines contradictions, préciser quelques-unes de ses limites, et proposer un schéma perceptif plus général dans lequel il pourrait s'inscrire cette fois en position de force.

A-T-ON BESOIN D'UN MECANISME D'INTEGRATION A LARGE BANDE POUR COMPRENDRE LES RESULTATS SUR CG ET F'2 ?

Dans un récent article, Bladon /9/ proposait de réaliser un modèle de prédiction de F'2 à partir d'un mécanisme simulant CG, c'est-à-dire d'un mécanisme d'intégration spectrale à 3.5 Barks. Cette hypothèse revient à considérer une structure du système auditif telle que celle décrite sur la Figure 1a, et reprise dans plusieurs articles (/8/, /10/). Mais d'autres hypothèses ont été émises concernant le paramètre F'2. Ainsi, Carlson et al. /2/ se sont demandés si leurs résultats expérimentaux n'étaient pas déduits de mécanismes de perception catégorielle. Ceci correspondrait à la structure b de la Figure 1, F'2 étant non un indice de la classification, mais au contraire un résultat de la classification. D'autre part, Bladon et Lindblom /11/ ont montré que certaines valeurs de F'2 (leur étude porte sur la région de [i] et [y]) pouvaient être prédites par un modèle simple d'estimation de distances entre spectres : les signaux référence (à 4 formants) et test (à 2 formants) utilisés dans les expériences correspondantes produiraient des spectres d'excitation interne (en sone/Bark) que le système de décision comparerait systématiquement pour choisir la valeur de F'2 produisant la distance minimum entre référence et test. Ceci correspondrait à la structure c de la Figure 1, et là encore F'2 ne serait pas un paramètre pour la décision phonémique mais un produit annexe des mécanismes de calcul et comparaison de distances assurant cette décision. Dans ces deux cas, il n'y aurait donc pas de mécanisme "câblé" de mesure de F'2 dans le système auditif. Considérons plus avant ces deux hypothèses.

F'2 est-il déduit de la perception catégorielle ? (Etude de la structure 1b)

Carlson et al. s'appuient, pour émettre leur hypothèse, sur des résultats de mesure de F'2 pour des stimuli de synthèse à 4 formants, avec F1=250 Hz, F2=2000 Hz, F4=3350 Hz, et F3 variant entre 2300 et 3000 Hz. Ils constatent une progression très rapide du paramètre F'2, qui passe d'une valeur proche de F2 à une valeur proche de F4, et même, pour certains ajustements, supérieure à F4. On observe notamment une forte

montée de F'2 autour de F3=2700 Hz. L'hypothèse des auteurs est que ce saut correspond non pas à une discontinuité "auditive", mais à une discontinuité de la perception phonémique, la voyelle perçue passant de [y] à [i] autour de cette position de F3. Or il existe une autre hypothèse : la position F3=2700 Hz correspond justement à peu près au point pour lequel F3 est, en Barks, équidistant de F2 et F4. Aussi, un modèle d'intégration spectrale peut prédire le saut de F'2, selon que la masse principale d'énergie soit constituée par l'ensemble F2-F3 ou F3-F4 (voir /10/, /12/).

Pour déterminer laquelle de ces deux hypothèses était la bonne, nous avons repris les expériences de Carlson et al. en réalisant à la fois des mesures de F'2 et des tests d'identification [i]-[y] avec les mêmes stimuli, et des expériences similaires avec une valeur de F1 égale à 450 Hz (toutes choses égales par ailleurs), de manière à nous déplacer dans la zône [e]-[ø] où la frontière d'identification risquait d'être fortement déplacée. Les résultats portés sur la Figure 2 sont clairs : si la frontière d'identification est, pour le continuum [i]-[y], au-delà de F3=2600 Hz et, pour le continuum [e]-[ø], inférieure à 2300 Hz, les courbes de F'2 dans les deux cas sont identiques. Le saut de F'2 n'est donc pas dû à un mécanisme de perception catégorielle.

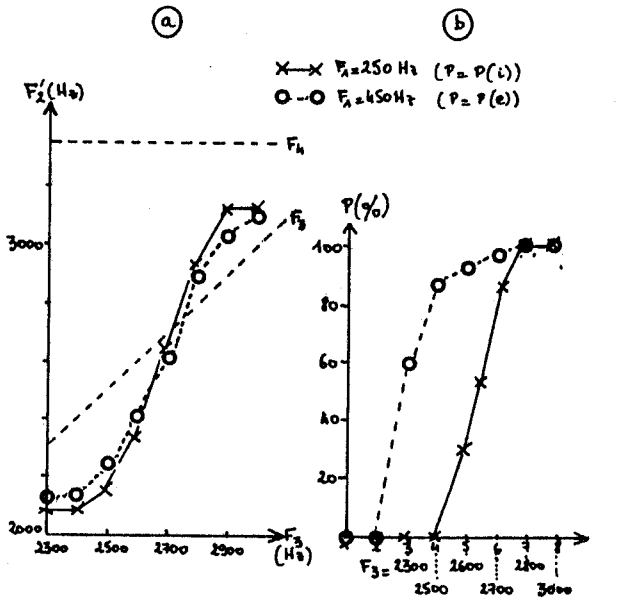


Figure 2 : F'2 et tests d'identification
On a utilisé des signaux de synthèse à 4 formants, avec F1=250 ou 450 Hz, F2=2000 Hz, F4=3350 Hz et F3 variable. En b, les stimuli 1 et 2, qui servent de référence, correspondent à :
1 F1=250 ou 450, F2=1700, F3=2000, F4=3350 Hz
2 F1=250 ou 450, F2=1700, F3=2300, F4=3350 Hz.

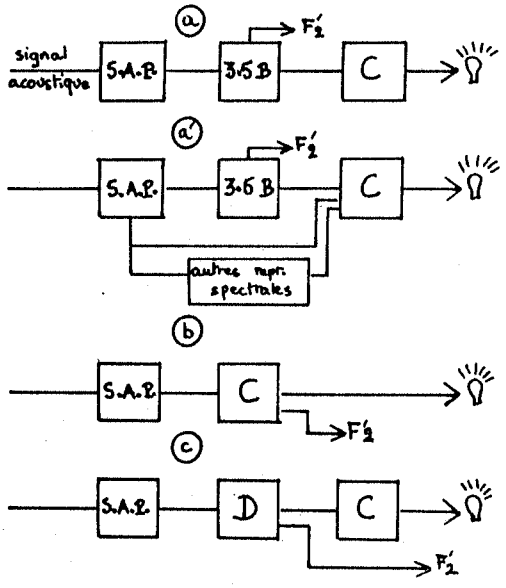


Figure 1 : Structures possibles pour le système auditif (voir texte).
S.A.P. → analyse du signal dans le système auditif périphérique
3.5 B → intégration spectrale à 3.5 Barks
C → mécanisme de classification phonémique
D → calcul de distances pour la classification

Les 3.5 Barks de Chistovich sont-ils déduits des bandes critiques du système auditif périphérique ? (Etude de la structure 1c)

Si l'on admet, par un indispensable souci d'économie, que F'2 et CG sont deux manifestations d'un même phénomène, le travail de Bladon et Lindblom conduit clairement à la question ci-dessus. Peut-on, donc, prédire la distance critique de 3.5 Barks proposée par Chistovich et al. par un simple modèle comparant les distances entre signaux à 1 et 2 formants, sans un intervention d'un mécanisme d'intégration à 3.5 Barks ? La Figure 3a montre les résultats de l'expérience de base de Chistovich. Dans cette expérience, un sujet ajuste le formant F^* d'un signal à 1 formant, pour qu'il soit aussi proche que possible d'un signal à 2 formants F1, F2 avec F2 fixe et F1 décroissant d'un test au suivant. La distance critique est définie par les auteurs par la valeur de F1 au-dessous de laquelle F^* cesse de décroître avec F1 pour revenir vers F2.

La Figure 3b montre la simulation que nous avons faite sur le modèle de Bladon et Lindblom, en déterminant systématiquement, pour un signal de F1 et F2 donnés, la valeur de F^* produisant la distance minimum au signal référence (F1, F2). Nous obtenons, en utilisant le critère de Chistovich et al., des distances critiques portées sur la Table I. Nous avons simulé l'ensemble des résultats quantitatifs concernant l'influence de F1, F2, des amplitudes et des formes des spectres au voisinage

des formants (voir /4/, /5/, /13/). Il est difficile de décider si l'hypothèse est tenable ou pas, les valeurs quantitatives des distances critiques étant fortement dépendantes des hypothèses quantitatives de la simulation (bandes passantes des formants, bandes critiques de l'analyse périphérique, forme exacte de la compression d'amplitude liée à la fonction sone, absence dans le modèle de Bladon et Lindblom de phénomènes de suppression à deux tons, ...). Nous reviendrons sur ce problème dans la conclusion.

Table I : distances critiques observées et simulées

F2 (kHz)	valeurs observées (d'après /5/)	valeurs simulées (modèle d'après /11/)
0.7	3.0 Barks	2.5-3.0 Barks
1.0	3.4 Barks	2.5-3.0 Barks
1.3	3.2 Barks	3.0-3.5 Barks

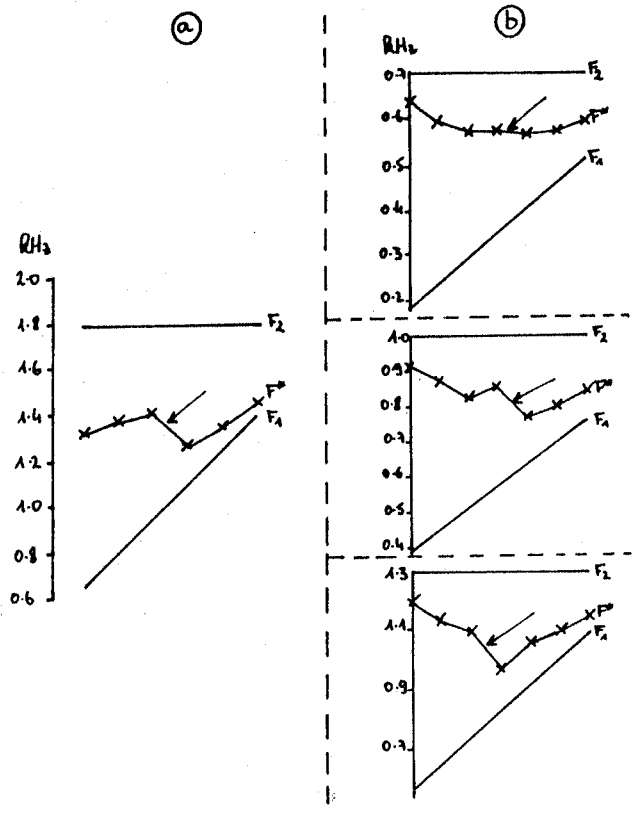


Figure 3 : phénomène du centre de gravité
a) résultats expérimentaux, d'après /4/
b) simulation sur le modèle de Bladon et Lindblom /11/ : pour chaque paire (F1, F2), on porte la valeur de F* produisant la distance minimale, au sens du modèle, entre signal à 2 formants et signal à 1 formant. La flèche signale la remontée de F* correspondant d'après Chistovich et al. à la distance critique entre F1 et F2.

APPORTS DE LA NOTION D'INTEGRATION SPECTRALE AVEC DISTANCE CRITIQUE DE 3.5 BARKS POUR LA COMPREHENSION DE CERTAINS PHENOMENES DE PERCEPTION

A la suite de Bladon /9/, nous avons proposé un modèle analogique /10/, puis une formule /12/ pour prédire F'2 à partir d'un mécanisme d'intégration à 3.5 Barks, modèles qui donnent d'excellents résultats de prédiction sur les corpus de résultats expérimentaux de Carlson et al. /2/ et Bladon et Fant /14/, et constituent un premier support pour cette notion.

Mais nous voulons revenir sur un point délicat qui semble porter une forte contradiction expérimentale : les amplitudes de F2, F3 et F4 jouent-elles un rôle dans la perception phonémique ? Carlson et al. montrent que ce n'est apparemment pas le cas, dans l'étude du continuum [i]-[y] sur des signaux à 4 formants. Cependant, le concept de centre de gravité passe par l'intervention des intensités, qui jouent d'ailleurs un rôle net dans les résultats d'Aaltonen /15/ sur des signaux à 3 formants. Or, la différence entre signaux à 3 et 4 formants pourrait jouer un rôle crucial, qui se comprend bien sur la Figure 2. En effet, on voit sur cette figure que pour les signaux à 4 formants utilisés précisément par Carlson et al. pour leurs tests d'identification, la frontière [i]-[y] est obtenue pour une valeur F3=2700 Hz qui correspond, on l'a vu, à un saut brusque de F'2. Ce phénomène, nous l'avons montré, est une coïncidence entre deux faits indépendants. Mais on comprend bien alors que même si les rapports d'intensité entre F2, F3 et F4 sont modifiés, le saut de F'2 se produira toujours à peu près à la même valeur F3=2700 Hz, milieu en Bark de F2 et F4 : la frontière [i]-[y] sera donc peu sensible à ces variations d'amplitude. Au contraire, dans la Figure 4, on observe les résultats d'une expérience d'identification sur le continuum [e]-[ø], avec F1=450 Hz, F2=1700 Hz, F4=3350 Hz et F3 variant entre 2000 et 2700 Hz, et pour des valeurs de A2 et A3 standards (déduites d'un modèle série), puis pour A'2=A2-12 dB, enfin pour A'2=A2-12 dB et A'3=A3+12 dB. Les résultats varient d'un sujet à l'autre, mais la tendance est évidente : la frontière varie nettement, dans ce cas, en fonction des amplitudes des formants. Ainsi, les amplitudes de F2, F3 et F4 interviennent dans la perception des voyelles avant, et la notion d'intégration spectrale à 3.5 Barks rend bien compte de l'ensemble des résultats expérimentaux sur le sujet.

Il reste sur cette figure à expliquer les caractéristiques non monotones clairement reproductibles d'un sujet à l'autre. L'observation des spectres montre que le fort niveau de réponses de [e] pour le 1er son de la 2ème série, et pour les 2 ou 3 premiers (selon les sujets) de la 3ème série correspond à des cas où F2 n'apparait pas dans le spectre, "écrasé" par F3. On retrouve donc un point clé de la théorie de Chistovich /16/ : avant l'intégration à 3.5 Barks, il existe un mécanisme de détection et de renforcement des petits pics spectraux. C'est probablement ce mécanisme qui intervient par exemple pour F2 pour le 2ème son de la 2ème série, et pas pour le 1er, qui explique le recul de F'2, et la chute de réponses [e].

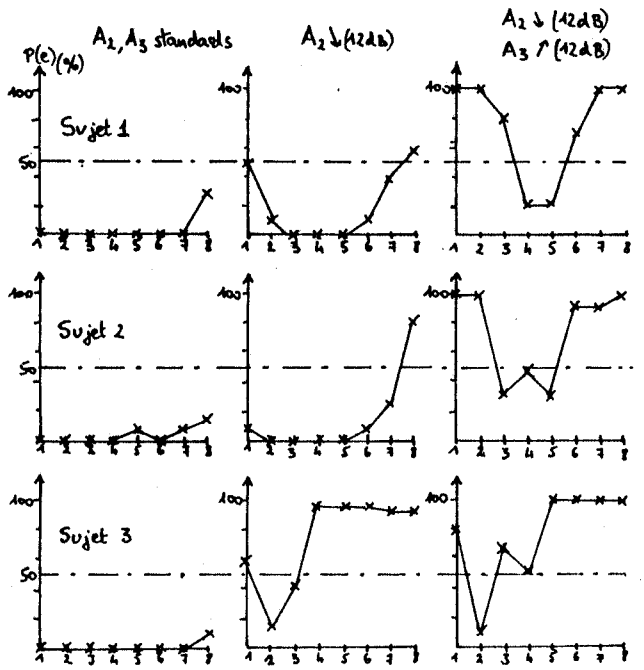


Figure 4 : rôle des amplitudes de F2 et F3 sur la frontière d'identification arrondi-non arrondi. Identification e - , pour 3 sujets (1 sujet par ligne), et 3 séries de stimuli correspondant à 3 caractéristiques d'amplitudes différentes (1 série par colonne).

PORTEE ET LIMITES DU CONCEPT D'INTEGRATION A 3.5 BARKS

Nous avons montré d'abord qu'une des hypothèses "réductionnistes" remettant le paramètre F'2 au rang de simple produit de la classification n'était pas viable. Dans un second temps, nous avons montré ce que l'hypothèse d'un mécanisme "câblé" d'intégration à 3.5 Barks dans le système auditif permettait de comprendre sur la perception des voyelles avant. D'autres travaux montrent d'autres applications en phonétique (/6/, /7/, /8/, /17/), en reconnaissance de parole /12/. L'ensemble de ces faits nous conduit à repousser la structure 1c (que nous n'avons pas pu mettre directement en défaut) au même titre que la structure 1b, et à postuler la réalité de la structure 1a, donc, en corrolaire, la réalité du paramètre F'2 comme paramètre de la perception des voyelles.

Néanmoins, il est nécessaire, si l'on veut comprendre de nombreux résultats de perception tels que ceux de Fujimura /18/, d'Aaltonen sur les temps de réponse de ses sujets /15/, de Klatt sur le rôle des vallées spectrales dans l'établissement des distances perceptives /19/, ou ceux concernant le rôle de B1 pour la perception de la nasalité /20/, de proposer une structure modifiée, nommée a' sur la Figure 1 : dans cette structure, l'intégration spectrale à 3.5 Barks ne fournit qu'une des représentations internes accessibles au système de décodage, représentation qui viendrait s'emboîter dans une série de représentations multiples du même signal. Ce n'est que dans ce cadre, essentiel selon nous pour décrire l'ensemble des mécanismes perceptifs de traitement et de décodage /21/, que l'on peut comprendre le phénomène de centre de gravité proposé par Chistovich et son équipe.

REMERCIEMENTS

De son origine à ses conclusions, ce travail doit beaucoup aux lumières tous azimuths de Christian Abry.

REFERENCES

/1/ P.C. Delattre, A.M. Libermann, F.S. Cooper, L.J. Gertsmann, "An experimental study of the acoustic determinants of vowel colour : Observations on one- and two-formant vowels synthesized from spectrographic patterns", Word, Vol.8, pp. 195-210, 1952.

/2/ R. Carlson, B. Granstrom, G. Fant, "Some studies concerning perception of isolated vowels", STL-QPSR 2-3, pp. 19-35, 1970.

/3/ R. Carlson, G. Fant, B. Granstrom, "Two-formant models, pitch and vowel perception", dans : Auditory analysis and perception of speech, édité par G. Fant et M.A.A Tatham, Academic Press, New-York, pp. 52-82, 1975.

/4/ L.A. Chistovich, R.L. Sheikin, V.V. Lublinskaya, "'Centers of gravity' and the spectral peaks as the determinants of vowel quality", dans : Frontiers of Speech Communication Research, édité par B. Lindblom et S. Ohman, Academic Press, London, pp. 143-158, 1979.

/5/ L.A. Chistovich, V.V. Lublinskaya, "The center of gravity effect in vowel spectra and critical distance between the formants : Psychoacoustical study of the perception of vowel-like stimuli", Hearing Research 1, pp. 185-195, 1979.

/6/ G. Fant, "Feature analysis of swedish vowels - a revisit", STL-QPSR 2-3, pp. 1-19, 1983.

/7/ A. Syrdal, "Aspects of a model of the auditory representation of american english vowels", Speech Communication 4, pp. 121-135, 1985.

/8/ H. Traunmüller, "Perceptual dimension of openness in vowels", JASA 69(5), pp. 1465-1475, 1981.

/9/ A. Bladon, "Two-formant models of vowel perception : shortcomings and enhancements", Speech Communication 2, pp. 305-313, 1983.

/10/ P. Escudier, J.L. Schwartz, M. Boulogne, "Perception of stationary vowels : internal representation of the formants in the auditory system and two-formant models", Symposium Franco-Suédois, Grenoble 1985.

/11/ A. Bladon, B. Lindblom, "Modeling the judgment of vowel quality differences", JASA 69(5), pp. 1414-1422, 1981.

/12/ M. Mantakas, J.L. Schwartz, P. Escudier, "Modèle de prédiction du '2ème formant effectif' - F'2- des voyelles . Application à l'étude de l'opposition de labialité pour les voyelles avant du français", présenté à ce congrès.

/13/ V.V. Lublinskaya, R. Carré, P. Escudier, "The study of some conditions determining the auditory perception of the phonetically meaning spectral cues of steady-state synthetic vowels", Symposium Franco-Soviétique, Grenoble 1981.

/14/ A. Bladon, G. Fant, "A two-formant model and the cardinal vowels", STL-QPSR 1, pp. 1-8, 1978.

/15/ O. Aaltonen, "The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels", Journal of Phonetics, 13, pp. 1-9, 1985.

/16/ L. Chistovich, "Auditory processing of speech", Language and Speech 23(1), 1980.

/17/ K.N. Stevens, "Spectral prominences and phonetic distinctions in language", Speech Communication 4, pp. 137-144, 1985.

/18/ O. Fujimura, "On the second spectral peak of front vowels : a perceptual study of the role of the second and third formants", Language and Speech 10, pp. 181-193, 1967.

/19/ D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra : a first step", Proc. IEEE, pp. 1278-1281, 1982.

/20/ S. Hawkins, K.N. Stevens, "Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels", JASA 77(4), pp. 1560-1575, 1985.

/21/ D. Béroule, J.L. Schwartz, "Essai de formalisation de faits et hypothèses de physiologie concernant le traitement de l'information pour la Reconnaissance Automatique de la Parole", présenté à ce congrès.

MODELISATION DU TRAITEMENT D'INFORMATION OPERE PAR LES PREMIERES COUCHES DE NEURONES DU SYSTEME AUDITIF

Thierry HERVE et Jean-Marc DOLMAZON

Institut de la Communication Parlee de GRENOBLE
I.N.P.G
46 av. Felix Viallet 38000 GRENOBLE CEDEX

ABSTRACT

We present a modelisation of a temporal neural net excited by a spatial stochastic point process; This process is obtained by a temporal sampling of stationnary renewal point processes, which describe the activity of several Auditory Nerve fibers responding to vowels. The activity of the net has been analysed like a Random Field one, so that we may apply the **GIBBS Measure** theory to characterise it. Our first results show that we can estimate the GIBBS Measure with only a few parameters and that the Field reaches within 50ms a stable state.

I INTRODUCTION

Les travaux sur le traitement de la Parole dans le Système Auditif humain commencent au niveau de la connaissance du prétraitement effectué par l'appareil auditif périphérique et se poursuivent par l'approche du traitement effectué dans les centres supérieurs; dans le cadre des travaux menés sur l'analyse du fonctionnement du Système Auditif, nous nous intéressons à la modélisation des premières couches de neurones du Cortex Auditif (celles qui reçoivent l'information directement issue ded cellules sensorielles).
Les tissus corticaux sont probablement très semblables d'une zone à l'autre du Cortex ; aussi utiliserons nous, pour définir notre modèle, les connaissances utilisées dans les modèles du Système Visuel (en effet,l'expérimentation électro-physiologique est plus accessible dans de domaine qu'elle ne l'est dans le Système Auditif).
Domaine majeur, en matière de simulation de réseaux de neurones, l'étude du Système Visuel a conduit nombre de Chercheurs à focaliser leur attention sur le problème de la Reconnaissance des Formes. Les machines neuromimétiques actuellement proposées sont, en général, des réseaux adaptatifs (1) (2), c'est à dire qu'ils règlent leur propre connectivité (efficacités synaptiques) au cours de leur évolution en fonction des circonstances extérieures (excitations) ; la plus part du temps cette régulation est non supervisée (auto-adaptation), c'est à dire que les efficacités synaptiques se modifient en fonction d'une règle d'interaction locale entre neurones (ou loi de plasticité synaptique), comme celle bien connue de HEBB (3) sans intervention d'un organne central de réglage.

Il convient de noter que ces systèmes fonctionnent en deux temps:

a) APPRENTISSAGE

On présente au réseau une série de Formes, dans un ordre quelconque jusqu'à ce que sa connectivité se stabilise ; On considère alors que la connectivité obtenue reflète la partition de l'ensemble des Formes en classes statistiquement indépendantes ;

b) RECONNAISSANCE

En réponse à une Forme quelconque le réseau active un représentant de l'une des classes qu'il a précédemment établies.

Les principales activités de recherche portent actuellement sur la modélisation des mémoires dites "associatives" (reconnaissance d'une image à partir d'une excitation clé - fragment de l'image en question-), l'extraction et la classification de traits pertinents (sur la base d'expériences sensorielles), la mémorisation de séquences temporelles de Formes, la Reconnaissance de Formes indépendamment de leur position spatiale.

Sur ces différents travaux nous remarquerons deux aspects importants:

a) Les Formes qui sont traitées n'ont pas de structure temporelle. Ce sont des images fixes dont on étudie une collection.

b) Ces Formes, en outre, sont significatives ; c'est à dire qu'elles relèvent d'un niveau conceptuel assez élevé ; nous dirons qu'elles appartiennent à un niveau cognitif.

Pour une transposition au signal de parole, deux différences majeures interviennent :
. la caractéristique essentielle du signal de parole est son aspect temporel, mis en evidence dans différentes expériences psycho-acoustiques (citons par exemple le rôle perceptif de la vitesse d'évolution d'un formant) ;
. nous souhaitons examiner le comportement du réseau pendant la durée d'un son (voyelle, par

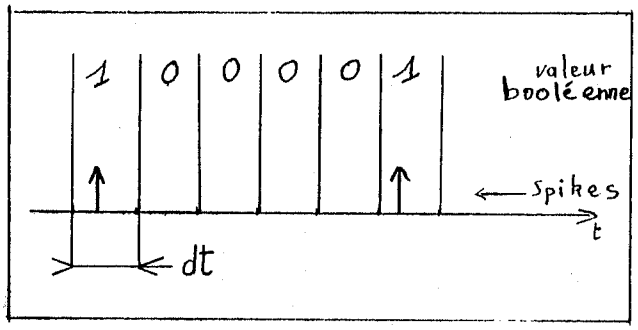
exemple). Les évènements excitant le réseau pendant ce temps ne seront donc pas indépendants mais bien la réalisation d'un même phénomène; à ce titre l'expérience de T. KOHONEN sur les voyelles du Finlandais (4) permet le classement des voyelles, mais il faut noter que ce réseau se nourrit d'une information spectrale statique sur chaque voyelle; il a donc d'emblée accès à une notion abstraite : **le Formant**. Nous souhaitons, au contraire, rester à un niveau proche du signal, en aval du Cortex, ce qui nous interdit ce type de démarche.
Nous présentons ici les résultats de nos premiers travaux. Ceux-ci concernent essentiellement la mise en place de notre modèle et une première estimation des résultats qu'il permet d'obtenir. Nous avons travaillé sur des densités d'intervalles inter-spikes artificielles, attendant de mieux connaitre nôtre outil avant de le connecter de façon effective à un modèle du système auditif.

II. DESCRIPTION DU MODELE

1. Excitation du réseau

Nous savons (5),(6),(7) que l'activité des fibres du nerf auditif en réponse à des voyelles stationnaires peut être modélisée par des processus ponctuels stationnaires de renouvellement. Prenant comme mesure de référence l'intervalle de temps séparant deux spikes sur une fibre, nous pouvons modéliser l'activité de la fibre par des réalisations aléatoires de ces intervalles. Ceux-ci suivront une densité de probabilité (dite de renouvellement) qui possède des propriétés caractéristiques de la voyelle d'excitation et de la Fréquence Caractéristique de la fibre considérée.

Nous pouvons discrétiser l'activité d'une fibre de façon à observer des tranches de temps n.dt ; Chaque tranche n.dt sera affectée de la valeur 0 si aucun spike n'a été émis durant l'intervalle (n.dt,(n+1).dt) et de la valeur 1 si un spike a été généré. Nous avons ainsi construit un processus de renouvellement alterné de 0 et de 1.



Pour décrire l'activité de fibres du nerf auditif en provenance de différents points de la membrane basilaire, nous répéterons cette opération avec différentes densités de probabilité liées à d'autres Fréquences Caractèristiques réparties le long de la membrane basilaire. Nous affectons alors à chaque fibre des coordonnées (I,J) qui correspondront à la répartition spatiale des fibres .

Nous pouvons alors, pour chaque instant n.dt fabriquer une matrice carrée binaire ; dans notre simulation (avec dt=0,1ms) nous fabriquons 2.000 matrices carrées (8x8) qui décrivent l'activité de 64 fibres pendant 200 ms (soit une durée légèrement supérieure à celle d'une voyelle).

La succession des 2.000 matrices binaires représente les réalisations d'un processus spatial dit "Champ aléatoire". Ce Champ est entièrement défini par sa **Mesure de GIBBS** (8).

2. Modèle de réseau

Le réseau sera composé de NxN neurones répartis sur une matrice binaire NxN ;
chaque neurone (I,J) possède :
- 1 sortie $Y_i(t)$ binaire;
- 1 entrée externe $X_i(n.(dt-1))$ recevant l'excitation binaire en provenance de la fibre (I,J) à chaque n.dt.
- 4 entrées latérales excitatrices $Y_{ij}(n.(dt+1))$: le neurone reçoit ainsi les excitations de ces plus proches voisins.
- 8 entrées latérales inhibitrices $Y_{ij}(n.(dt-1))$ provenant des neurones proches autres que les plus proches voisins.

Ce type d'excitations latérales reflète la connectivité des tissus neuronaux et nous nous réservons la possibilité d'en changer les caractéristiques topologiques. Chaque entrée est affectée d'un poids synaptique variable et la fonction neurone est binaire du type "seuil et saturation" .
Les équations de fonctionnement sont les suivantes:

a) fonction du NEURONE i :

$$Y_i(t+1) = U\Big(a.X(t) + \sum_j w_{ij}.Y_j(t) - S_i(t) \Big)$$

a et les w_{ij} sont les efficacités synaptiques de chaque liaison (i,j) (de **norme 1**, positives pour les excitations, négatives pour les inhibitions).
U est la fonction de Heavyside
$S_i(t)$ est le seuil du neurone i; il varit en fonction des dernières excitations/inhibitions reçues; il permet en quelque sorte de mémoriser l'activité pré-synaptique.

$$S_i(t+1) = S_i(t) + 0{,}25 \sum_j w_{ij}.Y_j(t) \ ;$$

$S_i(t+1) = S_0$ si $Y_i(t) = 1$
(décharge du neurone i à l'instant t);

S_i est borné sur l'intervalle $(S_{min}\ ,\ S_{max})$.

On remarque dans ces équations que S_i intègre une information temporelle sur les excitations, et Y_i intègre l'information de S_i avec en plus une information spatiale;

b) plasticité synaptique :

Nous avons pris une loi d'évolution rapide (9) qui permet une mise à jour des efficacités toutes les 3 ms :

$w_{ij}(t+1) = w_{ij}(t) - Epsilon.w_{ij}(t)$:

si l'excitation présynatique a été inefficace $w_{ij}(t+1)$ décroit;

$w_{ij}(t+1) = w_{ij}(t) + Nfois.Epsilon.w_{ij}(t)$:

si l'excitation présynaptique a été efficace w_{ij} croît; le facteur Nfois est un paramêtre qui permet de renforcer les "événements" efficaces.

$w_{ij} = w_{ij}(t)$:

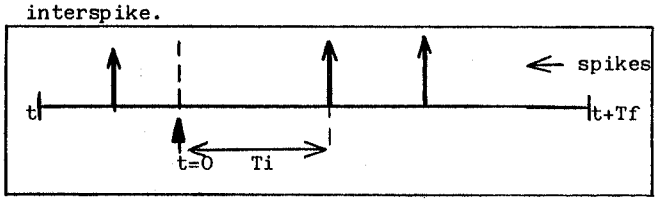
si t est dans l'intervalle (t_0, t_0+3ms), la connexion n'évolue pas (t_0 est multiple de 3ms).

III CARACTERISATION DE L'ACTIVITE DU RESEAU

1. Stationnarité des trains de spikes

D'un point de vue théorique si f(t) est la densité de renouvellement des intervalles inter-spike (de moyenne M), alors pour que le train de spike soit stationnaire, il faut et il suffit que la loi de la durée d'apparition du premier spike depuis le temps 0 soit décrite par la densité (1- F(t))/M (où F(t) est la fonction de répartion de f(t))

Physiquement nous avons interprété cette propriété comme suit : supposons que le processus stationnaire se déroule depuis un temps infini; un observateur définissant arbitrairement un instant 0 dans ce processus positionnerait cet instant de manière uniformément aléatoire dans un intervalle interspike.



Pour vérifier ce raisonnement nous avons mesuré de nombreuses réalisations de Ti par un tirage uniformément aléatoire sur la durée Tf (ces mesures ont été réalisées sur 10 densités quelconques de moyennes et écart-types différents) avec Tf = 1 s ; Nombre de tirages : 1500 . Nous avons alors montré que la densité de probabilité d'apparition de la durée T_i était une bonne estimation de la densité théorique (1-F(t))/M.
Nous avons donc retenu cette méthode pour la simulation d'un train de spikes stationnaire : nous générons un train de spike de 1 s et nous tirons au hasard l'instant 0 sur cette durée.

2 Mesure de GIBBS

Soit K une partie finie (ou boite finie) du plan Z^2, lorsque le processus spatial se déroule, nous obtenons une série de matrices qui représentent l'état de la boite. Les éléments (points (I,J)) de la boite qui sont à "1" forment une partie A; nous confondrons désormais A avec la matrice dont les éléments de A, et eux seuls, sont à l'état "1" (nous parlerons ainsi de la matrice A).
Un potentiel V est une fonction de la boite K vers l'ensemble des réels telle que: $V(\phi)=0$.

On définit l'état de GIBBS PI associé au potentiel V par:

$PI(A) = Z^{-1} \exp(V(A))$ où Z est une constante de normalisation;

si V est un potentiel alors on définit le potentiel d'interaction Jv par :

$$Jv(A) = \sum_{B \subset A} -1^{|A/B|} V(B), \text{ pour tout A de K}$$

(ou $V(A) = \sum_{B \subset A} Jv(B)$);

N.B. |A/B| = cardinal de l'ensemble A moins B.

La mesure de GIBBS est entièrement définie si l'on connait tous les potentiels d'interaction Jv. Nous donnerons une approche plus physique de cette mesure dans la suite du texte.

N.B. Dans notre cas les 64 processus de renouvellement étant générés de manière indépendante, le Champ aléatoire ainsi fabriqué est dit trivial, car il est exempt d'interactions spatiales locales (autrement dit les potentiels d'interactions seront faibles).

3. Paramètres du réseau

Comme nous l'avons dit, les potentiels Jv(A) caractérisent entièrement la mesure de GIBBS ; il y a C_N^n parties A de cardinal n, soit M=NxN=64 :

```
    64 Singletons ; Card (A) = 1 ;
  2016 Paires A ; Card (A) = 2 ;
 41664 Triplets A ; Card (A) = 3 ;
635 376 Quadruplets A; Card (A) = 4 ;
```

Ces nombres montrent que nous risquons d'être submergé par le mombre de potentiels Jv à calculer; Notre but, dès lors, est d'estimer le nombre de Jv significativement non négligeables. Etant donné que la valeur de Jv(A) est directement lié à la probabilité d'apparition de A , nous avons cherché à régler les paramètres du réseau de telle sorte que, statistiquement, CARD(A) soit inférieur ou égal à 2.
Ceci garantit alors que les potentiels de singletons et de paires décriront complètement la mesure et réduit d'emblée le nombre de potentiels d'interaction à calculer.

IV R E S U L T A T S :

Les simulations ont été réalisées à partir de trois densités définies sur l'intervalle (0-8ms)
a) densité uniforme sur l'intervalle (0-8ms) ;
b) densité somme de deux densités centrées respectivement sur 3 et 5 ms, avec chacune un écart-type de 1,5ms ;
c) densité construite identiquement à celle du b) mais avec des écart-types de 0,4ms ;

Pour chaque expérience, l'activité de chaque fibre est simulée à partir de la même densité ; nous obtenons donc 64 trains de spikes issus d'une même loi de probabilité; ceci assure une

homogénéité spatiale de l'excitation;

1. Valeur des paramêtres:

Les paramêtres Epsilon et Nfois décrits au II 2. ont été testés de manière systématique.
Le critère de bon fonctionnement du réseau est l'apparition de matrices de sortie de cardinal inférieur ou égal à 2; en effet, si les matrices de cardinale 3 ou plus sont rares, cela signifie que les potentiels d'interactions associés seront très faibles; nous n'auront donc pas à les calculer.
D'un point de vue physiologique, cela garantie que l'activité des neurones ne sera pas trop élevée.

La valeur de Epsilon optimale oscille autour de 0,05 et celle de Nfois est de l'ordre de 2 à 4.

Nous avons alors généralement 90 % des images dont le cardinal est inférieur ou égal à 2.

2. Potentiels d'interaction:

Dans ces conditions, nous avons calculé les Jv(A) et nous pouvons d'ores et déjà affirmer qu'un nombre assez faible de potentiels suffit à décrire la mesure ; en effet toutes nos simulations ont donné le même type de comportement.

Potentiels d'interaction de Singletons :
Sur 64, seulement 10 environ sont non négligeables.
Potentiel d'interaction de Paires :
Sur 2016, 30 environ sont non négligeables.
potentiels d'interaction de triplets :
ils sont complètement négligeables;

N.B. Le calcul des potentiels Jv pour les processus spatiaux entrée du réseau a montré ~~que tous~~ qu'ils sont tous homogènes et tres faibles comme nous l'annoncions au paragraphe II 2.

CONVERGENCE: Nos simulations montrent que la convergence de la mesure est obtenue au bout de 40ms lorsque les potentiels mesurés sont stabilisés. Cela signifie que dans l'espace des états, la mesure a convergé vers un état en moins de 50ms. Nous n'avons, pour l'heure, pas concentré nos efforts sur les problèmes de la répartion spatiale des potentiels non négligeables; toutefois nous pouvons déja dire que l'hétérogénéïté de cette répartition devra être étudiée car le processus d'excitation, les règles d'interaction latérales sont globalement homogènes;

V C O N C L U S I O N

- Notre simulation de l'activité stationnaires de fibres du nerf auditif s'appuie sur la théorie des processus de renouvellement ponctuels et permet en particulier de respecter le principe de stationnarité tel qu'il est énoncé d'un point de vue très théorique.
- La "fabrication" d'un processus spatial, pour artificielle qu'elle puisse paraitre, permet cependant la mise en place d'un réseau bidimentionnel avec les propriétés d'interaction spatiales connues.
- Notre modèle s'appuie sur le formalisme mathématique des mesures de GIBBS qui fournit un outil puissant d'estimation de l'activité par la mesure d'un nombre peu importants de potentiels d'interaction ; nous espèrons de plus condenser encore ces informations gâce à la combinaisons des informations sur l'amplitude des potentiels et leur localisation spatiale.

- Il est clair qu'un réseau de 64 neurones est très petit, aussi , dès que nous aurons établi complètement les propriétés qui caractérisent ce modèle, nous passerons à des tailles croissantes (128,...1.024) ; le but recherché étant alors, d'assimilér d'un point de vue mathématique la "boite" B à un plan; l'un des problèmes les plus importants sera de tester l'homogénéïté spatiale du processus, en fonction, par exemple, de l'état d'excitation initial de chaque neurone.

REMERCIEMENTS : ce travail a été conduit avec l'aide de Monsieur Jacques DEMONGEOT, Professeur à l'I.M.A.G de GRENOBLE, que nous tenons à remercier vivement.

R E F E R E N C E S :

(1) **K. Fukushima**, "Neocognitron", Biol.Cybern vol 36(4).1980.

(2) **T. Kohonen**, "Self-organisation and associative memory ", Springer Verlag (Berlin, Heidelberg, New York, Toqyo). 1984.
(3) D.O. Hebb, "the organisation of behavior" (Wiley). 1949.

(4) **T. Kohonen**, "Clustering, and topological maps of patterns", Proc. IEEE 82, october 1982.

(5) **B. Delgutte**, "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers", J. Acoust. Soc Am. 68, 843-857. 1980.

(6) **M.B. Sachs, E.D. Young**, "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate", J. Acoust. Soc. Am. 66, 470-479, 1979.

(7)**D.H. Johnson**, "The relationship between spike rate and synchrony in the response of auditory-nerve fibers to single tones", J. Acoust. Soc. Am., 68, 1115-1122, 1980.

(8) **J. Fricot**, "Champs aléatoires de renouvellement", Thèse D.I, IMAG Grenoble, 1985.

(9) **C. Von der Malsburg**, E. Bienenstock, "Statistical coding and short-term synaptique plasticité: a scheme for knowledge representation in the brain", to appear in "Disorded systems and biological organisation", Springer Verlag, (Berlin).

MODELISATION ACTIVE DE L'ENSEMBLE COCHLEAIRE : UNE NOUVELLE APPROCHE DES NON LINEARITES DE FONCTIONNEMENT

J.C. CAEROU *, J.M. DOLMAZON *, V.S. SHUPLJAKOV **

* I.C.P. Institut de la COMMUNICATION PARLEE, 46, Avenue Félix Viallet 38031 GRENOBLE CEDEX

** Institut de Physiologie PAVLOV, Acad. Sci.de l'U.R.S.S. LENINGRAD (U.R.S.S.)

ABSTRACT

This paper describes a new model of ear which simulates the non-linear behavior of the peripheral auditory system for real input sounds. The main assuption undelying the structure of the model is the presence of an active mechanism which works mainly near the resonance. The general structure of the model looks like classical lumped transmission line. Each cell has its own frequency resonance according to the coordinate-frequency correspondance observed along the basilar membrane. A non-linear local feedback from the resistive element, is mixed to the input signal in each cell. Such a circuit exibit interesting results concerning, mainly, the evolution of selectivity with respect to the input signal level, two-tone suppression effects and combination tones generation.

I-INTRODUCTION

Si le traitement du message acoustique dans les centres auditifs supérieurs est encore très mal connu, on commence, en revanche, à expliquer de façon beaucoup plus précise le fonctionnement du système auditif périphérique. Depuis quelques années, l'utilisation de nouvelles techniques a permis une progression rapide des connaissances dans ce domaine. De nombreuses données expérimentales fiables sont aujourd'hui disponibles et leur analyse détaillée amène des précisions intéressantes sur le rôle de telle ou telle structure.

Ainsi, la sélectivité relevée au niveau de la membrane basilaire semble aujourd'hui du même ordre de grandeur que celle observée dans les fibres isolées du nerf auditif (RHODES 1978, KHANNA et LEONARD 1982, etc.). Compte tenu des valeurs observées, il est probable qu'un phénomène actif intervienne dès ce niveau de la chaine auditive. D'autres découvertes récentes conduisent à la même conclusion (émissions auto-acoustiques, suppression à deux tons, produits de composition, etc.). De plus, contrairement à ce qu'on pourrait attendre, ces non-linéarités de fonctionnement interviendraient, de façon principale, pour les sons de très bas niveau et une approximation linéaire du fonctionnement ne serait pas possible (la sélectivité est meilleure pour les sons de faible intensité (RHODES)), la suppression à deux tons est un phénomène probablement présent en dessous du seuil de perception (JAVEL 1981, etc.).

Pour rendre compte de ce comportement, nous avons été amenés à développer un nouveau modèle de fonctionnement du système auditif périphérique. Ce modèle incorpore, dans son principe, deux caractéristiques fondamentales : il est non linéaire et il possède une boucle de rétroaction positive qui lui confère des propriétés très semblables à celles observées dans le système auditif.

II-DESCRIPTION DE LA CELLULE DE BASE

La membrane basilaire est simulée par des éléments à constantes localisées. Leur association forme une suite de N cellules de base. Comme pour les modèles développés dans le passé (PETERSON & BOGERT, HALL, DOLMAZON, SCHROEDER, etc.), l'ossature de base de chaque cellule est constituée d'un circuit oscillant RLC. Les caractéristiques de chaque circuit oscillant sont fonction du numéro de cellule. Le schéma ci-dessous représente une cellule de base du modèle.

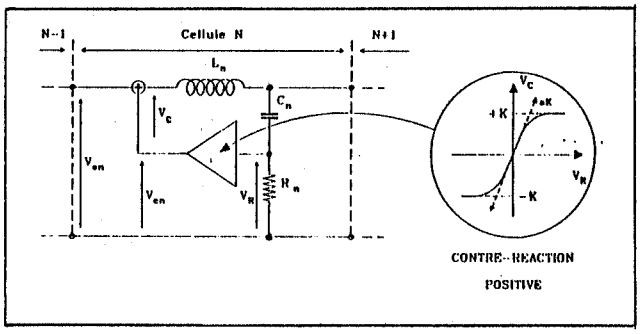


Figure 1 : Schéma de la cellule élémentaire du modèle

Au circuit RLC (passif), on ajoute une boucle de retour qui additionne à la tension d'entrée Ve_n

une tension de réaction Vc_n. Celle-ci est proportionnelle au courant dans le circuit oscillant, elle est prise aux bornes de la résistance et subit une **amplification** puis une **limitation** selon la loi :

Vc=K*th(alpha*Vr).

Où K*alpha représente le gain de ce circuit, th(alpha*Vr) la fonction de non linéarité et K la limitation en tension.

Si l'on étudie, en fonction de la fréquence pour un niveau d'excitation donné, le module de l'impédance d'entrée du circuit oscillant non bouclé, on constate qu'il passe par un minimum égal à R pour la fréquence de résonance. Le courant parcourant la résistance suit donc une loi inverse : il passe par un maximum à cette fréquence. Pour des fréquences éloignées de la fréquence de résonance du circuit, le courant dans la résistance est très faible: la tension de réaction est donc négligeable par rapport à la tension d'entrée : tout se passe comme s'il n'y avait pas de rétroaction. Quelle que soit l'entrée le comportement de la cellule sera **linéaire**.

Au fur et à mesure que l'on se rapproche de la fréquence de résonance, le courant dans la résistance croît, la tension Vr à l'entrée de la boucle de réaction croît dans les mêmes proportions: Pour de faibles niveaux de la tension d'entrée Ve, la tension Vr est également faible. Vc=Kth(alpha*Vr)-->K*alpha*Vr; La boucle de réaction fonctionne dans sa partie linéaire, la tension Vc est donc proportionnelle à Vr. Vc venant s'additionner à Ve, la tension de sortie sera plus importante qu'en absence de circuit de réaction. La **sélectivité** de la cellule est donc **augmentée**. Les paramètres influents sont K et alpha. Un accroissement du produit K*alpha peut provoquer la mise en **oscillation** de la cellule et donc la **génération spontanée** de signal à la sortie de la cellule.

Pour un Ve plus important, le circuit de réaction devient de plus en plus non-linéaire (Vc se rapprochant de la valeur K), entrainant l'apparition de **produits de composition.**

Enfin pour des niveaux d'entrée encore plus élevés, la tension Vc sortant du circuit de réaction est constamment limitée à +/-K. La contribution du signal Vc décroît (Vc<<Ve). La cellule retrouve alors un comportement analogue à celui où la contribution du circuit de réaction est négligeable.

Le modèle décrit ci-dessus peut être cablé puis étudié physiquement. Cette façon de procéder est peu pratique pour ajuster de façon précise les paramètres. Nous avons choisi la voie de la simulation sur ordinateur (au détriment de la vitesse de réponse !). La cellule de base du modèle se décrit par une équation différentielle du second ordre à coefficients variables. Pour des signaux d'entrée quelconques la résolution fait appel à des méthodes numériques : nous avons choisi celle de Runge Kutta pour sa simplicité de mise en oeuvre.

Dans le cas de signaux d'entrée sinusoïdaux, nous avons utilisé une méthode de calcul par itération qui donne des résultats plus rapidement.

III-RESULTATS

Une bonne connaissance du comportement de la cellule de base est nécessaire avant l'intégration du modèle complet. C'est pourquoi nous présentons une étude détaillée des principales caractéristiques de la cellule de base.

Pour l'ensemble de ces mesures les paramètres sont les suivants:

F_0=800 Hz; L=0,421 Hy; C=0,094 µF;
R=1000 Ohms; Q=2,116
K variant de 0,3V à 30V et K*alpha de 0 à 0,96

III-1) Réponse en fonction de la fréquence :

Nous avons appliqué à la cellule des signaux sinusoïdaux, variant lentement de 400 à 1600 Hz. Les tracés sont effectués à niveau d'entrée constant pour des valeurs variant de -50dB à +50dB par pas de 10dB (niveau de référence arbitraire).

La figure 2 représente la réponse de la cellule en fonction de la fréquence pour différents niveaux d'entrée.

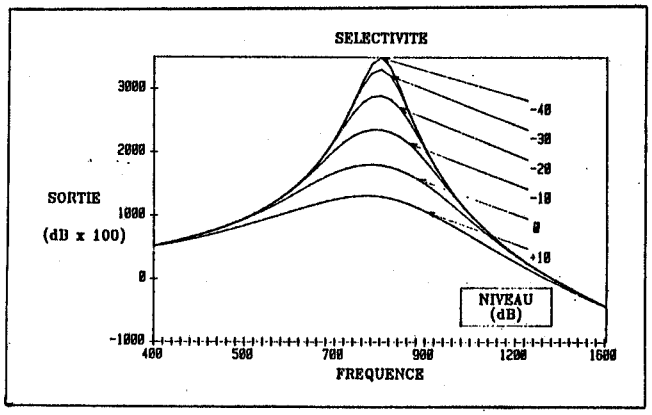


Figure 2: réponse de la cellule en fonction de la fréquence

Pour $F<<F_0$ ou $F>>F_0$ les réponses sont identiques quels que soient les niveaux et K*alpha. La contribution relative de la boucle de rétroaction reste bien négligeable.

Pour les faibles niveaux (en dessous de -40dB) on observe, une sélectivté indépendante du niveau d'entrée, le circuit présente un coefficient de qualité maximum.

Pour les niveaux les plus importants (au dessus de 30 dB), la sélectivité est également indépendante du niveau d'entrée, mais elle est minimale.

Pour des niveaux d'entrée compris entre les valeurs précédentes, et pour des fréquences voisines de F_0, la variation de la réponse est fonction du niveau d'entrée, d'autant plus fortement que F est plus proche de F_0. Plus le niveau d'entrée croît, plus la sélectivité baisse, ce qui est conforme aux mesures observées chez l'animal au niveau de la membrane basilaire (KHANNA 1982) et au niveau des réponses nerveuses (RUSSELL et SELLICK 1978, de BOER 1983, KEMP 1982, etc.).

III-2) Fonction d'entrée sortie :

Les figures 3 et 4 montrent les fonctions d'entrée-sortie obtenues pour $F=F_0$. On observe sur ces tracés les trois zones mentionnées ci-dessus : les deux zones linéaires pour les très faibles et les très forts niveaux d'entrée et une zone de "compression" entre les deux. Le phénomène de compression est d'autant plus fort que le produit K*alpha est proche de 1 (figure 3). La figure 4 montre que le paramètre de saturation K (à produit K*alpha constant) permet de translater la réponse sans la modifier, et donc de maitriser le niveau auquel intervient la compression. Ce niveau semble un paramètre important pour étudier les différentes populations de fibres présentant des fonctions de sensibilité différentes.

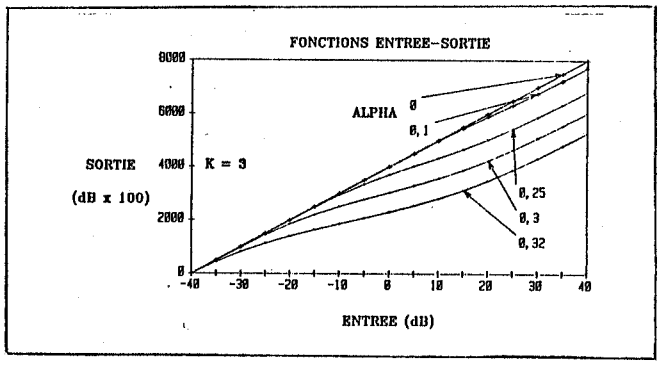


Figure 3 : Fonctions d'entrée-sortie pour K*alpha variable

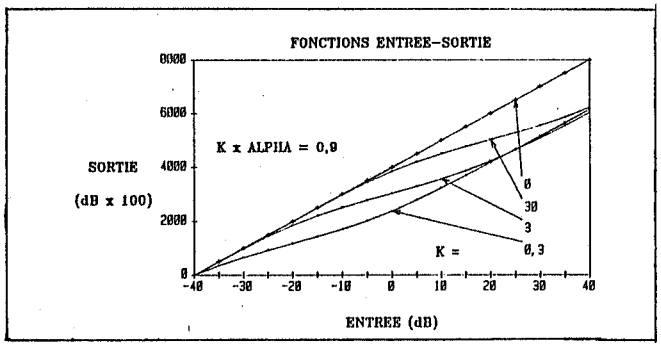


Figure 4 : Fonctions d'entrée-sortie pour K variable (K*alpha constant)

III-3) Réponse à deux fréquences :

Pour un signal incident à deux fréquences la cellule de base donne, dans certaines conditions, des produits de composition significatifs lorsque les niveaux et les fréquences des composantes du signal d'entrée correspondent au régime de fonctionnement non linéaire. Les termes $2F_1-F_2$ et $2F_2-F_1$ sont généralement les plus intenses, mais d'autres composantes peuvent apparaitre. La figure 5 montre (dans sa partie supérieure) l'allure du spectre de la réponse obtenue avec une stimulation bitonale (F_1=800 Hz et F_2=950 Hz). On observe bien la présence de produits de composition de part et d'autre des composantes F_1 et F_2. Les courbes situées dans la partie inférieure de cette figure montrent que l'amplitude des composantes $2F_1-F_2$ et $2F_2-F_1$ passe par un maximum dans la zone d'amplitude où la compression est maximale. Ces produits de combinaison se comportent comme des composantes fréquentielles nouvelles générées par la cellule. De telles composantes ont été observées à différents niveaux du système auditif périphérique (KIM et al, 1980, KEMP 1982).

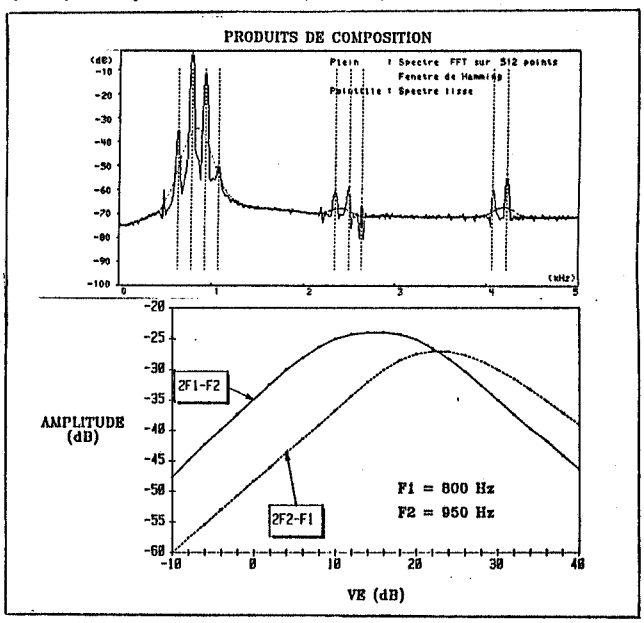


Figure 5 : Evolution des Produits de Composition générés par la cellule de base du modèle

III-4) Réponse impulsionelle :

Une impulsion de tension appliquée à l'entrée de la cellule provoque une oscillation amortie en sortie. L'amortissement de la réponse est d'autant plus faible que la réaction est forte (K*alpha croissant). On peut bien sûr provoquer la mise en oscillation spontanée pour K*alpha > 1.

Le niveau de l'impulsion a également une influence : une forte impulsion provoque un amortissement très rapide des oscillations, tandis qu'une impulsion très faible donne un

amortissement bien plus lent. Ces résultats sont conformes aux mesures de sélectivité : plus la sélectivité est importante et moins amortie est la réponse.

III-5) Suppression à deux tons :

Un calcul théorique montre que lorsque le signal d'entrée croit en amplitude, la fréquence pour laquelle le signal de sortie est maximum se décale vers les fréquences basses. Ce décalage peut atteindre 10 à 15% avec les valeurs des paramètres indiqués au début du paragraphe. Dans ces conditions il est possible qu'un signal sinusoïdal de fréquence F_1 et d'amplitude A_1, convenablement choisis, puisse subir une réduction d'amplitude, en présence d'un second signal F_2 et A_2. Pour une seule cellule cette réduction reste faible (0,2 à 0,3 dB), mais lorsque la rétroaction est "propagée" à d'autres cellules, la suppression est beaucoup plus importante. Cependant, le choix des paramètres de rétroaction nécessite une bonne connaissance de l'ensemble des observations pour d'autres signaux et le travail que nous présentons ici n'est pas assez avancé pour nous permettre actuellement une étude quantitative précise.

IV-CONCLUSIONS

Ces premiers résultats montrent que le comportement de la cellule de base est qualitativement conforme au comportement observé à différents niveaux du système auditif périphérique. Nous mettons en place actuellement la simulation d'un ensemble de cellules, avec pour objectif la description du fonctionnement d'une chaine de cellules pour caractériser le rôle de la boucle de rétroaction. Le "lieu" de bouclage est un paramètre que nous jugeons capital car il conduit à un comportement en fréquence particulier avec notamment dissymétrie des réponses de part et d'autre de F_0, et existence d'un rapport F_1/F_2 optimal pour les produits de composition et la suppression à deux tons. Lorsque les paramètres du modèle auront été ajustés, nous ajouterons un module d'adaptation pour simuler de façon complète l'aspect temporel des informations délivrées par le système auditif périphérique.

BIBLIOGRAPHIE :

ALLEN, J.B. (1985). "Cochlear Modeling", IEEE - ASSP Magazine, 2, 3 -29.

De BOER, E., (1983). "No sharpening ? A Challenge for Cochlear Mechanics.", J. Acoust. Soc. Am., 73, 567-573.

BROWN, A.M., and KEMP, D.T., (1984). "Suppressability of the $2F_1 - F_2$ Stimulated Acoustic Emissions in Gerbil and Man.", Hearing Research, 13, 29-37.

DOLMAZON, J.M., (1982). "Representation of Speech-Like Sounds in the Peripheral Auditory System in Light of a Model", in "the Representation of Speech in the Peripheral Auditory System", CARLSON, et GRANSTROEM Eds., Elsevier - Biomedical Press.

HALL, J.L., (1974). "Two-Tone Distorsion Products in a Non-Linear Model of the Basilar Membrane", J. Acoust. Soc. Am., 56, 1818-1828.

JAVEL, E., (1980). "Suppresion of Auditory Nerve Responses I : Temporal Analysis, Intensity Effects and Suppression Contours", J. Acoust. Soc. Am., 69,1735-1745.

KEMP, D.T., (1982). "Evidence of Mechanical Non Linearity and Frequency Selective Wave Amplification in the Cochlea.", Arch. Oto.-Rhino.-Laryngol. 224, 37-45.

KHANNA, S.M. and LEONARD, D.B.G. (1982). "Basilar Membrane Tuning in the Cat Cochlea.", Science, 215, 305-306.

KIM, D.O., MOLNAR, C.E. and MATHEWS, J.W., (1980) "Cochlear Mechanics: Nonlinear Behavior in Two Tone Responses as Reflected in Cochlear Nerve Fibre Responses and in Ear Canal Sound Pressure", J. Acoust. Soc. Am., 67, 1704-1721.

MOLLER, A.R., (1983). "Auditory Physiology.", Academic Press, New-York.

NELLY, S.T., and KIM, D.O., (1983). "An Active Cochlear Model Showing Sharp Tuning and High Sensitivity.", Hearing Research, 9, 123-130.

NELLY, S.T., (1985). "Micromechanics of the Cochlear Partition.", in 1985 Mechanics of Hearing Workshop.

RHODE, W.S., (1978). "Some observations on cochlear mechanics.", J. Acoust. Soc. Am., 64, 158-176.

RUSSEL, I.J., and SELLICK, P.M., (1978). "Intracellular Studies of Hair Cells in the Mammalian Cochlea", J. Physiol., 284, 261 - 290.

PERCEPTION

Président

François LONCHAMP

Institut de Phonétique de Nancy

MISE EN EVIDENCE DE FONCTIONS D'ANCRAGE ET DE DECLENCHEMENT EN PERCEPTION DE PAROLE

E. LHOTE*, J.M. DIAZ DE LEON**, S. VINTER***, V.E. OMOZUWA*

*Laboratoire de Phonétique de Besançon, **C.E.L.E de México, ***C.H.U. Besançon

ABSTRACT

Using an explicative model of speech recognition and decoding by the hearer, three listening situations where the processes of speech recognition and decoding are naturally retarded are considered in this study : listening to the speech of profound deaf children ; a contrastive phonemic perception task involving a group of Spanish mexicans learning french ; and the tonal perception of bisyllabic nouns in ẹ̀dǫ́ by eight native speakers of five different linguistic groups.

Certain zones of maximal density favoring the perception, comprehension, and language learning processes emerged from this study.

INTRODUCTION

Le travail présenté ici peut être considéré comme une contribution au traitement de la parole qu'effectue l'auditeur lors du processus d'écoute et plus précisément dans le passage des aspects perceptifs aux aspects cognitifs.

Nous commençons par différencier l'écoute et la perception de parole : percevoir la parole c'est recevoir et traiter un signal acoustique puis transformer celui-ci en un message à caractère linguistique et significatif. L'acte d'écoute ajoute à celui de percevoir une participation active de l'auditeur, une mise en situation du processus de traitement et une recherche de compréhension. Cette distinction très importante est nécessitée par le caractère finalisé de nos recherches qui visent soit à faire progresser la compréhension dans une langue étrangère, soit à améliorer la parole de l'enfant sourd. Ces objectifs peuvent sembler éloignés de ceux des acousticiens qui se consacrent à la reconnaissance de la parole. Or nous considérons quant à nous que les recherches qui sont très avancées dans la connaissance et la reconnaissance du signal acoustique ont besoin maintenant d'un changement de problématique linguistique, afin de mieux tirer profit du modèle humain dans l'acte de perception-compréhension de parole. Et inversement, les travaux que nous effectuons en linguistique appliquée et dont l'objectif principal est l'oral des langues et du langage ne peuvent plus ignorer les apports importants de la reconnaissance de parole et de l'intelligence artificielle. Haton (1) fait remarquer que le passage de la reconnaissance de mots isolés à celle de la parole continue constitue un "important changement de problématique". En choisissant de travailler sur l'écoute, nous changeons délibérément de problématique phonétique.

Les travaux effectués en perception de parole ont mis en évidence des résultats négatifs importants : d'une part l'invariance linguistique ne peut pas se trouver dans le signal acoustique ; d'autre part il n'existe pas de relation biunivoque entre l'information fournie par l'analyse linguistique et celle fournie par le signal acoustique (2). On comprend alors pourquoi Rossi (3), constatant que "l'identification acoustique est, d'un certain point de vue, meilleure que la performance de l'oreille", éprouve le besoin de rappeler la performance de l'auditeur humain qui, "pour compenser l'information somme toute grossière fournie par l'oreille, a recours aux contraintes phonologiques associées à des structures de haut niveau, pour l'essentiel non marquées acoustiquement".

La confrontation quotidienne aux difficultés de production et de compréhension éprouvées lors de l'utilisation d'une langue étrangère par des locuteurs d'origines diverses, nous a conduits à faire des constats qui remettent en cause des idées reçues et à émettre des hypothèses de travail :

1. une langue façonne l'écoute et la compréhension : on n'écoute pas en anglais comme on écoute en français et vice versa ;

2. les difficultés d'apprentissage varient avec la langue d'origine et avec la langue-cible ;

3. les études linguistiques et phonologiques contrastives permettent de comparer les systèmes d'encodage et de représentation, mais se révèlent peu efficaces pour expliquer pourquoi certaines différences résistent à l'apprentissage et pourquoi d'autres s'acquièrent facilement. Ces hypothèses constituent en fait des objectifs à long terme pour un groupe de travail plurilingue et pluridisciplinaire.

Nous allons dans le présent travail essayer d'apporter une contribution à l'hypothèse suivante : dans toute écoute de parole, il est des chemins obligatoires et des priorités à respecter ; comment changent les priorités d'écoute quand le langage est altéré par la surdité, quand on change la langue-cible et quand on change la langue maternelle ? En jouant soit sur une écoute ralentie, soit sur les contrastes linguistiques nous essayons de faire émerger des invariants de l'écoute et des spécificités liées à une langue.

Nous faisons appel à un modele de production-perception (4) que nous avons mis au point afin de tester l'efficacité de trois fonctions qui se dégageaient des observations antérieures et nous étudions plus particulièrement les fonctions d'ancrage et de déclenchement. Les situations retenues sont de trois types : la première est une situation de compréhension, la seconde fait appel à la perception phonémique et la troisième est empruntée à la perception tonale dans une langue tonale, l'ẹdó.

1. Le modèle de traitement utilisé.

Ce modèle vise à intégrer dans une représentation visuelle aussi simple que possible les aspects complexes qui contribuent au processus de perception-compréhension de parole. Etant donné la complexité du processus et la difficulté d'en donner une représentation globale à la fois efficace et correcte, nous nous sommes limités aux trois fonctions principales qui se dégagent de travaux antérieurs (4,5) et de ceux de Rossi (3) et Noizet (6). L'utilisation permanente d'un mode parallèle et d'un mode séquentiel, dégagée pour toute perception par Noizet (6), a conduit Lhote à mettre en évidence l'importance de la fonction de déclenchement (5) qui est certainement une spécificité du traitement cognitif.

Ces trois fonctions s'exercent selon des modes temporels différents :

- la fonction d'ancrage nécessite de la durée et correspond à une sorte de stationnement nécessaire à la fixation de formes, à leur stockage et rend l'observateur plus disponible pour explorer et analyser ;

- la fonction de repérage est mise en oeuvre dans les stratégies utilisées par l'auditeur pour traiter, décoder et sélectionner des hypothèses :

- la fonction de déclenchement s'exerce de façon quasi instantanée : la rencontre des deux forces d'ancrage et de repérage donne naissance à une force résultante qui déclenche la compréhension (fig. 1)

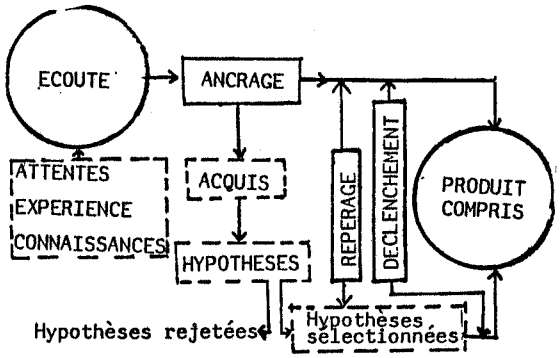


Figure 1: Le système de traitement de la parole par l'auditeur;les fonctions d'ancrage,de repérage et de déclenchement.

2. Expérience 1 : situation de compréhension

L'étude porte sur la compréhension de la parole d'enfants sourds par des phonéticiens et des orthophonistes.

2.1. Le matériel expérimental : 7 phrases ont été prononcées par 6 enfants sourds profonds de 10 à 12 ans appareillés, soit en lecture, soit en commentaire d'images. Les phrases ont été ensuite sélectionnées en fonction du critère de compréhension : ont été écartées celles qui ont été comprises et celles qui ont été incomprises par tous les auditeurs ; n'ont été conservées pour l'analyse que les phrases comprises par les spécialistes mais pas comprises d'emblée par les phonéticiens, de façon à observer au ralenti les processus de traitement.

2.2. Les auditeurs : 5 chercheurs phonéticiens dont 3 étrangers, 3 enseignants de l'université 3 étudiants étrangers.

2.3. Les conditions d'écoute et les tâches : chaque groupe d'auditeurs a été réuni en un même lieu et a subi ensemble l'épreuve. Tous ont eu à effectuer d'abord une tâche d'écoute individuelle consistant à consigner par écrit d'une part ce qu'ils entendaient, d'autre part ce qu'ils comprenaient (chaque phrase a fait l'objet de plusieurs écoutes), puis une écoute collective. Le contrôle de la situation d'expérience a été assuré par un enregistrement magnétoscopé, la présence de quatre observateurs et par une analyse acoustique des productions.

Pour chaque phrase, l'observation a consisté à dégager les effets et les produits de chaque fonction, à enregistrer la chronologie des faits chez chacun et à mettre en évidence les stratégies de compréhension.

2.4. Exemple analysé : "Maman boit du coca" (phrase reconstituée par le groupe de 5 phonéticiens).

- Après la première écoute :
  Produit de la fonction d'ancrage : maman
  Après repérage : _hypothèse perceptive
  mamamwatokoke
  _hypothèse linguistique :
  *maman moi ton coquin

  Pas de déclenchement, pas d'accès à la compréhension.

- Après la deuxième écoute :
  Deuxième point d'ancrage : coca

  Après repérage : _hypothèse rejetée : *ton coquin
  _Nouvelles hypothèses :
  - perceptive : (mamãmwa-koka)
  - linguistique : *maman moi-coca

  Réorientation avec marche arrière :

  maman (mwa) coca
  boit

  Déclenchement et compréhension :

  maman boit du coca

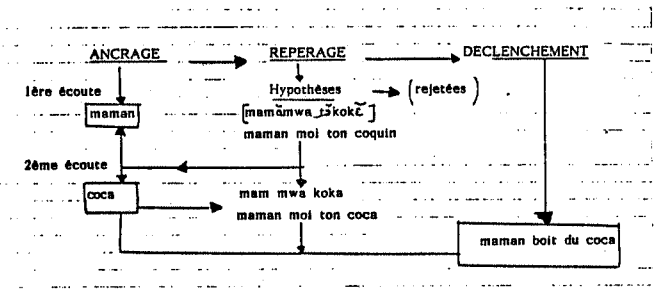


Figure 2 : Représentation du traitement de la phrase "Maman boit du coca", prononcée par un enfant sourd et décodée par des phonéticiens.

La spécificité de cette situation réside dans le retour aux premiers acquis de la fonction d'ancrage et à la réorientation effectuée avec le changement d'hypothèses.

Cette analyse telle qu'elle est présentée ne rend pas compte de l'ensemble des paramètres étudiés :

- l'effet du groupe, de la langue, de la connaissance de la parole d'enfant sourd ;
- le poids de la structure syntaxique, syllabique sur le décodage ;
- la fréquence d'occurrence des unités lexicales dans la langue ;
- l'effet de l'écoute collective sur l'écoute individuelle ;
- les stratégies individuelles de chaque auditeur par rapport à la perception phonétique et à la compréhension.

Nous ne signalerons ici que ce qui émerge et nous paraît fructueux pour tous ceux qui s'intéressent à la reconnaissance de parole :

1. la fonction d'ancrage dégage le plus souvent des unités lexicales facilement accessibles (en termes de structure phonétique), le plus souvent dissyllabiques, qui sont souvent sous l'accent ; quand elles ne sont pas sous l'accent, l'analyse acoustique montre qu'il s'agit de productions accentuellement déformées.

2. L'efficacité dans le traitement est liée étroitement aux stratégies de déclenchement : un auditeur souple dans ses critères d'acceptation (un étranger a tendance à être plus souple qu'un francophone dans cette tâche d'écoute de parole d'enfant français sourd) utilisera mieux les acquis de l'ancrage pour parvenir à la compréhension.

3. Expérience 2 : perception phonémique

Connaissant les difficultés qui existent entre les deux systèmes phonétiques de l'espagnol et du français, nous avons voulu tester et comparer la perception de phonèmes français absents dans le système de l'espagnol. Ces phonèmes sont les suivants: |ã, ɛ̃, ɔ̃, y, ø| et |v, z, ʃ, ʒ|

3.1. Le matériau expérimental : nous avons choisi des mots monosyllabiques du français à structure simple CV en veillant d'une part à opposer les phonèmes nouveaux à des phonèmes espagnols voisins et d'autre part à n'introduire qu'un phonème nouveau par syllabe.

3.2. Les auditeurs et les tâches : tous les auditeurs ont subi en laboratoire de langues un test de discrimination et un test d'identification. La consigne consistait à mettre une croix dans des grilles. Dans le premier cas, on demandait aux sujets d'indiquer seulement si deux sons étaient identiques ou non, afin de tester leur aptitude à distinguer au sein de deux séquences CV si un son différait d'une séquence à l'autre. Dans le second cas, l'auditeur devait choisir, parmi plusieurs phonèmes, celui qu'il identifiait.

3.3. Les résultats, résumés dans la figure 3, font apparaître une grande différence de scores selon la tâche et selon le groupe de phonèmes. On peut en déduire que les problèmes perceptifs sont plus liés à la tâche d'identification qu'à celle de discrimination.

Les résultats obtenus par l'ensemble des auditeurs dans l'identification de chaque phonème montrent que certains phonèmes résistent plus à la perception que d'autres (figure 4).

La comparaison des résultats selon le niveau d'apprentissage permet de dégager les points suivants :

- certains phonèmes, absents du système de l'espagnol, constituent des difficultés perceptives maximales, car ils résistent à l'apprentissage ; d'autres perturbent beaucoup moins l'apprenant hispanophone ;
- les phonèmes à difficulté perceptive maximale

sont les voyelles /ɛ̃/ et /ɑ̃/ ;
- parmi les phonèmes nouveaux qui perturbent moins la perception phonémique, on remarque la présence de la voyelle nasale /ɔ̃/.

Ces résultats permettent déjà de hiérarchiser des difficultés phonémiques pour un groupe linguistique donné.

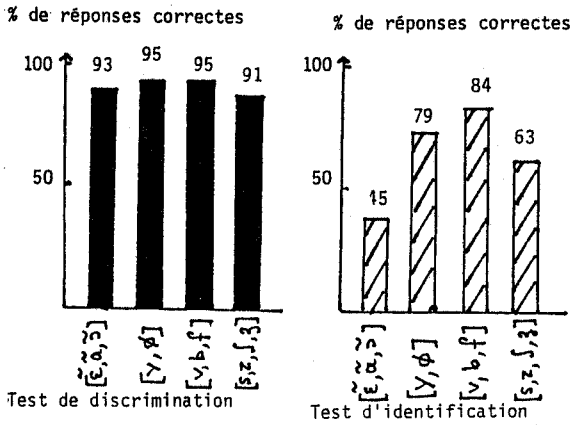


Fig.3: Pourcentage de réponses correctes dans le test de discrimination et dans le test d'identification.

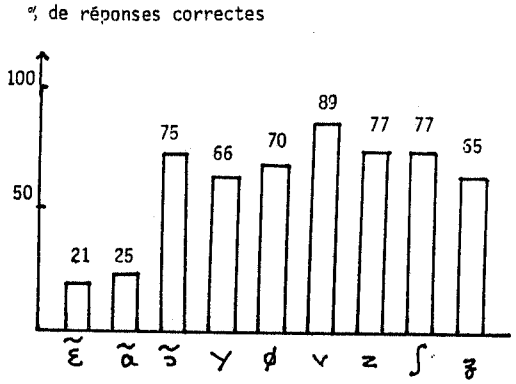


Fig.4 : Résultats obtenus par l'ensemble des auditeurs dans l'identification des phonèmes.

4. Expérience 3 : Analyse de perception tonale

L'objectif de cette expérience est la détermination du (ou des) paramètre(s) acoustique(s) qui, parmi ceux utilisés par les locuteurs, sont indispensables pour l'auditeur ẹ̀dó, dans la perception tonale de mots dissyllabiques du type V_1 CV_2 de la langue.

L'ẹ̀dó est une langue tonale parlée dans l'état de Bendel au sud du Nigéria ; elle comporte deux tons de base, Haut et Bas, phonologiquement pertinents dans des substantifs dissyllabiques ou polysyllabiques de la langue. Ces deux tons sont combinés ici selon les séquences $\acute{V}_1$ $C\acute{V}_2$, $\grave{V}_1$ $C\grave{V}_2$, $\acute{V}_1$ $C\grave{V}_2$ et $\grave{V}_1$ $C\acute{V}_2$ dans des mots dissyllabiques.

4.1. Procédure expérimentale

L'expérience a été menée sur un corpus de 7 mots pour chacune des 4 structures, ce qui représente un ensemble de 28 items prononcés par un locuteur ẹ̀dó et enregistrés. Ces huit auditeurs de cinq groupes linguistiques différents (Yoruba, Ika, Hausa, Français, Chinois) ont écouté puis répété les 28 items ; les 228 stimuli obtenus ont été analysés acoustiquement à l'aide de l'analyseur de mélodie PM 100.

Les 228 stimuli ont été présentés à deux auditeurs natifs ẹ̀dó qui avaient pour tâche de donner par écrit le sens exact de ce qu'ils entendaient : il s'agit là d'un exercice de reconnaissance à caractère linguistique qui s'appuie sur un processus cognitif.

Certains items ont été acceptés et d'autres écartés par les auditeurs ẹ̀dó. Nous avons alors comparé les formes acoustiques des items acceptés et des items rejetés chez chaque locuteur.

4.2. Résultats.

Notre analyse acoustique montre que la variation de F_0 est le paramètre acoustique le plus important dans la discrimination fonctionnelle des tons de l'ẹ̀dó.

Cette variation peut se manifester de façon différente selon la structure tonale considérée ainsi dans la séquence $\acute{V}_1$ $C\acute{V}_2$, un ton Haut est perçu Haut seulement si V_1 et V_2 sont réalisées au même niveau de hauteur et ceci indépendamment de la durée et de l'intensité de V_1 et V_2 ; un écart de 1/4 de ton (ou plus) entre F_0 final de V_1 et F_0 initial de V_2 entraînera automatiquement la perception de cette séquence comme $\grave{V}_1$ $C\acute{V}_2$ même si V_1 et V_2 sont réalisées avec une courbe plate ; V_1 et V_2 pourraient être réalisées avec une courbe montante si F_0 initial de V_1 ne dépasse pas d'un ton F_0 initial de V_2. Le ton de V_1 sera perçu comme un ton montant si cette différence fatidique est supérieure à 1 ton. Le ton bas sur V_2 dans la séquence $\acute{V}_1$-$C\grave{V}_2$ est perçu comme phonologiquement bas à condition qu'il soit réalisé avec une pente descendante du fait de l'assimilation du ton bas dans cette position au même niveau ou à un niveau légèrement supérieur à celui du ton haut sur la syllabe qui précède, avant sa chute. Toute autre réalisation de ce ton dans cette position aboutit à une distorsion de la perception tonale des mots de ce type.

Dans la séquence $\grave{V}_1$-$C\grave{V}_2$, le ton bas sur V_1 pourrait être réalisé de façons différentes tout en étant phonologiquement bas : soit de façon légèrement descendante (tendance générale) par rapport à la descente beaucoup plus rapide du ton bas sur V_2, soit comme un palier plat, auquel cas un écart de 1/4 de ton ou plus sépare F_0 final de V_1 et F_0 initial de V_2. Dans tous les cas le bon bas sur V_2 est toujours réalisé avec une chute brutale de F_0. C'est aussi dans cette séquence qu'on peut constater une corrélation constante et proportionnelle entre la variation fon-

damentale et la variation d'intensité. En d'autres termes l'intensité décroît avec F_0 dans ce type de séquence tonale.

4.3. Interprétation et conclusions

Cette étude expérimentale qui allie l'analyse acoustique et l'analyse perceptive met en évidence des indices acoustiques prioritaires, dont l'auditeur èdõ ne peut se passer dans la perception de séquences VCV :

- la réalisation de chacun des deux tons peut varier et même altérer le sens de la variation de hauteur dans certains cas sans entraîner pour autant de confusion de signification pour l'auditeur natif.

- en revanche il apparaît dans certaines séquences des seuils critiques de variations tonales dont le dépassement fait passer d'une signification à une autre.

L'étude détaillée des quatre séquences tonales possibles sur V_1 CV_2 permet de dégager des indices acoustiques prioritaires pour l'identification par un natif:

1. pour la séquence H - H : l'indice est la pente montante (ou le palier plat) de F_0

2. pour la séquence B - B : la pente descendante de F_0

3. pour la séquence B - H : l'écart tonal entre F_0 final de V_1 et F_0 initial de V_2.

4. pour la séquence H - B : F_0 initial de V_2 doit être au niveau de F_0 final de V_1 avant de commencer une chute tonale longue (8)

5. Conclusion

Nous avons utilisé un modèle explicatif du traitement de la parole par un auditeur pour essayer d'analyser des aspects spécifiques de l'écoute en relation avec la production, avec les résultats combinés de l'analyse acoustique et de l'analyse contrastive. Ce modèle a comme principal mérite d'aider le chercheur à garder la maîtrise des aspects multiples et complexes du processus complet de Production-Perception-compréhension.

L'étude parallèle de trois situations nous a permis de préciser et de nuancer nos hypothèses initiales :

- il existerait une priorité de la démarche cognitive sur la démarche linguistique et une seconde priorité de la démarche linguistique sur la démarche perceptive (à caractère phonétique)

- l'utilisation du niveau phonétique par rapport au niveau linguistique ne se fait pas de façon simple, ni de façon ordonnée ; l'activité de traitement suppose un va-et-vient permanent entre les niveaux.

- le passage d'une langue à l'autre suppose des priorités : certaines différences phonétiques constituent des zones à densité maximale de difficulté et d'autres résistent peu à l'apprentissage.

- tout auditeur utilise dans sa langue des indices de reconnaissance dont le taux d'information varie ; ceci laisserait entendre qu'il peut se passer de certains éléments et pas d'autres... Il ne reste qu'à trouver lesquels.....

Références

(1) HATON, j.P., Intelligence artificielle en compréhension automatique de parole : état des recherches et comparaison avec la vision par ordinateur. Technique et Science Informatique 0752-4072, 1985/03, 265-23.

(2) PISONI, D.B.; SAWUSCH, J.R., Some Stages of Processing in Speech Perception, in Structure and Process in Speech Perception; ed. by A. COHEN & S.G. NOOTEBOOM, Communication and Cybernetics 11, Springer Verlag, 1975, 16-35.

(3) ROSSI, M., Les contraintes phonologiques dans un système de reconnaissance de la parole, 6èmes Journées d'Etude sur la Parole, Toulouse, 1975, 161-190.

(4) LHOTE, E., La compréhension orale, Le paysage sonore d'une langue, le français, Buske Verlag, 1986, chapitre 3. (à paraître)

(5) LHOTE, E., La parole et la voix, Buske Verlag 1982, chapitre 5.

(6) NOIZET, G., De la perception à la compréhension du langage, P.U.F., Paris, 1980.

(7) DI CRISTO, A., Prolégomènes à l'étude de l'intonation. Microprosodie. Ed. du CNRS, Paris, 1982.

(8) ROSSI, M., Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole, Phonética, 1971, 23, 1, 1-33.

CONTRIBUTION A L'ETUDE DES SYSTEMES VOCALIQUES
LE CAS DU VIRI (SUD-SOUDAN)

J.M. Hombert* et R. Point**

*Université Lyon 2 et Lacito-CNRS, ** INSA, Lyon

ABSTRACT

Viri, an Adamawa-Eastern language spoken in Southern Sudan, has a seven vowel system : [i, ɪ, ɛ, a, ɔ, ʊ, u].

In a vowel perception experiment, 53 synthetic vowel stimuli - F1, F2 and F3 values being selected in such a way that these stimuli optimally covered the acoustic vowel space - were presented to a Viri speaker. His task was to identify which stimuli "sounded like" one of his 7 vowels. The vowels ɪ and ʊ were never recognized by the subject. An acoustic analysis revealed that F1, F2 and F3 values were very similar for u and ʊ and quasi identical for i and ɪ. It is suggested that relatively unusual acoustic parameters such as the relative amplitude of the first two formants and the proximity of F3 and F4 may play a major role in the perception of Viri vowels (especially in the case of the i/ɪ distinction).

1. LE SYSTEME VOCALIQUE DU VIRI

Le viri est une langue du groupe adamawa-oubanguien (famille nigéro-congolaise) parlée au sud Soudan. Son système vocalique comprend les sept voyelles suivantes : i, ɪ, ɛ, a, ɔ, ʊ, u. La figure 1 montre que ces voyelles qui, dans les exemples donnés, portent toutes un ton moyen représenté par un tiret au-dessus de la voyelle, peuvent se trouver dans le même contexte.

lɪ̄	"rève"	lʊ̄	"danse"
lī	"soleil"	lū	"arriver"
lɛ̄	"mon oeil"	lɔ̄	"oeil"
	lā "vous"		

Fig. 1 - Les 7 voyelles du viri

2. VALEURS FORMANTIQUES

Afin de préparer le sujet à l'expérience perceptuelle décrite dans le paragraphe suivant, il lui a été demandé de produire des séquences lV puis de répéter la voyelle seule. Après une digitalisation du signal à 10 kHz, les valeurs des formants des voyelles du viri ont été obtenues par la méthode de la prédiction linéaire (autocorrélation - 12 coefficients) sur des segments de 25 msec pris au centre des voyelles prononcées en isolation. Les valeurs de F1 et F2 pour 5 répétitions de chacune des 7 voyelles sont présentées en figure 2. Ces valeurs sont regroupées - pour chaque voyelle - dans une ellipse de dispersion (intervalle de confiance de 90 %).

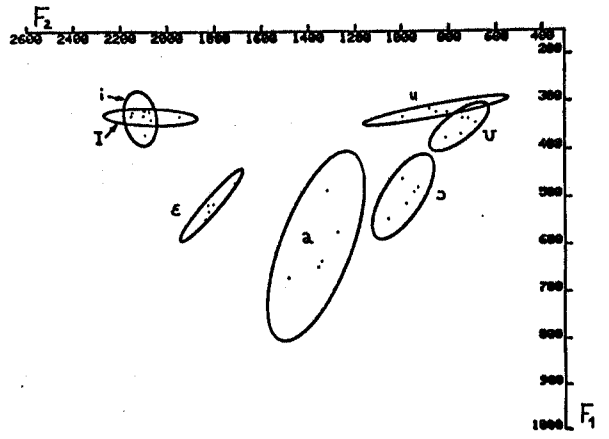


Fig. 2 - Distribution acoustique des voyelles du viri dans un espace F1/F2 (1 locuteur, 5 répétitions par voyelle)

On remarquera que ces ellipses de dispersion de i et ɪ principalement et de u et ʊ dans une certaine mesure se recoupent. Ce chevauchement dans l'espace F1/F2 ne disparaît pas si l'on introduit les valeurs de F3 comme le montre la figure 3. En effet les valeurs de F3 sont très proches pour u et ʊ et se recouvrent complètement dans le cas de i et ɪ.

		F1	F2	F3
ɪɪ	1	320	2109	3046
	2	371	2105	2941
	3	331	2112	3085
	4	323	2158	2740
	5	322	2088	2872
ɪɛ	1	340	2078	2854
	2	321	2104	2949
	3	331	2162	2940
	4	329	2057	3085
	5	332	1960	2803
ɪɐ	1	546	1849	2972
	2	531	1836	
	3	517	1838	2695
	4	515	1818	2931
	5	470	1725	2723
ɪa	1	671	1494	
	2	648	1368	2609
	3	636	1354	2584
	4	574	1287	2561
	5	486	1332	2557
ɪo	1	515	993	2682
	2	547	1071	2577
	3	489	959	2511
	4	479	944	2469
	5	462	1008	2511
ɪɒ	1	376	825	2435
	2	337	727	2499
	3	345	697	2500
	4	335	752	2428
	5	369	759	
ɪu	1	301	688	2248
	2	322	864	2408
	3	322	818	2304
	4	332	1006	2296
	5	315	891	2232

Fig. 3 - Valeurs de F1, F2 et F3 des 7 voyelles du viri (1 locuteur, 5 répétitions par voyelle). Les valeurs de F1 et F2 sont les mêmes que celles reportées dans la figure 2

3. EXPERIENCE PERCEPTUELLE

Pour essayer de mieux cerner les paramètres acoustiques utilisés par les locuteurs du viri pour distinguer les 7 voyelles du système, nous avons procédé à une expérience perceptuelle dont le protocole a déjà été présenté [1], [2].

En bref il s'agit de faire découper l'espace vocalique par les locuteurs eux-mêmes, en fonction de leur organisation perceptuelle.

Après avoir été familiarisé avec la tâche qui consiste à isoler une voyelle (à partir d'une séquence lV), le sujet écoute une bande sur laquelle est enregistré un jeu de 53 stimuli synthétiques dont les valeurs de F1, F2 et F3 sont présentées dans la figure 4.

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

Fig. 4 - Valeurs des formants F1, F2 et F3 pour les 53 stimuli synthétiques (les stimuli 42 à 47 représentent les voyelles antérieures arrondies, les stimuli 48 à 53 représentent les voyelles postérieures non arrondies).

Après la présentation de chaque stimulus le sujet doit indiquer quel mot, choisi parmi ceux présentés dans la figure 1, contient le même "son" (timbre vocalique) que le "son" (stimulus synthétique) qu'il vient d'entendre. Le sujet a aussi la possibilité de répondre qu'aucun des mots de la figure 1 ne correspond au stimlus qui lui est proposé. Chaque stimulus a été présenté 10 fois.

La présentation des résultats de ce test perceptuel est donnée en figure 5. Nous avons regroupé dans un même cadre les stimuli qui au moins 9 fois sur 10 ont été identifiés comme la "même voyelle viri" (les stimuli 42 à 53 représentant les voyelles antérieures arrondies et postérieures non arrondies n'ont pas été représentés sur le graphique de la figure 5 par souci de clarté). L'examen de ces résultats révèle que deux des sept voyelles - ɪ et ʊ - n'ont pas été "reconnues".

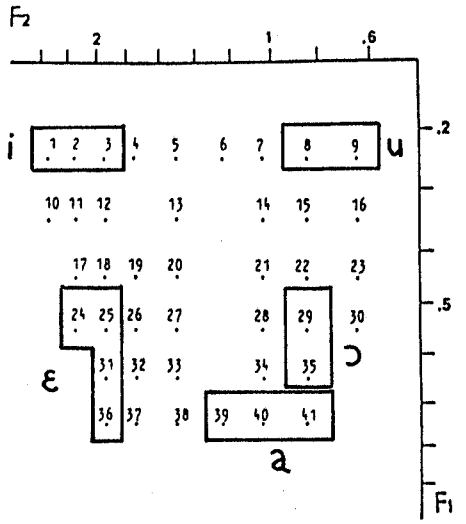


Fig. 5 - Résultats du test perceptuel (les zones encadrées regroupent les stimuli identifiés comme mêmes voyelles dans au moins 9 présentations sur 10)

4. ANALYSE SPECTRALE

Si, au lieu de se limiter à la comparaison des pics formantiques de F1, F2 et F3, nous prenons en considération l'ensemble du spectre, il devient alors possible de faire apparaître des indices acoustiques permettant de différencier par exemple les voyelles i et ɪ.

La figure 6 montre des spectres représentatifs de ces deux voyelles. On s'aperçoit alors que ces deux spectres se distinguent par :

- la différence d'amplitude relative du premier et du second formant : l'amplitude du deuxième formant de ɪ est beaucoup plus faible que celle du i ;

- la proximité - en amplitude et en fréquence - du troisième et du quatrième formant du ɪ comparativement aux formants correspondants du i.

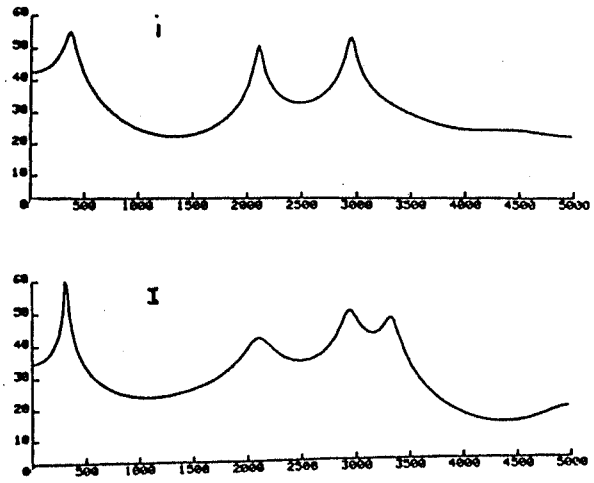


Fig. 6 - Spectres représentatifs de i et ɪ

Il est probable que ces indices jouent un rôle important dans la discrimination entre i et ɪ. Puisque notre ensemble de voyelles synthétiques ne prenait pas en compte les variations de ces paramètres, qui ne sont pas généralement considérés comme jouant un rôle prépondérant dans la perception des voyelles - il n'est pas surprenant que certaines voyelles du viri n'aient pas été identifiées par le sujet locuteur soumis au test perceptuel.

5. CONCLUSION

Nous sommes conscients du caractère préliminaire de cette étude. Il est bien évident que les problèmes qu'elle pose et les éléments de réponse que nous avons tenté d'apporter devront être approfondies par des travaux qui s'attacheront prioritairement
- à mettre en évidence le rôle perceptuel des indices acoustiques présentés au paragraphe précédent en préparant une nouvelle expérience perceptuelle qui incorpore ces nouveaux indices comme paramètres des stimuli ;
- à étendre cette étude à un plus grand nombre de locuteurs viri ;
- à cerner les réalisations articulatoires responsables de ces distinctions acoustiques peu courantes.

Les langues pour lesquelles nous disposons de bonnes données phonétiques sont de l'ordre de quelques dizaines, c'est-à-dire environ 1 % des langues du monde. De ce fait la collecte de données, même préliminaires, sur des langues "exotiques" est une tâche prioritaire pour les phonéticiens si l'objectif visé est la compréhension de tous les mécanismes possibles en production et perception de la parole.

Le cas du viri confirme l'importance de l'étude de l'ensemble de la distribution spectrale par rapport à la simple localisation des pics formantiques. Il apparaît, par ailleurs, qu'un locuteur viri, comme cela a déjà été signalé pour d'autres langues de la même zone géographique, utilise l'amplitude relative des deux premiers formants comme discriminant.

De nouveaux tests perceptuels devraient permettre de décider le poids relatif de ce discriminant par rapport à d'autres discriminants possibles.

REFERENCES

[1] J.M. Hombert, "Universals of vowel systems : the case of centralized vowels", Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhague, 1979, vol. 2, 27-32.

[2] J.M. Hombert et G. Puech, "Espace vocalique et structuration perceptuelle en swahili", Pholia 1, 1984, 199-208.

INTEGRATION PERCEPTIVE DE LA FREQUENCE FONDAMENTALE ET DE LA DUREE D'UN INTERVALLE DE SILENCE LORS DE L'IDENTIFICATION D'ITEMS LEXICAUX

N. Bacri

Laboratoire de Psychologie Expérimentale
54, Boulevard Raspail, 75006 Paris

In two experiments, bearing on the identification of mono- and dissyllabic isolated words, two kinds of cues are analysed: duration of a silence inserted between a vowel and a fricative noise, flat or rising F_0 movements. Identification functions and decision times differ for the two classes of words. Context effects and a trading relation between cues are found only for monosyllabic "ambiguous" stimuli. Generally, F_0 changes shorten decision times. The data support the auditory integration hypothesis of phonemic and non phonemic cues at an early stage in perception, when auditory F_0 contrasts trigger off anticipatory mechanisms, whereas stress pattern on dissyllables is independently processed. Identification, specifically in this last case, is determined by a matching between acoustic input and lexical prototypes.

L'identification d'items lexicaux implique des stratégies de codage du signal acoustique, qui supportent l'interprétation des indices pertinents pour la reconnaissance phonétique et la segmentation du signal. Le premier objectif de cette recherche est d'étudier le type d'information disponible pour une prise de décision sur l'identité de mots, en précisant si les informations spécifiquement phonémiques sont seules traitées ou si l'intégration perceptive porte sur toutes les informations acoustiques, qu'elles aient valeur phonémique ou non. Parmi ces dernières, la fréquence fondamentale a un effet sur la perception du voisement d'une occlusive pré-vocalique <1,2> ou d'une fricative <3>. Ses variations peuvent interagir avec celles d'une durée vocalique. Mais son effet n'est pas alors stable: le degré de la compensation, son existence même sont dépendants des valeurs contextuelles <4>.

Par suite, le statut des procédures qui fondent les processus d'identification fait problème. Elles peuvent relever strictement de l'analyse auditive pré-syllabique ou comporter un niveau post-syllabique. En ce dernier cas, les "sorties" de l'intégration auditive pourraient être modulées en fonction, par exemple, de la structuration syllabique de l'item lexical et éventuellement de sa représentation prototypique.

Un second objectif de la recherche est donc d'analyser le jeu entre niveaux de traitement et d'explorer certaines des conditions auditives nécessaires à la mise en place de processus cognitifs qui interviennent dans la prise de décision, comme par exemple les processus d'anticipation. Le choix de confronter la fréquence fondamentale et un intervalle de silence précédant un bruit de frication présente de ce fait un intérêt supplémentaire. Il permet d'étudier les modalités de fonctionnement d'un paramètre intonatif, F_0, dans ses relations avec une seconde dimension acoustique, le silence, qui, selon sa valeur et le contexte, est intégré perceptivement comme tenue d'une occlusive <1, 5> ou reçoit une valeur intonative <6>. L'hypothèse est que les contrastes auditifs créés sur les variations de F_0 seront l'objet de deux types de traitements. Ils contribueront à l'intégration auditive du signal et/ou modifieront les conditions d'accès aux mots en entrant dans la formation d'un pattern accentuel.

Nous avons soumis au jugement d'auditeurs, lors de deux expériences successives, deux groupes de mots, monosyllabiques et dissyllabiques, de structure $(C_1)\ V_1\ C_2$ et $(C_1)\ V_1\ C_2\ V_2$, (C_2 = fricative), comportant un silence de durée variable situé entre V_1 et C_2. La fondamentale était maintenue stationnaire ou bien variait. La montée linéaire de F_0 sur les monosyllabes est susceptible de créer un contraste auditif. Cette même montée sur la première syllabe des dissyllabes, suivie d'un tracé stationnaire sur la seconde syllabe, réalise la forme minimale d'une configuration intonative, le pattern accentuel. L'analyse des jugements d'identification permettra d'apprécier l'interaction entre F_0 et silence, et de voir si elle est sensible à la structure syllabique de l'item lexical. L'étude des temps de décision, qui expriment le degré de certitude des auditeurs, conduira à une évaluation des modalités de traitement des contrastes auditifs, selon que l'information qu'ils apportent sur la forme sonore de l'item est ou non pertinente à la tâche.

METHODE

Sujets: 16 adultes (6 femmes et 10 hommes), disant ne pas présenter de déficience auditive, ont participé aux deux expériences, en deux sessions

séparées par un laps de temps d'au moins une heure.

Stimuli: 5 monosyllabes et 5 dissyllabes ont été enregistrées par un locuteur masculin, puis analysées et synthétisées (système dérivé de LPC). La fréquence d'échantillonnage était de 10 kHz. A chaque item de départ correspondait un mot n'en différant que par la présence de l'occlusive vélaire non voisée /k/ précédant la fricative. Par exemple, à "as" correspond "axe", à "assez" "accés".

Durées	as	tasse	es	bosse	fils
V_1	170	220	183	170	190
$(C)V_1$	170	240	183	350	287
/s/	245	235	368	289	320
D. totale	415	475	551	639	607
	assez	essai	session	aussi	fission
V_1	122	134	80	126	48
$(C)V_1$	122	134	265	126	156
/s/	208	156	256	214	181
D. totale	578	643	648	600	650

Caractéristiques temporelles des stimuli (en ms).

Le silence, variant de 0 à 50 ms par pas de 10 ms, était situé à la fin de la dernière période correspondant à la transition vocalique et précédait immédiatement le bruit de frication. Pour chaque expérience, deux types de variations de F_0 ont été réalisés:
- variations stationnaires sur les parties voisées du signal: 150 Hz
- monosyllabes: montée linéaire de 150 Hz à 250 Hz, à partir du premier segment voisé
  dissyllabes: montée linéaire de 150 Hz à 250 Hz sur la partie vocalique de la 1ère syllabe, suivie d'un tracé stationnaire à 150 Hz sur la voyelle de la seconde syllabe.

Les 12 stimuli (6 stationnaires et 6 montants) construits à partir d'un même mot sont présentés en ordre aléatoire à l'intérieur d'un bloc. L'intervalle inter-stimuli était de 7 s, entre blocs de 2 à 3 minutes.

Procédure: Les stimuli étaient présentés binauralement (écouteurs). Les sujets ajustaient leur niveau d'écoute librement. La passation, individuelle, eut lieu dans un milieu non bruyant. Une période d'entrainement, avec un matériel différent, précédait chaque expérience. Les sujets devaient décider le plus rapidement possible s'ils entendaient /s/ ou /ks/; ils étaient familiarisés, pour chaque mot, avec les deux termes de l'alternative. Les temps de décision ont été calculés à partir de la fin de l'intervalle de silence, en soustrayant du temps total la durée écoulée entre le déclenchement de l'horloge et le début du bruit de frication.

JUGEMENTS D'IDENTIFICATION

L'analyse globale des réponses aux items monosyllabiques montre un déplacement significatif de la fonction d'identification: 23,5 ms de silence sont nécessaires pour que le groupement /ks/ soit perçu lorsque F_0 est stationnaire, 16 ms suffisent en présence d'une montée linéaire de F_0 (Fig. 1). Ce déplacement est interprétable en référence à l'établissement d'une relation de compensation entre la durée de l'intervalle de silence et la forme des variations de F_0, sous la réserve que le concept de compensation contextuelle soit étendu aux cas de formation d'un percept intégré, même en l'absence d'une équivalence phonétique au sens strict.

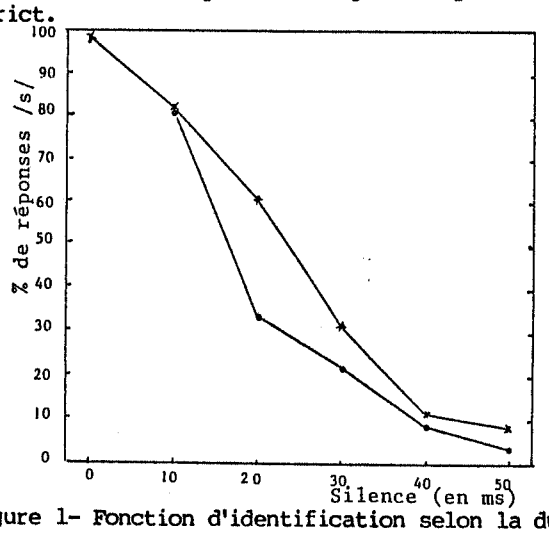


Figure 1- Fonction d'identification selon la durée du silence, pour chaque pattern intonatif (-x- stationnaire, -•- montant). Monosyllabes.

L'évaluation, par interpolation linéaire, des "frontières" pour chaque auditeur, confirme l'analyse globale. Les moyennes dérivées pour l'ensemble des sujets diffèrent significativement (F_0 stationnaire: 24,6 ms; σ :7,65; F_0 montant: 19,3 ms; σ :6,1; $t(30)=2,05$, $p<.05$). On notera que d'autres indices acoustiques ont pu intervenir dans l'identification des mots, en particulier la durée de la voyelle et celle de la fricative. La conjonction de durées relativement brèves retarde l'apparition du groupement /ks/, ce qui traduit le recouvrement entre indices successifs et suggère un traitement simultané des différents indices composant le signal.

Les items dissyllabiques donnent lieu à une toute autre distribution des réponses: l'effet du silence se maintient, mais on constate l'absence d'un déplacement significatif de la fonction d'identification (Fig. 2). Le percept /ks/ se forme pour 19,5 ms de silence avec F_0 stationnaire, 16 ms avec un pattern accentuel. L'étude sujet par sujet va dans le même sens: l'intégration du silence comme tenue d'une occlusive s'effectue sans interagir avec la fondamentale.

Il est intéressant de souligner qu'une interaction n'apparait entre F_0 et durée de silence que lorsque les stimuli, monosyllabiques, sont ambigus. Aux extrêmes du continuum de silence,

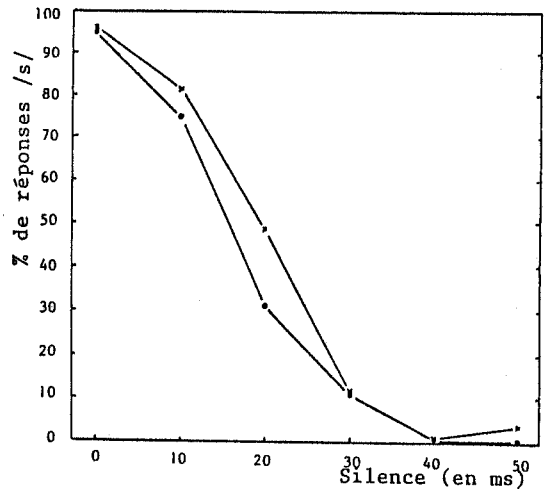


Figure 2- Fonction d'identification selon la durée du silence, pour chaque pattern intonatif (-×- stationnaire, -•- contrastif). Dissyllabes.

l'indice phonémique apporte une information suffisante. Ce n'est que lorsqu'un silence de durée moyenne rend plus difficile l'appariement entre données acoustiques et représentation auditive qu'on constate un décalage systématique de l'identification vers des valeurs plus brèves du silence, en présence d'une variation rapide de F_0. Malgré cette réserve, il n'en reste pas moins qu'alors l'interaction entre indices phonémique et intonatif apparait clairement.

La dissymétrie entre les réponses aux mono et aux dissyllabes sera interprétée en référence à la différence de position du moment critique à l'intérieur du mot. Cette différence renvoie à la structuration syllabique, qui sous-tend la distinction entre contrastes auditifs, formant ou non un pattern accentuel. Les modalités de leur traitement seront évaluées à partir de l'analyse des temps de décision.

TEMPS DE DECISION

Les analyses de variance portant sur l'ensemble des réponses aux mono et aux dissyllabes respectivement montrent l'effet significatif du pattern d'intonation (mono: $F(1,15) = 22,10$, $p<.005$; dissyl.: $F(1,15) = 42,38$, $p<.001$) et de l'intervalle de silence (mono: $F(5,75) = 5,11$, $p<.01$; dissyl.: $F(5,75) = 13,21$, $p<.001$). Mais l'interaction entre les deux facteurs n'est significative que pour les monosyllabes ($F(5,75) = 2,96$, $p<.05$) (Fig. 3 et 4).

L'analyse séparée des deux conditions intonatives montre à nouveau une dissymétrie entre monosyllabes et dissyllabes: pour les monosyllabes, le silence n'a d'effet significatif qu'en présence d'un tracé F_0 montant, alors que cet effet est manifeste dans les 2 cas pour les dissyllabes.

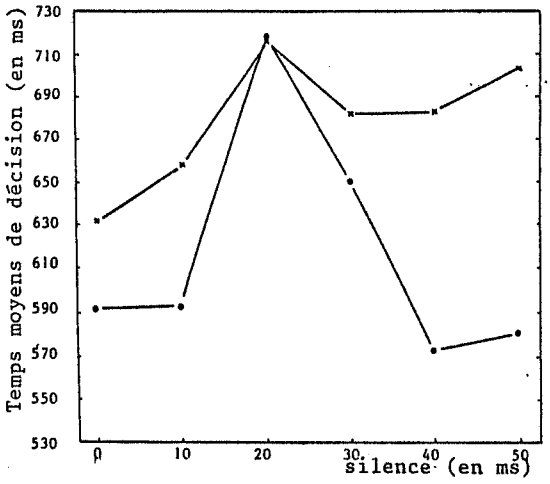


Fig. 3- Temps moyens de décision selon la durée du silence et le pattern intonatif (-×- stationnaire, -•- montant). Monosyllabes.

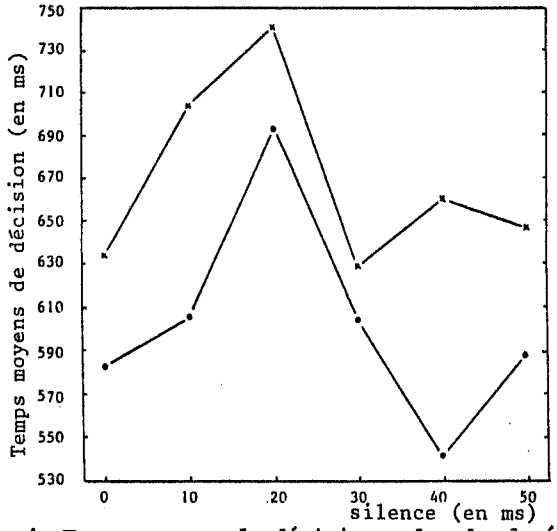


Fig.4- Temps moyens de décision selon la durée du silence et le pattern intonatif (-×- stationnaire, -•- contrastif). Dissyllabes.

L'identification de mots se différencie nettement selon la structuration syllabique de l'item. Si le moment critique est en position pénultième, une interaction s'établit entre F_0 et silence, mais la vitesse de la décision n'est pas influencée par la durée du silence quand F_0 est stationnaire. Ce résultat montre que les deux dimensions du signal sont l'objet de deux traitements distincts. On peut estimer en première approximation que la réalisation d'un fort contraste auditif (montée de 100 Hz sur 1 syllabe) précédant le silence entraine une surestimation de ce dernier, ce qui faciliterait la prise de décision et expliquerait le déplacement de la fonction d'identification. Mais cette interprétation des résultats en référence aux seules conditions de l'analyse auditive, ne rend pas compte de deux aspects des données:

- La décomposition de l'effet de F_0 sur les T.D. montre que sa significativité provient de la différence entre les seules valeurs extrêmes du

silence(10 ms, 40 et 50 ms). La rapidité de la décision est globalement liée à sa probabilité. Aux réponses aléatoires correspond une forte incertitude, quelle que soit la forme de F_0.

- Aucune interaction ou compensation entre F_0 et silence n'apparait lorsque le silence est en position médiane: les variations de F_0 n'influencent pas les jugements d'identification et l'effet de chacun des 2 facteurs sur la rapidité de la décision est indépendant.

Ces faits suggérent que les procédures d'analyse auditive sont plus complexes qu'il n'apparaissait au premier abord. Identifier une monosyllabe est une tâche simple, comparable à une détection de phonème. La présence d'un contraste auditif accroit la sélectivité de l'information et entraine la focalisation de l'attention sur l'indice pertinent pour l'identification du mot: le silence, dans sa relation avec le bruit de frication. La variation de F_0 active l'identification, centrée sur un indice qui fonctionne comme dominant. Les compensations entre F_0 et silence renforcent cette dominance en ce qu'elles modifient la probabilité de l'identification et facilitent la formation d'un groupement consonantique. En l'absence d'un contraste auditif, le calcul porte sur tous les indices présents dans le signal, ce qui ralentit la décision.

Lorsque l'auditeur a à traiter de dissyllabes, les conditions d'élaboration de sa réponse sont différentes: passer de "fission" à "fiction" implique le repérage de la structure syllabique de la séquence, en même temps que la construction d'un pattern accentuel (succession d'une montée linéaire de 100 Hz et d'un tracé stationnaire). L'absence d'effet du type de variations de F_0 sur l'identification, son influence sur la rapidité de la prise de décision, qui ne s'accompagne toutefois pas d'une interaction entre les 2 facteurs, conduisent à penser que les informations portées par le pattern accentuel ont été l'objet d'un traitement spécifique. Le contraste auditif marque la limite intersyllabique; il facilite la décision. Cet effet est indépendant de la durée du silence. On en infère qu'il y a eu dissociation entre informations phonémiques et informations intonatives. En ce dernier cas, les variations de F_0 sont traitées sur une durée relativement longue comme un pattern accentuel, ce qui modifie les procédures d'accès à l'item. Le repérage de la structuration syllabique globale s'effectue indépendamment de l'identification syllabique.

CONCLUSIONS

L'étude de l'identification de mots isolés, mono- et dissyllabiques, présentant une fondamentale stationnaire ou bien un contraste auditif qui avait ou non valeur intonative, permet en premier lieu de conclure que l'intégration perceptive porte sur toutes les informations distribuées au long du signal. Une décision phonémique peut prendre en compte un indice non phonémique, si celui-ci, comme c'est le cas pour une variation rapide de F_0, modifie la sélectivité de l'information phonémique.

Le traitement des variations de F_0 et de la durée de silence s'effectue en parallèle. En effet, un accés séquentiel à ces deux types d'information entrainerait une compétition qui aurait ralenti la décision. Il est probable que leur intégration auditive en un percept ne relève pas seulement de contraintes physiologiques. On peut certes admettre que le traitement des changements continus du spectre acoustique et des événements que constituent les contrastes se réalise dès le niveau auditif. Mais l'intégration auditive s'est avérée sensible au contexte linguistique. Nous ne pouvons ici préciser si la structuration syllabique intervient pendant l'analyse auditive ou module ses sorties. Les réponses aux monosyllabes suggérent que l'intégration de F_0 aux autres indices précéde l'identification. Mais les dissymétries entre les réponses aux mono- et aux dissyllabes, leur différenciation selon que les stimuli sont ou non de "bons exemplaires" du mot vont dans le sens de l'intervention dans la prise de décision d'un niveau de traitement post-syllabique, spécifiquement lexical. Le percept formé auditivement sera re-évalué en fonction de la représentation que les sujets ont de la forme sonore des items lexicaux.

Une seconde conclusion est que la fréquence fondamentale est elle-même l'objet de deux types de traitement, situés probablement à deux niveaux distincts, pré- et post-syllabique. Corrélat acoustique de distinctions phonémiques, elle contribue à l'intégration des segments. En ce sens, la perception de l'intonation a son fondement dans les mécanismes d'intégration des configurations spectro-temporelles. Base de la formation de configurations intonatives, elle modifie, non l'identité du mot, mais les conditions de son accés. Sa variation influence le degré de certitude des auditeurs, et permet la mise en place de mécanismes d'anticipation. Ces mécanismes modifient les processus de codage des stimuli, en ce qu'ils régissent les conditions de l'ajustement de l'auditeur au signal, et corrélativement les prises de décision sur l'identité des mots.

REFERENCES

<1> M. Haggard, Q. Summerfield, M. Roberts, "Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F_0 cues in the voiced-voiceless distinction", J. of Phonetics, 9, 49-62, 1981.

<2> R.N. Ohde, "Fundamental frequency as an acoustic correlate of stop consonant voicing", J.A.S.A., 75(1), 224-230, 1984.

<3> M.A. Derr, D.W. Massaro, "The contribution of vowel duration, F_0 contour and frication duration as cues to the /juz/-/jus/ distinction", Perception & Psychophysics, 27(1), 51-59, 1980.

<4> A. Nicaise, N. Bacri,"Durée vocalique, variations de F_0 et perception de frontières", 14è J.E.P., 175-178, 1985.

<5> M.F. Dorman, L.J. Raphael, A. Liberman, "Some experiments on the sounds of silence in phonetic perception", J.A.S.A., 65, 1518-1532, 1979.

<6> N. Bacri, A. Nicaise, "Prosodie et intégration phonétique du silence", 13è J.E.P., 227-228, 1984.

LA PERCEPTION DES TONS DU CHINOIS PAR DES FRANCOPHONES ET PAR DES CHINOIS

P.W. CHI LEE * et M.H. PERON * *

*Université Tamkang, Taïwan, **Laboratoire de Phonétique de Besançon

ABSTRACT

The phonetic common sense generally associates a good production with a good perception; it appears logical therefore to expect that the Chinese would correctly perceive the tones of their language.
A twofold experiment was carried out in Taïwan and Besançon. The Taïwan investigation which was chiefly concerned with the acquisition of french prosody, aims at checking the ability of the Chinese to identify tones in their language. The Besançon study focused on the influence of musical experience on chinese tones production and perception by French listeners.
The results from these experiments tend to differ from the "common sense" expectation: they attest to the priority of a linguistic task in the perceptual process, and to the lack of a systematic correspondence, in a native speaker, betwen production and discrimination-identification.

INTRODUCTION

"Chaque langue engendre chez ceux qui la parlent un ensemble d'habitudes perceptives, linguistiques et culturelles" (1)

A partir d'une langue tonale, le chinois, nous avons mené une double approche visant à étudier la perception tonale chez des Chinois et des Français. Les expériences menées du côté chinois et du côté français présentent des objectifs différents :

- du côté chinois, nous aimerions savoir comment les Chinois perçoivent la mélodie tonale de leur langue. Notre but est pédagogique, il vise à vérifier l'existence d'une corrélation entre la perception tonale du chinois et la perception mélodique du français chez des Chinois apprenant le français ;

- du côté français, notre travail présente un double objectif : nous avons essayé d'étudier la perception des tons du chinois par des francophones et nous nous sommes ensuite intéressées à la perception tonale chez des sujets d'expérience musicale différente.

1. Expérience 1 : Perception et production des tons du chinois par des Français.

Pour l'ensemble des tests, les tons de référence correspondent aux cinq tons décrits par CHAO (2) comportant quatre tons phonologiques auxquels s'ajoute un ton 3', qui est un ton bas descendant, variante du ton 3 (Figure 1)

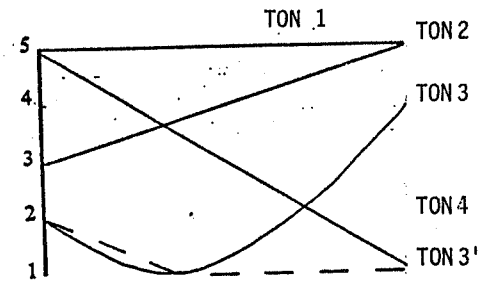


Figure 1 : Représentations graphiques des tons du chinois mandarin

LES CINQ NIVEAUX DE Y.R. CHAO*

TON 1 : 5 5
TON 2 : 3 5
TON 3 : 2 1 4
TON 4 : 5 1

TON 3' : 2 1 1

1.1. Test n° 1 : exercice d'identification.

Le corpus comprend 100 unités monosyllabiques correspondant aux voyelles /a-o-i-u-y/ ; les cinq tons ont été répartis au hasard.

L'enregistrement a été effectué par une étudiante chinoise parlant le chinois mandarin. La parole et le glottogramme, captés respectivement par un microphone et un glottomètre, ont été enregistrés sur les deux voies d'une bande magnétique, la mélodie obtenue à partir du glottogramme par l'analyseur de mélodie a été transcrite à l'aide d'un enregistreur graphique de type OSCILLOMINK.

Nous avons choisi neuf sujets répartis en trois groupes égaux : groupe 1 (musiciens confirmés ayant fait du solfège et jouant d'un instrument), groupe 2 (musiciens jouant ou ayant joué d'un instrument), groupe 3 (non musiciens).

Nous avons demandé aux auditeurs de noter les tons entendus suivant un code graphique préalablement établi. Une phase d'apprentissage précédait le test , ceci afin de familiariser les auditeurs avec les tons.

Discussion et interprétation des résultats.

Nous avons constaté que les tons les mieux reconnus étaient le ton 1 (10,5 % d'erreurs), le ton 3' (24 % d'erreurs), le ton 4 (24 % d'erreurs) et les tons 3 (28 %) et 2 (34,5 %).

Il semble donc que le ton statique (ton 1) est le mieux identifié.

Les travaux de HOMBERT (3) ont montré que les glissandos descendants étaient mieux perçus que les glissandos montants. Par contre, ROSSI (4) a démontré qu'il n'y avait pas de différence de perception entre les deux. Pour ce qui nous concerne nous ferons simplement remarquer que les tons 2, 3' et 4 sont de sens opposé mais pas de même pente. Par conséquent, si dans notre test nous observons que les tons descendants sont mieux perçus que les tons montants, nous ne pouvons affirmer que cela est dû au seul sens de leur glissando.

Nous avons également observé que les sujets musiciens reconnaissaient mieux les tons que les autres sujets. Ils ne font pas d'erreur sur le sens des glissandos mais tendent plutôt à confondre les tons descendants (tons 3' et 4). Ainsi l'expérience musicale semble jouer un rôle important pour l'identification des tons (figures 2 et 3)

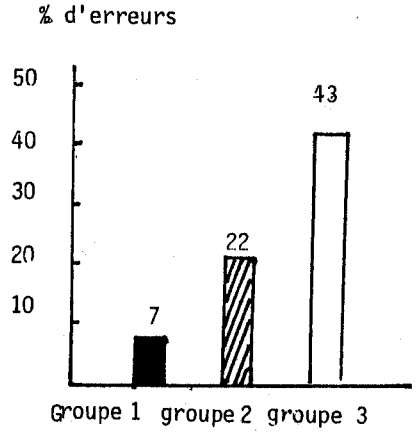


Figure 2 : Pourcentages d'erreurs par groupe :

groupe I musiciens confirmés
groupe II musiciens jouant d'un instrument
groupe III non musiciens

ton 2 identifié à un ton 3 : 18,8 %)		erreurs faites
ton 3 identifié à un ton 2 : 9,6 %)	28,4 %	par les groupes I, II, II
ton 3 identifié à un ton 4 : 18,8 %)		erreurs faites
ton 4 identifié à un ton 3' : 5,5 %)	23,8 %	par les groupes I, II, III
ton 3 identifié à un ton 4 : 6,4 %)		erreurs faites
ton 4 identifié à un ton 3 : 6,4 %)	12,8 %	par les groupes II, III
ton 4 identifié à un ton 2 : 6,0 %)		erreurs faites
ton 2 identifié à un ton 4 : 3,2 %)	9,2 %	par les groupes II, III
ton 1 identifié à un ton 2 : 6,4 %)		erreurs faites
ton 2 identifié à un ton 1 : 1,4 %)	7,8 %	par le groupe I

Figure 3 : Analyse, ton par ton, des erreurs faites par chacun des trois groupes.

2. Test n° 2 : exercice de répétition

Le corpus comprend 25 unités syllabiques. Nous avons demandé à trois locuteurs, choisis parmi les neuf auditeurs du test n° 1 et représentant chaque groupe, de répéter les tons entendus.

Discussion et interprétation des résultats.

Nous avons remarqué que les tons 1,2 et 3 donnent lieu à des erreurs altérant leurs propriétés. Pour les tons 3 et 4, le sens du glissand est respecté ; seule varie la hauteur finale et/ou initiale.

La réalisation la plus fidèle au modèle correspond à celle du locuteur du Groupe 2. Bien que ce sujet ne possède pas une expérience musicale optimale d'après notre définition, il est le seul à avoir appris le chant. La pratique du chant, paramètre oublié au début de notre expérience, devrait être prise en compte dans l'analyse de la production tonale.

2.3. Synthèse des tests n° 1 et n° 2.

Nous avons comparé les résultats obtenus aux deux tests par les trois sujets auditeurs/locuteurs et nous avons regardé s'il existait une relation entre l'identification et la répétition des tons chez un même sujet. Nous avons constaté que le sujet du groupe 1 obtenait de meilleurs résultats au test d'identification grâce à la pratique du solfège ; le sujet du groupe 2 répétait mieux les tons qu'il ne les identifiait et celui du groupe 3 présentait les résultats les moins bons aux deux tests.

Il semble donc, que pour tester la perception tonale d'1 sujet, les deux catégories de tests soient nécessaires.

2. Expérience n° 2 : Perception mélodique et perception linguistique des tons chinois par des Chinois. (5)

2.1. Processus expérimental :

Notre corpus est composé de 34 stimuli naturels comprenant des mots et des logatomes d'une à quatre syllabes. Une locutrice chinoise a été invitée à prononcer le corpus. Les mêmes processus et appa-

reillages d'enregistrement que dans l'expérience précédente ont été employés. 45 étudiants chinois du département du français à l'université de Tamkong à Taïwan ont été invités à subir nos deux tests.

Le premier est un test de perception mélodique dans lequel la tâche consiste à faire dessiner la courbe mélodique perçue après l'écoute de chaque stimulus ; ce test a été précédé d'un exercice d'entraînement à la représentation de mélodies tonales par le dessin. Le deuxième est un test de perception tonale dans lequel on demande d'identifier les tons de chacune des syllabes de ces stimuli à l'aide des quatre chiffres 1, 2, 3, 4, qui représentent respectivement le premier, le deuxième, le troisième et le quatrième tons chinois. Nous pouvons donc prétendre que le premier test fait appel à une perception purement acoustique, tandis que le deuxième suppose une mise en relation des faits physiques et des éléments d'un système linguistique.

2.2. Résultats et discussion :

A notre grande surprise, nos auditeurs qui peuvent tous produire correctement les tons chinois, ne maîtrisent pas forcément la perception mélodique des tons de leur langue. Le taux moyen de réponses correctes ne s'élève qu'à 43,9 %. Nous avons relevé 3 types de fautes pour la perception mélodique :

- Incohérence du sens de variation de hauteur : Certains auditeurs dessinent une courbe montante pour le premier ton et le quatrième ton. Quant au troisième ton, il est aussi bien représenté par le dessin d'une courbe montante que par celui d'une courbe descendante.
- Interférence du système linguistique : 50 % de nos auditeurs ont dessiné un "V" pour représenter le 3e ton. Ce troisième ton qui est noté linguistiquement par un "V" ne présente en fait cette forme mélodique que dans le mot isolé fort accentué ; or dans notre corpus aucun stimulus n'est prononcé de cette façon.
- Faible sensibilité à la variation relative de hauteur d'une syllabe à l'autre pour les stimuli polysyllabiques : 43 % seulement des auditeurs ont une bonne perception de la forme et du sens de la variation de hauteur à l'intérieur de la syllabe, mais ne sont pas du tout sensibles à la relation de hauteur d'une syllabe à l'autre.

Quant au test d'identification linguistique, nous avons remarqué que nos auditeurs se sentaient beaucoup plus à l'aise que dans le test précécent, parce qu'ils pouvaient se servir de leurs acquis linguistiques. Et c'est encore le troisième ton qui présente le plus de confusion. Ceci correspond aux résultats des autres travaux (6). Le taux moyen de perception correcte s'élève à 87,25 %, ce qui représente un taux élevé par rapport à celui du premier test (43,9 %)

Dans le permier test, en faisant dessiner les courbes mélodiques, nous croyions faire abstraction du phénomène linguistique. Or nos résultats montrent tout de même l'influence importante du système linguistique, en particulier dans la perception du 3e ton.

Le chinois étant une langue à ton lexical et monosyllabique, les variations de hauteur au sein d'un mot et d'une syllabe jouent un rôle prépondérant ; il n'est donc pas étonnant que les Chinois n'apprécient pas la hauteur relative entre les syllabes. Ce phénomène nous aidera peut-être à expliquer la production monotone de l'intonation française chez une partie des Chinois.

Conclusion

La démarche perceptive n'est pas la même chez les Français et chez les Chinois. Les Chinois sont influencés par le contexte linguistique ; par contre les Français assimilent davantage les tons à de la musique.

Notre première expérience a montré que la perception tonale des sujets français dépendait de la forme mélodique des tons : le ton 1 statique est mieux identifié que les tons 2, 3, 3' 4 à glissando. Nous avons également remarqué que le niveau de connaissance musicale des Français en solfège favorise l'identification des tons et une pratique du chant améliore la répétition.

Notre deuxième expérience a fait apparaître que les Chinois n'éprouvaient pas de grande difficulté à noter les tons par leurs numéros (1, 2, 3, 4) mais qu'ils avaient beaucoup plus de difficulté à reconnaître leur forme mélodique. Nous avons constaté également le rôle prépondérant du système linguistique, car la moitié de nos auditeurs a fait appel à des propriétés des éléments linguistiques dans la perception mélodique, en particulier dans la perception du 3e ton.

Références

(1) LHOTE, E., Des langues, des paysages sonores, A la découverte des paysages sonores des langues, Annales de l'Université de Franche-Comté, Belles Lettres, 1986, chapitre 1 (à paraître)
(2) CHAO, Y.R., Mandarin Primer, Harward University Press, Cambridge, 1948.
(3) HOMBERT, J.M., Perception of contour tones : An experimental investigation. Proceedings of the Ith Annual Meeting of the Berkeley Linguistics Society, 1975, 221-232.
(4) ROSSI, M., La perception de glissandos descendants dans les contours prosodiques, Phonetica 1978, 35, 11-40.
(5) PERON, M.H., CHI LEE, P.W., Le paysage chinois pour un francophone et le paysage français pour un sinophone, A la découverte du paysage sonore des langues, chapitre 3 (à paraître)
(6) KIRILOFF, C., On the auditory perception of tones in Mandarin, Phonetica, 1969, 20, 63-67.

EVALUATION DE DISTANCES EN UTILISANT DES SONS SYNTHETIQUES ET LA PERCEPTION HUMAINE

Haiyan Yé Denis Tuffelli

Institut de la Communication Parlée (LA. CNRS. No.368)
46, Av. Félix-Viallet 38031 Grenoble France

ABSTRACT

In this paper, we try to compare several distortion measures with the human's perception using synthetic sounds. Correlation and another measure of coherence is used. The goal of this research is to study the coherence between mathematical distortion measures and the human's perception. The results show there are some differences between them. But Itakura distortion measure and cepstral are the best in the case of our isolated vowels.

I. INTRODUCTION

La mesure de distances est très importante pour le traitement du signal. En analyse c'est une mesure d'erreur ou une mesure d'approximation (1). En synthèse ou codage c'est une mesure objective de qualité de sons ou une mesure de distortion de sons(2,3). En reconnaissance c'est une mesure sur laquelle se base la décision(4). Et en perception c'est un moyen de modéliser la dissimilarité perceptive(5). Donc suivant les différents cas, on utilise différents termes: distance; distortion; dissimilarité ou dissimilitude. Quelles soient les termes, les fonctions employées ne sont pas en général des métriques mathématiques qui vérifient les propriètés de symétrie ou d'inégalité triangulaire.

En général, il existe 2 sortes de distances: 1) un premier type de distance est défini par un critère mathématique sans utiliser des connaissances perceptives, par exemple: Itakura-Saito(1); Itakura (4); cepstral(6); rapport de vraisemblance(6) et Itakura-Saito pondéré(7) ... etc. 2) un autre type de distance est défini à partir de connaissances perceptives, par exemple: métrique de pente pondérée WSM(8); distance euclidienne de spectres en bandes critiques(9) et distance interspectrale avec notion de formants (10,11) ... etc. La première approche a une base mathématique solide mais sans contraintes perceptives. La deuxième approche prend en compte les connaissances en perception. Une caractéristique commune à peu prés générale de ces distances est que la comparaison (mise en correspondance) entre spectres Y et X se fait entre les valeurs ayant la même fréquence du type Y(f)-X(f) ou Y(f)/X(f). Si Y(f) et X(f) n'ont qu'un seul pic, la distance présentera une "saturation" quand les fréquences des pics s'éloigneront (à la limite la fonction distance sera elle-même un pic inversé).

Les relations de ces 2 sortes de distances avec la réalité de la perception n'ont pas été beaucoup étudiées. La première étude a été faite entre les distances et les seuils différentiels(12). Des études récentes ont été faites en vue de définir une nouvelle distance: une distance bien corrélée avec la distance phonétique perceptive (8), une distance interspectrale (10) ...

Une comparaison plus globale peut se faire aussi entre les matrices de confusion obtenues de 2 façons: avec des auditeurs puis avec un système de reconnaissance automatique (13).

Cet article va comparer les 2 sortes de distances précédentes avec les histogrammes d'une série de tests psychoacoustiques conçus spécialement dans ce but-là.

II. EVALUATION DE MESURES DE DISTORTION

A - Les Différentes Mesures de Distortion Testées

Nous allons décrire très rapidement les 12 mesures que nous avons comparées au moyen de nos tests:

* La mesure de distortion d'Itakura(4) est la mesure d'Itakura-Saito optimisée en gain qui fut à l'origine introduite comme une fonction d'erreur (mise en correspondance) pour l'estimation de modèles spectraux autorégressifs.

$$d_{ita}(x,x') = \log(\alpha/\alpha_m)$$

où α est une énergie résiduelle et α_m est l'énergie résiduelle minimale.

* La mesure de distortion cepstrale(6) est une approximation de la norme L_2 de la distortion spectrale logarithmique au moyen des N premiers termes d'un développement en série:

$$d_{cep}(x,x') = \sum_{k=-N}^{N} (c_k - c'_k)2$$

Deux variations de cette distance sont:

$$d_{kc}2(x,x') = \sum_{k=-N}^{N} (k * (c_k - c'_k))2$$

C'est une distance euclidienne des dérivées de phase du spectre.

$$d_{kc}1(x,x') = \sum_{k=-N}^{N} k * (c_k - c'_k)2$$

* Deux autres mesures de distortion, à priori mauvaises ont été testées ici: Une distance euclidienne sur des coefficients de prédiction linéaire et une distance euclidienne sur des coefficients d'autocorrélation. On a utilisé pour cela une chaine d'analyse à codage prédictif.

* La métrique de pente pondérée proposée par Klatt est une mesure de distortion avec des bases en perception (8).

$$d_{wsm}(x,x') = Ke|E-E'| + \sum_{i=1}^{Q} K(i)*[S(i)-S'(i)]^2$$

Ke et K(i) sont des coefficients pour lesquels nous avons pris: Ke=0 et K(i)=1 (suivant (14) les performances de reconnaissance de mots isolés sont meilleures avec ces valeurs). Ici on a pris aussi, pour des raisons de chaine d'analyse disponible,Q=18 (Cette valeur diffère sensiblement de celles utilisées par Klatt qui prenait de plus grandes valeurs pour Q).

*Une autre mesure de distortion avec des bases en perception a été proposée par Plomp(9). Plus tard elle fut utilisée par Carlson(15) et Blomberg(16).

$$d_{plm}(x,x') = (\sum_{i=1}^{Q} |L_i - L'_i|^p)^{1/p}$$

où les L_i représentent un spectre en bandes critiques (i est le numéro de bande) et où p a la valeur 1 ou 2.

* Nous proposons une distance

$$dprop(x,x') = \Sigma (dS^2(f) - dS(f+1)*dS(f))$$

où $dS(f) = S(f) - S'(f)$

On peut montrer par le calcul et par l'expérimentation que cette distance est presque équivalente à celle de Klatt ce qui veut dire qu'une distance du type Klatt n'est pas unique. En comparant les expressions de ces 2 distances on peut voir que la distance euclidienne sur la pente (Klatt) peut s'exprimer par la différence entre une distance euclidienne normale et l'autocorrélation de la différence entre spectres. Ce qui peut permettre une généralisation de la distance de Klatt en adoptant différents pas pour l'autocorrélation.

* Une autre mesure de distortion simple (appelée ici D_{hsl}) utilisant la pente spectrale est définie par une distance de Hamming sur un jeu de paramètres Fn. On a:

$$F_n = \begin{cases} 1 & \text{si } X(n+1) >= X(n) \text{ et } X(n+1) >= \text{seuil} \\ 0 & \text{autrement} \end{cases}$$

X(n) est un spectre lissé soit en échelle linéaire soit en échelle des Mel.

A notre avis, quelquefois certaines distances peuvent avoir des valeurs qui peuvent surprendre. Par exemple dans le cas de la dernière distance évoquée on peut démontrer qu'avec certains spectres cette distance est proportionnelle à l'écart entre la première résonance (pic) du premier spectre et la deuxième résonance du deuxième spectre !!

B - Le Problème de l'Evaluation

C'est un problème difficile et fondamental. Par exemple une méthode classique d'évaluation de différentes mesures de distortion est de les tester par le biais d'un système de reconnaissance de parole. Ainsi l'on peut juger des performances suivant les différents taux de reconnaissance obtenus. C'est une façon de procéder lourde: il faut un nombre de mots suffisant en entrée du système de reconnaissance, ce système n'est pas léger à mettre en oeuvre ... etc. De plus le choix des mots n'est pas évident. Mais même ce choix effectué, l'apprentissage n'est lui non plus pas facile à faire. En définitive c'est une technique couteuse en temps.

On cherche ici à utiliser des résultats déjà obtenus sur des voyelles (utilisation de voyelles représentatives) afin de pouvoir évaluer de manière peu couteuse différentes mesures. De plus nos tests permettront de "voir", au sens propre du terme, le comportement de ces différentes mesures de distortion.

C - Les Tests Psychoacoustiques Mis en Oeuvre

Deux sortes de tests ont été utilisées. Ils produisent des histogrammes qui peuvent être facilement comparés avec des courbes obtenues avec les mesures de distortion. Ils ont été menés sur des voyelles synthétiques isolées. Parmi toutes les paires possibles, 12 paires de voyelles ont été choisies. Ceci afin d'éviter que, dans le plan F1-F2 , une troisième voyelle puisse exister sur la droite reliant les deux voyelles de la paire testée. Une série de 11 sons a été synthétisée, pour chaque paire de voyelles, par une interpolation linéaire des formants entre 2 voyelles. Le synthétiseur est celui de Klatt et les données de formants sont celles de Mrayati(17) avec 120Hz de fréquence fondamentale.

Durant un test, un auditeur doit écouter les séries précédentes de sons et doit discriminer chaque son présenté entre 2 références avec un choix forcé.

Pour la première sorte de test ces 2 références sont des références phonétiques, c'est-à-dire que ce sont des étiquettes de voyelles (les 2 noms de la paire de voyelle utilisée); Donc ici, les sons références des 2 voyelles ne sont pas données à l'auditeur. On appellera cette sorte de test, un test X.

Pour la deuxième sorte de test les 2 références sont des références acoustiques. Avant d'entendre un son intermédiaire entre 2 voyelles, l'auditeur entend ces 2 voyelles. Des recommandations données en début de test lui demande de juger uniquement en fonction des sons références entendus. On appellera cette sorte de test, un test ABX.

Pour chaque sorte de test 12 histogrammes ont été traçés avec 12 auditeurs sur 132 sons (11 sons intermédiaires pour 12 paires).

En fait ce que les auditeurs réalisent ici est une mesure de similarité phonétique entre le son intermédiaire présenté et les 2 références.

L'auditeur reconnaitra un tel son s'il lui semble plus semblable à une référence qu'à une autre référence.

D - Les Courbes de Mesures de Distortion

Les mêmes signaux que précédemment ont été utilisés avec les mesures de distortion. Pour des raisons de facilité de comparaison nous calculons:

$$D_s(x,V1,V2) = d(x,V2)-d(x,V1)$$

où V1, V2 sont les 2 références et x est n'importe quel son de la série de sons synthétisés par interpolation linéaire entre V1, V2. d est une mesure de distortion.

Les évaluations des mesures de distortion sont faites par corrélation et pourcentage d'erreurs. Evaluations qui seront définies dans le prochain paragraphe.

III. RESULTATS EXPERIMENTAUX

A - Mesure de Corrélation Normalisée r

Elle est souvent utilisée pour comparer une mesure de distortion et la perception humaine.

Si x et y sont des vecteurs on a:

$r = \cos\theta = (x,y) / (||x||.||y||)$

B - Pourcentage d'Erreurs

* Pour la perception humaine, il y a une frontière statistique (la flèche dans la fig.1) entre 2 voyelles. Chaque auditeur peut faire quelques erreurs vis à vis de cette frontière. La moyenne de cette erreur pour tous les auditeurs est dénotée par E_h. Par exemple, un histogramme est présenté à la Fig 1, E_h est la somme de la région noircie.

* Pour les mesures de distortion, un pourcentage d'erreur E_{dis} est défini. Ce pourcentage est calculé par le rapport de 2 longueurs: la longueur de l'intervalle entre la frontière de la mesure de distortion (passage par zéro de la courbe) et la frontière statistique humaine, la longueur entre 2 références(voir Fig 2).

C - Quelques Figures

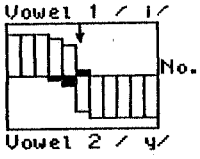


Fig 1.

Test ABX

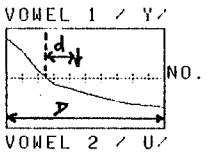
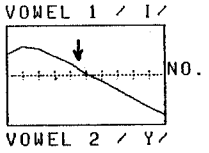
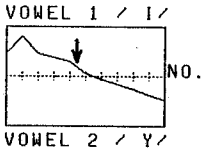


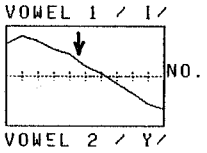
Fig 2. $E_{dis}=d/D$



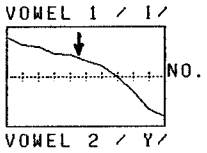
(a) Itakura



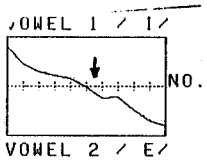
(b) Cepstral (Ck)



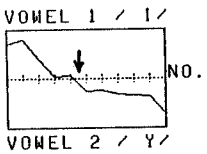
(c) $k^{1/2}$*Ck



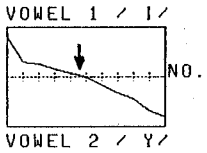
(d) k*Ck



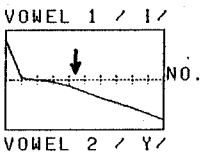
(e) Ri



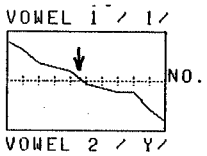
(f) Ak



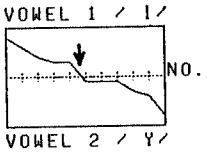
(g) Plomp(euclidienne)



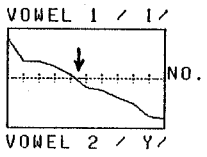
(h) Plomp(city-block)



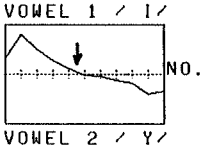
(i) hsl(linéaire)



(j) hsl(Mel)



(k) Dis proposée



(l) WSM

Fig 3. ↓ est la frontière statistique des tests

D - Quelques Résultats

Nous présentons maintenant une partie des résultats obtenus sur les corrélations et les pourcentages d'erreurs.Tous les résultats qui vont suivre sont des moyennes sur 12 tests. La corrélation a été calculée sur tous les points des 2 courbes: D_s et l'histogramme de la perception humaine.

Pour les tests X:

	Corrélation	Pourcentage
Auditeurs		9.6%
Itakura	0.86	10.1%
Cepstral	0.83	13.7%
$d_{kc}1$	0.85	15.1%
Plomp(city-block)	0.82	15.3%
Ri(euclid.)	0.77	17.0%
Plomp(euclid.)	0.80	17.8%
$d_{kc}2$	0.79	18.3%
Hamm. slop(lin.)	0.74	18.3%
Hamm. slop(Mel)	0.77	19.6%
WSM	0.70	21.0%
Ak(euclid.)	0.64	22.6%

Pour les tests ABX:

	Corrélation	Pourcentage
Auditeurs		8.8%
Cepstral	0.90	8.7%
Itakura	0.90	9.5%
$d_{kc}1$	0.91	9.5%
$d_{kc}2$	0.87	13.5%
Plomp(city-block)	0.86	14.7%
Ri(euclid.)	0.83	16.3%
Plomp(euclid.)	0.84	16.7%
dprop	0.81	19.0%
Hamm. slop(lin.)	0.76	19.6%
WSM	0.78	19.6%
Ak(euclid.)	0.71	19.6%
Hamm. slop(Mel)	0.71	20.0%

Un autre type de corrélation peut être calculé à partir des différentes frontières. Par exemple la corrélation entre les frontières obtenues par Itakura et Cepstral sur 12 tests est 0.99, elle correspond à un angle de 6.7°; la corrélation entre Itakura et Ri est 0.9, elle correspond à un angle de 25.8°.

Les résultats sur la distance WSM sont peut-être dûs à une application un peu différente de la méthode originale.

IV. PROPOSITIONS DE TRAITEMENTS SPECTRAUX AVEC ALIGNEMENT NON LINEAIRE

Avec les mesures de distortion précédentes " la comparaison de base" se fait souvent à la même fréquence pour le spectre référence Y et le spectre inconnu X. On retrouve, dans l'expression de telle mesure, des termes Y(f)-X(f) ... etc. Nous parlerons à ce propos d'alignement naturel et nous allons exprimer cela en termes de graphe. On définit un graphe dont les noeuds sont des doublets (Y(f1),X(f2)). Un doublet est un "point de comparaison" entre la composante de Y à f1 et celle de X à f2. Avec l'alignement naturel on suit un chemin diagonal (Fig.4). La définition du graphe doit être complétée par la définition d'arcs: (Y(f),X(f))—(Y(f+1),X(f+1)). Enfin ce graphe doit être valué par des pondérations associées aux noeuds c(Y(f),X(f)). c est une comparaison élémentaire entre Y(f) et X(f). La mesure de distortion d s'exprime comme une somme:

$$d = \sum_{f=1}^{N} c(Y(f),X(f))$$

Cett technique d'alignement est simple (nous avons vu qu'il peut y avoir des problèmes de saturation) mais efficace en temps de calcul.

On peut penser à d'autres alignements plus compliqués: Par exemple Paliwal(18) utilise un alignement avec programmation dynamique inspiré de l'alignement classique temporel en reconnaissance et constate que cette technique compliqué n'est pas efficace en taux de reconnaissance. Ce résultat est intéressant car il nous parait logique. Les alignements temporel et fréquentiel sont deux problèmes de nature différente. Dans le premier cas on cherche à minimiser la variation normale de durée (une voyelle isolée de 300 ou 400 ms reste toujours la même, l'écart en durée doit être ignoré tout au moins en premier approximation car la durée peut être aussi un indice de discrimination).

En revanche pour une comparaison spectrale il nous semble que la proposition précédente s'inverse et que tout écart est à priori significatif. Prenons l'exemple extrême et caricatural de 2 spectres constitués chacun d'une raie d'amplitude A en f1 et f2 (Fig.5). Il est clair que le chemin optimal passe par le point P et que la distance interspectre résultante est nulle quel que soit f1 et f2! Une distance utilisant la diagonale n'aurait pas donné de bons résultats non plus.

Nous pensons que l'écart entre 2 spectres ne doit pas se mesurer principalement par une mesure de distortion le long d'un chemin mais surtout par la forme du chemin optimal g de coïncidence. On peut caractériser cette forme par l'aire située entre g et la diagonale Blomberg(1986). Ceci ne nous parait pas souhaitable: Avec l'exemple précédent des solutions équivalentes (mais d'aires différentes) sont possibles. Toutes ces solutions passent par le point P (où Y et X sont non nuls). Nous proposons donc de mesurer la distortion par la longeur l de P à la diagonale (l est proportionelle à f1-f2) mais pondérée d'une manière variable par une fonction ß des quantités Y(f2) et X(f1) donc de la forme ß*l. En généralisant ce procédé à l'ensemble du chemin g on obtient une mesure de distortion "proportionelle à l'écart entre formants" sans que l'on soit astreint à la détection des formants.

Notre deuxième proposition concerne la création d'un spectre "milieu" entre 2 spectres Y et X. Ce problème nous semble lié à celui des mesures de distortion. Avec l'exemple de la Fig.5 une moyenne sans précaution conduit à un spectre moyen à 2 raies! Une projection sur la diagonale (fréquence fm à définir) nous semble préférable car moins destructive de la forme.

Notre troisième proposition concerne la comparaison avec des résultats de perception: par exemple l'approximation d'un spectre avec plusieurs formants par un spectre à deux formants peut probablement être étudié à travers l'étude du chemin optimal.

Enfin notre quatrième proposition concerne l'étude de l'adaptation au locuteur en considérant des points d'arrivés de chemin non situés sur la diagonale (et propres à des classes de locuteur).

Nous allons donc mettre en oeuvre maintenant un programme de recherche pour explorer les 4 voies précédentes. Les premières études seront réalisées avec les sons synthétiques précédents.

Eventuellement une réalisation électronique peut être envisagée (structure systolique).

V. CONCLUSIONS

Les principales mesures de distortion mathématiques obtiennent de meilleurs résultats que les mesures basées sur la perception mais les tests que nous avons menés sont favorables aux mesures mathématiques (les sons varient seulement par des décalages de formants). Comme on s'y attendait les coefficients Ak ont de mauvais résultats. Quelque fois de très mauvaises frontières sont obtenues qui sont difficiles à expliquer. Une très grande corrélation entre les mesures d'Itakura et Cepstral est observée. A notre avis une distance idéale devrait pouvoir mesurer les distortions des formants sans les détecter car cela est difficile. Il a été prouvé qu'une distance utilisant un alignement fréquentiel (avec une programmation dynamique) donnait de mauvais résultats. Cela nous senble normal, nous pensons que ce n'est pas la distance minimale qu'il faut utiliser mais la distortion du chemin d'alignement qu'il faut mesurer.

Nous avons fait dans cette étude plusieurs propositions nouvelles (généralisation de la distance de Klatt et traitements spectraux non-linéaires).

Le choix le plus difficile, dans ce travail, est l'ensemble des formants des sons références. Cette difficulté est inhérente à tout apprentissage. Plutôt que d'effectuer, à l'image des systèmes de reconnaissance de la parole, un apprentissage long, couteux et aléatoire nous avons préféré utiliser des références existantes considérées comme représentatives des voyelles françaises. D'une manière surprenante d'une part elles sont très bien adaptées aux mesures de distortion d'Itakura et cepstral, d'autre part elles ne donnent pas de trop mauvais résultats avec les coefficients d'autocorrélation.

REFERENCES

1. F.ITAKURA, S.SAITO, "A Statistic Method for Estimation of Speech Spectral Density and Formant Frequencies" Elec. and Commu. in Japan, Vol. 53-A, No.1, 1970, pp36-43
2. B.H.JUANG, "On Using the Itakura-Saito Measures for Speech Coder Performance Evaluation" AT&T Bell Lab. Tech. J. Vol.63, No.8, 1984, pp1477-1498
3. R.M.GRAY, A.BUZO, A.H.GRAY, Y.MATSUYAMA, "Distortion Measure for Speech Processing" IEEE Trans. ASSP-28, No.4, pp367-376, 1980
4. F.ITAKURA, "Minimum Prediction Residual Principle Applied to Speech Recognition" IEEE ASSP-23, No.1, 67-72, 1975
5. R.A.W.BLADON, B.LINBLOM, "Modeling the Judgment of Vowel Quality Differences" JASA 69(5), 1981, pp1414-1422
6. A.GRAY, J.D.MARKEL "Distance Measures for Speech Processing" IEEE Trans. ASSP-24, No.5 1981 pp380-391
7. P.L.CHU,D.G.MESSERSCHMITT, "A Frequency Weighted Itakura-Saito Spectral Distance" IEEE Trans.ASPP-30, No.4, pp545-560, 1982
8. D.H.KLATT "Prediction of Perceived Phonetic Distance from Critical Band Spectra: a first step" ICASSP pp1278-1281, 1982
9. R.PLOMP "Timbre as a Multidimentional Attribute of Complex Tones" FREQUENCY ANALYSIS AND PERIODICITY DETECTION IN HEARING Ed. PLOMP, 1970
10. M.J.CARATY, X.RODET, "Distance Interspectrale à Critères Perceptifs" 14eme JEP, 1985, pp87-90
11. M.HUNT, "A Robust Formant-Based Speech Spectrum Comparison Measure" ICASSP 1985, 19.9
12. R.VISWANATHAN, J.MAKOUL, W.RUSSELL "Towards Perceptually Consistent Measure of Spectral Distance" ICASSP pp485-489, 1976
13. M.ESKENAZI, J.S.LIENARD, "Recognition of Static State French Sounds Pronounced by Several Speakers; Comparison of Human Performance and an Automatic Recognition Algorithm" Speech Communication 2(1983) pp173-177
14. N.NOCERINO, F.K.SOONG, L.R.RABINER, D.H.KLATT, "Comparison Study of Several Distortion Measures for Speech Recognition" Speech Communication 4(1985) pp317-331
15. R.CARLSON, B.GRANSTROM, "Model Prediction of Vowel Dissimilarity" STL-QPSR 3-4, 1979, pp84-104
16. M.BLOMBERG, R.CARLSON, K.ELENIUS, B.GRANSTROM, "Auditory Models and Isolated Word Recognition" STL-QPSR 4, 1984, pp1-15
17. M.MRAYATI, Contribution aux Etudes sur la Production de la Parole, Thèse d'Etat, INPG 1976
18. K.K.PALIWAL, W.A.AINSWORTH, "Dynamic Frequency Warping for Speaker Adaptation in Automatic Speech Recognition" J. Phonetics (1985) 13, 123-134
19. M.BLOMBERG, K.ELENIUS "Nonlinear Frequency Warp for Speech Recognition" à paraître à ICASSP-86

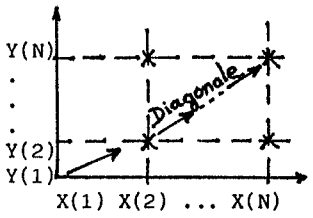


Fig. 4

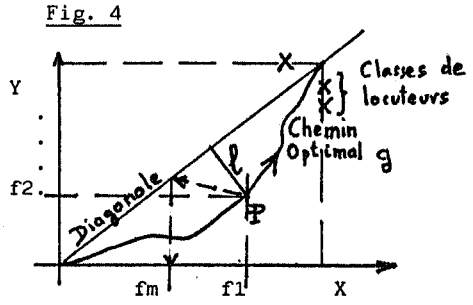


Fig. 5

RELATIVITE DU POIDS DES VARIABLES DANS L'ANALYSE EN COMPOSANTES PRINCIPALES

C. Gélinas-Chebat*, M. Rossi**, J.L. Chandon ***

*UQAM, **Institut de Phonétique d'Aix-en Provence, ***IAE d'Aix-en-Provence

ABSTRACT

Multidimensional scaling analysis in phonetics, was used for the first time by Shepard(1). His objective was to reconstruct a psychological pattern of speech perception. But this kind of statistical analysis even if many authors have proposed modification, obliges to estimate similarities between two pairs of stimuli. In fact, in phonetics we know that so often reality is really quite different. The distance between two points in an Euclidian space can not describe adequately the similarity or dissymilarity between two variables in a phonetic perception task. So researchers have to consider very carefully this fact in the interpretation of the factors in a multidimensional scaling analysis. This paper will describe succinctly the methodological problem and we will try to propose an acceptable statistical solution .

RESUME

Shepard(1) a été le premier à utiliser des analyses statistiques multidimensionnelles dans l'interprétation de données phonétiques. Or un tel traitement statistique, comme le souligne Goldstein (2) bien qu'extrêmement révélateur, introduit un certain biais par l'obligation d'établir une estimation de la similarité (ou dissimilarité) entre deux paires de stimuli dans le calcul de la distance. Après avoir souligné les difficultés d'interprétation des données phonétiques d'une matrice de confusion par des analyses multidimensionnelles, nous proposerons une méthode statistique alternative.

ETAT DE LA QUESTION:

Les stratégies perceptives des auditeurs suscitent encore aujourd'hui de nombreuses interrogations. L'étude menée par Miller et Nicely (3) est encore aujourd'hui tout à fait pertinente et est encore largement utilisée comme source de référence. Il s'agit d'une étude des confusions de perception de mots émis, auprès d'une population d'individus auditivement normaux .

Les confusions, provoquées par différentes conditions de bruit et de filtrage ont été recueillis sous forme de matrices de confusion où pour chaque mot émis il est possible de voir le nombre de bonnes réponses et le nombre de fois où il est confondu avec un autre mot.

Ces auteurs ont tenté de dégager un certain nombre d'informations relatives aux traits distinctifs par une mesure de l'intelligibilité de la covariance c'est-à dire de l'entropie. Il s'agit en fait dans le cadre de la théorie mathématique de la communication, la théorie cybernétique, de la mesure de la dispersion de l'information comme indice du pourcentage d'information transmise pour les cinq traits distinctifs de voisement, de nasalité, de stridence, de duré et de lieu d'articulation.

Cette mesure de l'entropie ne fournit malheureusement pas d'information sur les "erreurs" de perception. Il n'y a pas de réponses attendue favorisée, toutes les possibilités de réponses étant équiprobables. La véritable dimension qu'il est possible de dégager est qu'un mot émis, caractéristique d'un trait distinctif donné, présente des réponses confuses ou organisées sous telle condition expérimentale. Une mesure du degré d'organisation peut aussi se faire par un simple calcul du Khi-deux. Mais, dans un cas comme dans l'autre, il n'est pas possible de préciser quantitativement quel mot parmi les mauvaises réponses est choisi de préférence par les auditeurs. Shepard (1) a tenté de contourner ce problème en proposant des analyses factorielles multidimensionnelles à partir des résultats de Miller et Nicely (3).

Ces analyses présentent d'énormes attraits. Bouroche et Saporta (4) en décrivent d'ailleurs certains avantages par opposition à des traitements statistiques plus traditionnels. Ces analyses statistiques permettent de manipuler de très importants fichiers de données, de n'avoir à poser qu'un minimum d'hypothèses à priori et donc de laisser parler les donnés par elles-mêmes et enfin, d'obtenir une vision globale des résultats c'est-à-dire de l'information.

La "Multidimensional Scaling Analysis" (MDS), comme nous l'a proposé Shepard (1), est l'une de ces méthodes qui cherche à réduire l'information d'une matrice de confusion à n éléments en situant dans un espace à n-1 dimensions tous les éléments de cette matrice.

Outre le problème généralement soulevé d'interprétation des dimensions, liée au poids des variables sur les différents facteurs, la base de calcul de cette méthode nécessite l'uniformisation des distances entre deux points de la matrice c'est à dire que la distance entre ces deux points doit être la même:

"we require that the proximity between i and j be the same in **both** directions" (Shepard, 1 ,p.72) Pour mesurer cette similitude entre deux points de la matrice, il déterminait une moyenne de distance entre ces deux points en divisant le nombre total de confusions de i et de j par le nombre total de bonnes réponses des i et j, c'est-à-dire:

$$S_{ij} = \frac{P_{ij} + P_{ji}}{P_{ii} + P_{jj}}$$

Mais, comme la figure 1 suivante l' illustre, le passage des donnés d'origine de Miller et Nicely (3), à la mesure déterminée par Shepard (1) oblige à perdre une partie de l'information de nos données de base.

	p	t	k		p	t	k
p	697	171	335	p	---		
t	190	882	193	t	0,229	----	----
k	278	194	721	k	0,432	0,241	----

figure 1: Exemple du passage des données de Miller et Nicely(4) aux donnés de Shepard (1)

Kruskal (5) a préféré introduire la notion de stress afin de résoudre cette difficulté. Mais il considérait qu'en fait "our procédure is substantially faster than Shepard's, but this probably reflects programming improvements rather than anything more fundamental" (Kruskal, 5, p.25)

Aussi Caroll et Chang (6) ont repris essentiellement la méthode qu'avait proposé Shepard (1), mais en apportant une certaine modification au calcul de la distance entre les points. Il s'agit d'une distance euclidienne "modifiée" qui se décrit ainsi

$$d^{i}_{jk} = \sqrt{\sum_{t=1}^{r} w_{it} (x_{jt} - x_{kt})^2}$$

où w est un indice de poids. Cet indice améliore en effet la mesure de Shepard (1) mais ne résout pas avec satisfaction, à notre avis, le problème d'établir une "moyenne" entre deux cases.

Pols et Stoop (7) au congrès international de phonétique à Utrecht en 1983, avait préféré la méthode dite TUCKALS de Kroonenberg et De Leeuw (8), qui traduisait mieux la réalité phonétique de consonnes du néerlandais que la méthode dite INDSCAL ne pouvait le faire. Cette méthode TUCKALS, "a mean-squared loss function is used to minimize the difference between the low-dimensional model and the original data" (Kroonenberg et De Leeuw, 8, p. 96) tente de diminuer les différences, sans plus.

Mais, il n'est pas toujours possible, à l'exemple de Lonchamp(9), d'analyser quatre banques de données pour les comparer aux siennes afin de voir quelle méthode d'analyse multidimensionnelle métrique c'est-à-dire de type INDSCAL ou non-métrique c'est-à-dire de type KRUSKAL, permet le mieux de saisir l'espace perceptif des sons vocaliques. Il s'agit d'une contribution extrêmement intéressante où l'auteur admet que "les progrès de nos connaissances dans le domaine de la perception des sons du langage ne sont pas si rapides qu'on puisse négliger une source potentielle d'informations" (Lonchamp, 9, p.417)

NOTRE CONTRIBUTION:

1: La collecte des données:

Les résultats d'une épreuve de perception de mots, filtrés et non-filtrés, (Gélinas-Chebat, 10), ont été recueillis sous forme de matrices d'occurence telles que définis par Chandon et Pinson (11). La condition sans-filtre représente les résultats d'un groupe contrôle (6 matrices) alors que 60 matrices représentent les conditions de perception des mots sous l'effet de dix filtres différents.

2: Le traitement statistique:

Comme nous voulons à la fois rendre compte des bonnes et mauvaises réponses de chaque matrice, nous avons procédé au calcul d'un indice de confusion qui se décrit ainsi:

$$t = \frac{Kn - \sum_{i=1}^{k} n_{ii}}{Kn}$$

où kn est le nombre total de réponses et n_{ij} correspond à la diagonale pour l'ensemble de la matrice.

Le taux de confusion pour chaque ligne, c'est à dire pour les mots à l'émission, et chaque colonne, c'est à dire pour les mots à la réception, peut se traduire ainsi:

$$t_i = \frac{n_{i\bullet} - n_{ii}}{n_{i\bullet}} \quad \text{et} \quad t'_i = \frac{n_{\bullet i} - n_{ii}}{n_{\bullet i}}$$

C'est à partir de cet indice du taux de confusion que nous avons procédé à différentes analyses dont des analyses de variances (MANOVA du programme SPSS) pour dégager le poids de différents traits distinctifs sur la quantité d'information transmise, de même qu'une hiérarchie de certains de ces traits.

3. RESULTATS

Le tableau 1 suivant résume les différentes analyses de variance multiples (MANOVA): il indique la valeur des F dans la relation entre les variables dépendantes, à savoir les taux de confusion à l'émission et à la réception, et les variables indépendantes, à savoir les dimensions continu/discontinu (occlusive/constrictive), voisé/non-voisé (sourde/sonore), grave/aigüe, et enfin l'influence du voisinage de la voyelle /A-I-U/.

TABLEAU 1: Résumé des analyses de variances

	OCC/CONS	SD/SO	G/A/C[1]	A/I/U
Emission	7.62	n.s.	6.40	2.19
Réception	9.25	n.s.	3.44	2.73

4. DISCUSSION ET CONCLUSION

Malgré la critique souvent émise à propos des analyses factorielles où il s'agit "d'un modèle a priori car on postule l'existence de facteurs communs et on cherche à vérifier le modèle à partir des données expérimentales" (Bouroche et Saporta, 4, p.26) nous pensons qu'il est justifiable dans un cadre phonétique si nous prenons en considération les différents caractères possibles de nos données. Ainsi, nous avons pu constater, que certains traits contribuent d'avantage à la variance. Le trait distinctif relatif à la dimension occlusif/constrictif est celui qui provoque le plus de dispersion des réponses aussi bien à la réception qu'à l'émission. De plus, il est possible de noter une dispersion plus marquée des réponses lorsque la consonnne initiale de notre expérimentation est une occlusive que pour les mots dont l'initiale est constrictive. Cette dispersion est d'autant plus marquée sous la dimension émission que réception. Ce comportement des occlusives trouverait son explication par le fait qu'elles sont plus instables acoustiquement et donc beaucoup plus fragiles aux manipulations du message par différents procédés phonétiques.

Le trait de gravité qui vient en deuxième lieu nous permet de remarquer que la valeur du F est beaucoup plus importante à l'émission qu'à la réception. Encore une fois, il semble se dégager un comportement distinct des consonnes lorsqu'elles ont un rôle à jouer sous la dimension réception et émission. Certaines consonnes semblent facilement confondues mais, ne semblent pas particulièrement attirer les mauvaises réponses. Les voyelles contribuent significativement à la variance aussi bien à la réception qu'à l'émission mais dans une proportion moins marquée que pour les deux autres traits déjà identifiés. Quant au trait de voisement, il n'a pas d'effets significatifs, ni à l'émission ni à la réception.

Cette analyse permet d'établir la hiérarchie suivante des traits que nous avons étudiés selon leur contribution décroissante à la confusion: continu/discontinu > grave/aigu > voisé/non-voisé. Soulignons que cette hiérarchie est stable aussi bien sous la dimension émission que réception. Ainsi la dimension continu/discontinu ou occlusif/constrictif est extrêmement importante pour expliquer la confusion. Cette contribution est plus importante que ne semble l'être la contribution du lieu d'articulation, lui-même plus important que la contribution du voisement.

[1] Nous employons le terme de "compact" pour décrire les consonnes pour lesquelles l'opposition grave/aigu est redondante.

L'intérêt des recherches en perception est, entre autre, de contribuer à la théorie des traits distinctifs telle que Jakobson, Fant et Halle (12)l' ont élaborer pour décrire `les sons de la parole aussi bien au plan articulatoire, qu'acoustique et perceptuel. Comme Rossi (13) le fait remarquer "une théorie des traits doit s'insérer dans un modèle de la perception du langage" (Rossi, 13, p.82).

Il s'agit d'une étude exploratoire, où pour le moment nos résultats permettent de proposer une alternative aux analyses classiques multidimensionnelles dans le but de libérer l'analyse phonétique des contraintes statistiques artificielles, imposées par ces méthodes.

Notre méthode permet de souligner le poids relatif de chacun des traits distinctifs ce qui n'est pas nécessairement réalisable avec des analyses multidimensionnelles où l'interprétation des facteurs est laissée aux chercheurs. Les résultats que l'on obtient par la méthode que nous avons proposée sont d'autant plus fiables qu'ils sont exempts d'artifices statistiques.

REFERENCES

(1) R.N. Shepard, "Psychological representation of speech sounds" in E.E. David et P.B. Denes, ed., Human Communication: A Unified View, New York, McGraw-Hill Book Company, Ch.4, pp.67-113, 1972.

(2) L. Goldstein, "Categorical features in speech perception and production", JASA, vol.67 no.4, pp.1336-1348, 1980.

(3) G. A. Miller et P.E. Nicely, "An Analysis of perceptual confusions among some english consonants", JASA, vol. 27, no. 2, pp.338-352, 1955.

(4) J.M. Bouroche et G. Saporta, "L'analyse des données", Pour la science, ed. française de: Scientific American, pp. 23-35, 1978.

(5) J.B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", PSYCHOMETRIKA, vol 29 no. 1, pp. 1-27, 1964.

(6) J.D. Caroll et J.J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition", PSYCHOMETRIKA, vol. 35, pp. 383-419, 1970.

(7) L.C.W. Pols et I. Stoop, "Confusion between dutch consonants under various conditions of noise and reverberation" in Van den Broecke et Van Heuven, ed., Proceedings of the Tenth International Congress of Phonetic Sciences, Dordrecht, Foris publications, pp.455-458,1984

(8) P. M. Kroonenberg et J. De Leeuw, "Principal component analysis of three-mode data by means of alternating least squares algorithms", PSYCHOMETRIKA, vol. 45, pp. 69-97, 1980.

(9) F. Lonchamp, " Analyse multidimensionnelle de l'espace perceptif des sons vocaliques" 12e journées d'étude sur la parole, Montréal, pp. 406-418, 1981.

(10) C. Gélinas-Chebat, "Les propriétés acoustiques et les indices perceptuels des traits de lieu d'articulation chez l'auditeur normal et le mal-entendant" Thèse de 3e cycle présentée à l'université d'Aix-Marseille I, 397p. 1983.

(11) J.L. Chandon et S. Pinson, Analyse Typologique théories et applications, Paris, Masson, 254p.

(12) R. Jakobson, G.M. Fant, M.Halle, "Preleminaries to speech analysis. The distinctive features and their correlates, 1er éd., Cambridge, M.I.T. Press, 1951.

(13) M. Rossi, "Les traits acoustiques" La Linguistique, vol.13, no.1, pp. 63-82, 1977.

MODELE DE PREDICTION DU "DEUXIEME FORMANT EFFECTIF" F'2 ET APPLICATION A L'ETUDE DE LA LABIALITE DES VOYELLES AVANT DU FRANCAIS

M. Mantakas, J.L. Schwartz, P. Escudier

Institut de la Communication Parlée - Grenoble

ABSTRACT

A new formula for the estimation of vowel "effective second formant" F'2 is presented and tested, incorporating the critical 3.5 Bark-distance and based on formant frequencies. It is supposed to provide a way of transforming the acoustic space (F1, F2, F3, F4) into a "perceptual space" (F1, F'2). A first test of the ability of this formula to provide an efficient parameter for vowel classification is proposed for the study of labiality in French front vowels.

INTRODUCTION

Le meilleur modèle disponible jusqu'à maintenant pour l'estimation du "second formant effectif" F'2 des voyelles est celui proposé par Bladon et Fant en 1978 [1]. Selon ce modèle, F'2 est calculé implicitement comme la valeur moyenne des fréquences formantiques F2, F3 et F4, pondérées par leurs puissances spectrales, ce qui constitue son principal intérêt vis-à-vis d'autres modèles plus ou moins ad-hoc ([2], [3]).

Pourtant, la formule de Bladon et Fant est mise en cause pour deux raisons principales. D'une part, cette formule repose sur un premier regroupement de F3 et F4, remplacés par leur moyenne géométrique, avant leur interaction avec F2. Ceci résulte en une estimation peu claire de la puissance spectrale du groupe F3-F4, et conduit à l'introduction d'un paramètre K, dont le choix s'avère à la fois très difficile a priori et souvent critique [4]. D'autre part, comme Bladon [5] l'a montré ultérieurement, ce modèle prédit des valeurs de F'2 identiques pour des paires de voyelles phonétiquement très distinctes. Ces paires de voyelles, précisément construites pour montrer les limitations du modèle proposé, se caractérisent par des valeurs de F1 et F4 identiques, et de F2 et F3 très différentes. C'est précisément le regroupement F3-F4 qui permet, dans le cas où F2 est bas, de "tirer" F'2 fortement vers le haut.

Bien que la logique du modèle de Bladon et Fant soit celle du "centre de gravité" spectral, il n'incorpore pourtant pas la distance critique de 3.5 Bark associée à ce phénomène depuis 1979 par Chistovich et son équipe [6]. Or l'introduction de cette distance critique dans le modèle, comme Bladon lui-même le suggère [5], peut, nous le verrons, éliminer le second problème et aider à clarifier le premier.

LE NOUVEAU MODELE PROPOSE

Deux hypothèses fondamentales sont à base du travail suivant. D'abord, nous supposons que F'2 recouvre une réalité perceptive, et constitue donc un paramètre perceptif des voyelles, effectivement mesuré par un mécanisme de traitement spectral "câblé" dans le système auditif, et utilisé pour les décisions phonémiques. Il existe d'autres hypothèses, qui peuvent conduire à proposer que F'2 ne recouvre aucun mécanisme réel de traitement transformant un ensemble de paramètres (F1, F2, F3, F4) en un ensemble réduit (F1, F'2) : nous nous attachons à les écarter dans un autre travail présenté à ce congrès [7]. La deuxième hypothèse est que F'2 et "centre de gravité" sont deux manifestations d'un mécanisme unique. En effet, les expériences d'ajustement de type "centre de gravité" ou F'2 sont tout à fait de même nature.

Sur ces bases, un premier modèle de traitement global du spectre par intégration à large bande (3.5 Barks) a été élaboré et testé [8]. Nous proposons ici une formule directement déduite des propriétés de ce modèle, et qui s'articule sur les points suivants.

1. Cette formule doit pouvoir opérer sur les seules caractéristiques formantiques. Nous reviendrons plus loin sur les avantages et inconvénients de ce choix. Les principaux corrélats de F'2 sont les fréquences formantiques F2, F3, F4 et parmi elles, F2 joue un rôle prédominant. Les puissances spectrales correspondantes n'interviennent qu'au second ordre. Le premier formant n'intervient que par son influence sur les niveaux des autres formants.

2. Les fréquences formantiques sont exprimées en Bark et les puissances spectrales en décibels, unités "naturelles" de la perception [9]. En ce qui concerne les puissances spectrales, la conversion logarithmique est faite avec un niveau référence de 0 dB SPL. Dans tout ce qui suit, nous emploierons

la notation Z2, Z3, Z4 et Z'2 pour les fréquences F2, F3, F4 et F'2 exprimées en Bark et I2, I3, I4 pour les puissances spectrales en décibels.

3. Les fréquences F3 et F4 sont comparées à F2. Si leur distance à F2 est inférieure à la distance critique de 3.5 Bark, elles sont intégrées avec F2, sinon F'2 est égal à F2. Pour éviter des discontinuités brutales de prédiction autour de la distance critique, pour des raisons pratiques - problèmes de mesure des formants - autant que théoriques - légères variations éventuelles de la distance critique le long de l'echelle de fréquences (voir [6]) - nous introduisons les coefficients A_i, i=3,4, tels que:

$$A_i = \begin{cases} 1 & \text{pour } Z_i - Z2 \leq 3.0 \text{ Bark} \\ 0 & \text{pour } Z_i - Z2 \geq 3.5 \text{ Bark} \\ \text{variant linéairement entre 0 et 1 entre ces deux limites.} \end{cases}$$

4. Une modification de la logique d'intégration simple est introduite pour tenir compte de la particularité des voyelles avant non arrondies. Pour ces voyelles, la proximité de F3 à F4 plutôt qu'à F2 crée une proéminence spectrale caractérisée plutôt par le groupe F3-F4 que F2-F3-F4. Donc, dans ce cas, la formule n'intègre que F3 et F4 (voir plus bas). Cette idée est supportée par les résultats de l'expérience d'ajustement de F'2 dans la région [i]-[y], effectuée par Carlson et al. [10], et présentant un saut brusque des valeurs de F'2 selon la proximité de F3 à F2 ou F4. L'importance de cette frontière est reconfirmée par Schwartz et al. [7], cette fois en plus pour des voyelles de type [e]-[ø]. De plus, le modèle de traitement "global" du spectre par intégration à large bande reproduit ce basculement [8].

5. Les niveaux I2, I3, I4 sont soit mesurés, soit calculés d'après le modèle série tout-pôles de Fant [11], après avoir estimé les bandes passantes selon un modèle de dépendance à partir des fréquences formantiques (par exemple, formules (34), (35) dans Fant [12]).

6. On applique une correction isosonique pour convertir les décibels I_i en phones L_i, d'après les courbes isosoniques classiques (voir [9]).

7. Une première étude montre une surestimation légère mais systématique des valeurs de F'2 prédites, ou, en d'autres termes, un rôle quantitatif de F2 supérieur à ce que l'on attendrait d'une simple pondération de F2, F3 et F4. Ceci peut être produit, dans le système auditif, par les phénomènes de suppression à deux tons, de nombreuses études montrant les effets suppresseurs d'un formant sur les formants de fréquences plus élevées ([13], [14]). Faute de connaissances sur l'aspect quantitatif de ces phénomènes pour des formes spectrales si complexes, nous nous contentons de multiplier les niveaux L3 et L4 par les coefficient a3 et a4, dont les valeurs sont fixées respectivement à 0.7 et 0.6 .

La nouvelle formule proposée est la suivante:

- Si Z3 - Z2 supérieure à la distance critique
  Z'2 = Z2
- Si Z3 - Z2 inférieure à la distance critique et Z4 - Z2 supérieure, alors
  Z'2 = (L2·Z2 + A3·a3·L3·Z3) / (L2 + A3·a3·L3)
- Si Z4 - Z2 inférieure à la distance critique, on distingue deux cas:
  1/ Si Z3 - Z2 ≤ Z4 - Z3, alors
  Z'2 = (L2·Z2 + A3·a3·L3·Z3) / (L2 + A3·a3·L3)
  2/ Si Z3 - Z2 > Z4 - Z3, alors
  Z'2 = (A3·a3·L3·Z3 + A4·a4·L4·Z4) / (A3·a3·L3 + A4·a4·L4)

La formule peut s'écrire sous la forme plus compacte suivante:

$$Z'2 = \frac{C2 \cdot L2 \cdot Z2 + A3 \cdot a3 \cdot L3 \cdot Z3 + C4 \cdot a4 \cdot A4 \cdot L4 \cdot Z4}{C2 \cdot L2 + A3 \cdot a3 \cdot L3 + C4 \cdot A4 \cdot a4 \cdot L4}$$

où A_i sont définis ci-dessus et C_i comme suivant:

$C4 = 1 - C2$

et

$$C2 = \begin{cases} 1 & \text{si } Z4 - Z2 > 3.5 \text{ Bark} \\ f((G - Z3 + (3.5 - 3.0)/2) / (3.5 - 3.0)) & \text{si } Z4 - Z2 \leq 3.5 \text{ Bark} \end{cases}$$

avec

$G = (Z2 + Z4) / 2$

et

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x > 1 \\ x & \text{si } 0 \leq x \leq 1 \end{cases}$$

TEST DE L'EFFICACITE DE PREDICTION AVANTAGES ET INCONVENIENTS DU MODELE

Nous avons testé notre modèle sur les corpus expérimentaux des deux principales expériences de perception dans le domaine, comprenant l'un les 9 voyelles longues du suédois (Carlson et al. [10]) et l'autre les 18 voyelles cardinales (Bladon et al. [1]). Nous reproduisons les valeurs des formants et de F'2 pour ces 2 corpus dans les tables 1 et 2.

Les valeurs de F'2 prédites par notre modèle et par celui de Bladon et Fant [1] apparaissent sur les colonnes 6 et 7. Pour ce dernier modèle nous avons choisi la valeur optimale du paramètre K obtenue "a posteriori" par une minimisation de l'erreur totale sur chaque ensemble de voyelles [4]: K= 0.5824 x F2 pour le premier et 0.4909 x F2 pour le second (l'erreur totale est définie par la somme des erreurs absolues, en Barks: $\sum |Z'2\text{ajusté} - Z'2\text{calculé}|$). Les deux dernières colonnes donnent l'erreur de prédiction en Bark: Z'2ajusté - Z'2calculé). Les voyelles sont organisées selon un critère de Z'2 - Z2 croissant, qui représente en général le passage des voyelles arrière vers les voyelles avant.

Il faut tout d'abord remarquer que les voyelles [ɒ] et [ʉ] du deuxième ensemble ont des valeurs de F'2 ajusté nettement inférieures à celles du F2 correspondant. C'est également le cas de 5 autres voyelles du même ensemble, mais dans une marge de 0.4 Bark qui semble raisonnable.

	F1	F2	F3	F4	F'2a	F'2c NM	F'2c BF	Z'2a-Z'2c NM	Z'2a-Z'2c BF
u	310	730	2250	3300	730	730	730	0.00	0.00
o	400	710	2460	3150	720	710	711	0.07	0.06
ø	360	1690	2200	3390	1720	1879	1737	-0.59	-0.07
ɑ	580	940	2480	3290	960	940	944	0.12	0.10
y	255	1930	2420	3300	2010	2119	2156	-0.35	-0.47
ɯ	280	1630	2140	3310	1730	1817	1674	-0.33	0.22
e	375	2060	2560	3400	2370	2288	2370	0.23	0.00
ae	605	1550	2450	3400	1960	1835	1638	0.44	1.20
i	255	2065	2960	3400	3210	3079	3064	0.26	0.30

table 1
Test de la performance du modèle sur l'ensemble de voyelles de syntèse de Carlson et al. [10]
Notation:
F'2a : valeur ajustée du F'2 (réf. précedente)
F'2c : " estimée " "
Z'2a-Z'2c : erreur de prédiction en Bark
NM : nouveau modèle proposé
BF : modèle de Bladon et Fant [1]

	F1	F2	F3	F4	F'2a	F'2c NM	F'2c BF	Z'2a-Z'2c NM	Z'2a-Z'2c BF
ɒ	670	1050	2900	3490	947	1050	1068	-0.62	-0.72
ɑ	660	1170	2770	3650	1103	1170	1180	-0.37	-0.42
u	290	700	2550	3280	669	700	701	-0.23	-0.24
ɔ	570	840	2640	3310	806	840	842	-0.23	-0.25
o	370	730	2670	3240	700	730	733	-0.22	-0.24
ɯ	360	1550	2430	3030	1503	1859	1910	-1.41	-1.60
ɯ	300	1320	2480	3440	1300	1320	1351	-0.10	-0.25
ʌ	620	1260	2390	3610	1284	1260	1266	0.12	0.09
OE	700	1430	2390	3350	1458	1498	1464	-0.18	-0.03
ɣ	450	1300	2640	3470	1326	1300	1343	0.13	-0.08
ø	460	1520	2290	3290	1570	1786	1570	-0.86	0.00
a	770	1400	2460	3710	1452	1400	1409	0.24	0,20
ɛ	380	1690	2460	3570	1754	1966	1747	-0.76	0.03
e	470	2180	2720	3790	2361	2390	2326	-0.08	0.10
ɛ	680	1890	2580	3940	2076	2145	1925	-0.22	0.51
y	300	1890	2250	3000	2101	2031	2084	0.23	0.05
oe	640	1450	2330	3030	1637	1676	1589	-0.16	0.20
i	300	2300	3070	3590	3095	3290	3154	-0.39	-0.12

table 2
Test de la performance du modèle sur l'ensemble de voyelles de syntèse de Bladon et al. [1]
Notation comme pour la table 1

Ce fait demande confirmation expérimentale. En tout état de cause, notre formule comme celle de Bladon et Fant ne peut prédire que des valeurs de F'2 supérieures à celles de F2.

Pour les autres voyelles, la nouvelle formule est relativement satisfaisante. En général, les performances des deux modèles sont comparables.

Notre formule présente deux avantages par rapport à celle de Bladon et Fant, qui correspondent précisément aux problèmes soulevés dans l'introduction. D'une part, elle repose exclusivement sur des hypothèses classiques sur le traitement spectral dans le système auditif : échelles naturelles des fréquences et des amplitudes, correction isosonique, suppression à deux tons, intégration cnetrale à 3.5 Barks. Le regroupement artificiel de F3 et F4 est ainsi évité, et on dispose d'un modèle dont tous les éléments peuvent être justifiés. D'autre part, l'utilisation de la distance critique introduite par Chistovich résout, comme Bladon l'avait annoncé ([5]), le cas des paires de voyelles qu'il a construites pour mettre en défaut la formule de Bladon et Fant.

L'inconvénient de notre formule est le problème fondamental de tous les modèles reposants sur les données formantiques. En particulier, nous devons remarquer la grande sensibilité à la mesure de F2 et de F4. Néanmoins, l'intérêt d'une telle formule par rapport au modèle global que nous avons proposé [8] est qu'elle peut être employée sur des corpus acoustiques constitués uniquement de mesures de formants, ce qui en fait un outil très "portable".

PREMIERE APPLICATION DU MODELE A L'ETUDE DE LA LABIALITE DES VOYELLES AVANT DU FRANCAIS

La question générale que nous nous posons est la suivante: qu'est-ce que F'2 peut apporter à la classification et la normalisation de voyelles du Français? Nous abordons ici exclusivement le problème de la classification, en employant le modèle proposé.

Nous avons choisi un corpus élaboré par les chercheurs de l'Institut de Phonétique de Grenoble. Il comprend les voyelles [i], [y], [e], [ø] précédées par les consonnes [s], [z], [ʃ], [ʒ]. Chacun des 16 contextes du type CV est représenté par 3 logatomes prononcés une fois par 5 locuteurs: 3 hommes et 2 femmes. Ce corpus est donc constitué de voyelles avant arrondies et non arrondies, plongées dans des contextes maximisant et minimisant cette opposition d'arrondissement, et il a fait l'objet d'une étude très poussée au niveau articulatoire [15]. Il est très bien adapté à notre problème, puisque c'est dans cette zone de l'espace vocalique que nous attendons l'écart maximum entre F2 et F'2, donc entre "espace acoustique" et "espace perceptif".

Pour un locuteur masculin (DA), nous avons mesuré les fréquences et les niveaux des quatre premiers formants, par la méthode du cepstre en faisant 2-3 "coupes" sur toute la durée de la voyelle mesurée (pour le logiciel employé se référer à [16],[17]).

Nous visualisons sur la figure 1 le plan (F1, F2) (axes en Bark) pour toutes les coupes des voyelles de ce locuteur. Nous remarquons la bonne séparation des voyelles [e]-[ø] et le léger recouvrement de [i]-[y].

Nous visualisons sur la figure 2 l'espace "perceptif" (F1, F'2), F'2 étant estimé par notre notre modèle. La séparation des voyelles [i]-[y] est nettement améliorée, et devient aussi nette que la séparation [e]-[ø] : on peut parler d'une "normalisation" de l'opposition de labialité. Cependant, on constate une augmentation des dispersions intraclasse, particulièrement pour la classe [ø], ce qui peut diminuer l'efficacité de la classification.

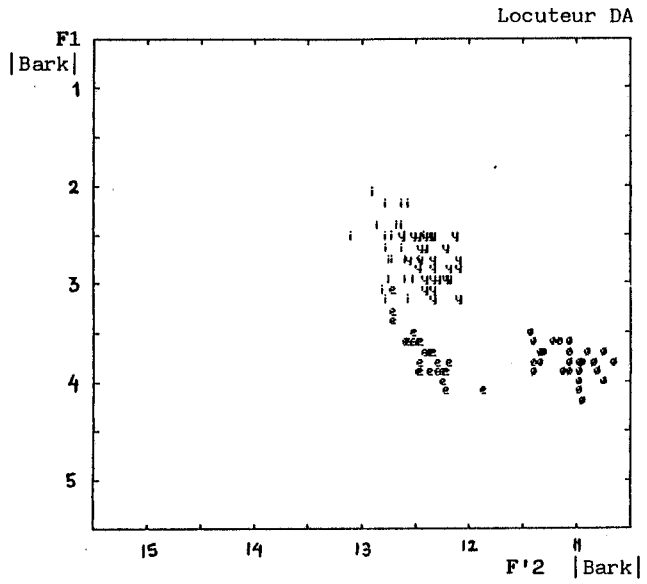


figure 1
Espace acoustique (F1, F2) pour le locuteur DA (axes en Bark)

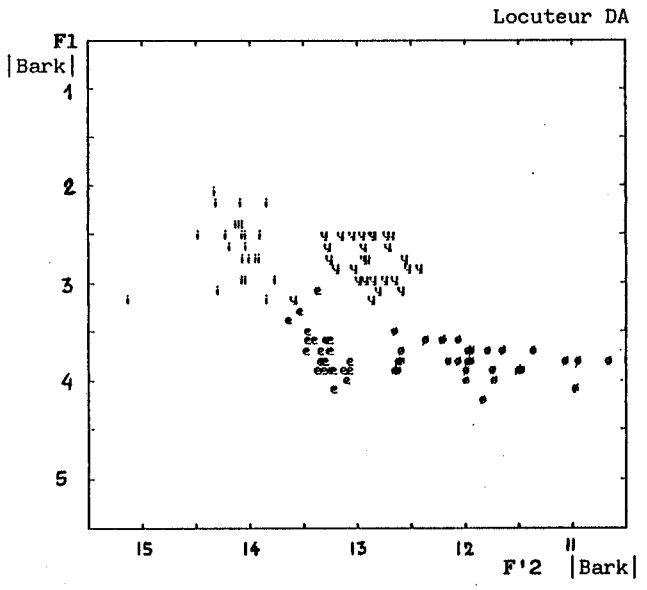


figure 2
Espace "perceptif" (F1, F'2) pour le locuteur DA (axes en Bark)

Pour ce locuteur, l'implémentation de la formule présente les problèmes suivants.

Pour la voyelle [i], comme il était prévu, F3 est plus proche de F4 que de F2. Pourtant, la grande distance de F4 par F2 conduit à rejeter le groupe F3-F4 comme corrélat principal de F'2, au profit du groupe F2-F3, ce qui conduit à des valeurs de F'2 relativement basses. D'autre part, pour [y] la petite distance de F2 à F4 a un effet opposé pour un sous-ensemble minoritaire de points caractérisés par une valeur de F3 plus proche de F4 que de F2. Pour éviter le recouvrement des valeurs de F'2 pour cette paire de voyelles, on doit proposer un compromis subtil pour la valeur exacte de la distance critique (on a choisi une zone critique de 3.0 à 3.3 Barks pour la figure 2).

En ce qui concerne le [ø], F4 est trop éloigné de F2 pour contribuer à F'2. D'autre part, la distance entre F2 et F3 est tantôt supérieure, tantôt inférieure à la distance critique pour cette voyelle, ce qui constitue la cause principale de dispersion.

Ce premier résultat de notre étude, bien qu'insuffisant pour nous permettre de conclure, nous semble pourtant positif. Pour d'autres locuteurs, un certain nombre de problèmes se posent, liés au problème classique de la mesure des formants. On aborde par là un autre intérêt majeur de la notion de centre de gravité, que nous ne voulons explorer que dans un second temps : une intégration à large bande peut précisément permettre d'éviter de passer par la détection et la mesure fine des formants. Un tel mécanisme peut donc nous fournir une série de paramètres stables pour la détection de certains traits vocaliques, tels que les traits grave-aigu ou bémolisé-non bémolisé.

Cette étude doit être continuée plus systématiquement et pour plus de locuteurs. La mise en évidence des valeurs du F'2 par une expérience de perception pourrait s'avérer particulièrement importante.

REMERCIEMENTS

Nous remercions C. Abry et L.J. Boë de nous avoir fait bénéficier de leur précieux corpus, et le second pour les innombrables heures qu'il a passées avec nous pour l'analyse de ce corpus.

REFERENCES

[1] R.A.W. Bladon, G. Fant, "A two formant model and the cardinal vowels", Speech Transmission Lab. QPSR 1/1978.

[2] R. Carlson, G. Fant, B. Granström, "Two-formant models, pitch and vowel perception", in "Auditory Analysis and Perception of Speech", G. Fant et M.A.A. Tatham eds, Academic Press, London, pp. 55-82, 1975.

[3] K.K. Paliwal, D. Lindsay, W.A. Ainsworth, "A study of two-formant models for vowel identification", Speech Communication 2, pp. 295-304, 1983.

[4] M. Mantakas, "Perception de voyelles stationnaires: Etude d'un modèle à deux paramètres (Modèle de Bladon et Fant), Rapport DEA, ENSERG, 1985.

[5] A. Bladon, "Two-formant models of vowel perception: Shortcomings and enhancements", Speech Comm. 2 (1983), pp.305-313.

[6] L.A. Chistovich, R.L. Sheikin, V.V. Lublinskaja, "'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, 1979.

[7] J.L. Schwartz, P. Escudier, "Le système auditif humain comprend-il un mécanisme d'intégration spectrale à large bande?", article présenté dans ce Congrès, 1986.

[8] P. Escudier, J.L. Schwartz, M. Boulogne, "Perception de voyelles stationnaires: Représentation interne des formants dans le système auditif et modèles à deux formants", Symposium Franco-Suédois, Grenoble, 1985.

[9] E. Zwicker, R. Feldkeller, "Psychoacoustique", Masson, 1981, traduction en Français de "Das Ohr als Nachrichtenemfärger", 1967.

[10] R. Carlson, B. Granström, G. Fant, "Some studies concerning perception of isolated vowels", Speech Transmission Lab. QPSR 2-3/1970.

[11] G. Fant, "Acoustic Theory of Speech Production", 's-Gravenhage 1960.

[12] G. Fant, "The vocal tract in your pocket calculator", Speech Transmission Lab. QPSR 2-3/1985.

[13] M.B. Sachs, E.D. Young, "Encoding of steady-state vowels in the auditory nerve : representation in terms of discharge rate", J. Acoust. Soc. Am. 66, pp. 470-479, 1979.

[14] P.G. Stelmachowicz, A.M. Small, P.J. Abbas, "Suppression effects for complex stimuli", J. Acoust. Soc. Am. 71, pp. 410-420, 1982.

[15] Institut de Phonétique de Grenoble, "Labialité et Phonétique", Publ. de l'Univ. des Langues et Lettres de Grenoble, 1980.

[16] C. Benoit, "EDISIG : encore un éditeur de signal ?!!", 13ème JEP , pp. 211-214, 1984.

[17] G. Feng, "Analyse cepstrale, visualisation sonagraphique et détection des formants", Séminaire GALF : Traitement du signal de parole, pp. 207-216, 1983.

ESSAI DE FORMALISATION DE FAITS ET HYPOTHESES DE PHYSIOLOGIE CONCERNANT LE TRAITEMENT DE L'INFORMATION POUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

D. Béroule - J.L. Schwartz

LIMSI Orsay - ICP Grenoble

Abstract

We present herein some physiological properties of the auditory system liable to provide a new insight in the Signal and Information Processing for Automatic Speech Recognition. We mainly discuss three points: signal analysis in the peripheral auditory system, spectral processing in the higher levels, connexionist models for cognitive processing.

INTRODUCTION

Comme les problèmes en reconnaissance de parole et notamment au niveau du décodage acoustico-phonétique restent considérables, mais aussi parce que cette recherche se doit par nature d'être pluridisciplinaire, de nombreux travaux dans le domaine se sont inspirés de faits et d'hypothèses de physiologie et de neurophysiologie. Or, si de nombreux progrès peuvent être attendus de ces domaines, il apparait (ce fut parfois le cas à nos propres dépens) que les pièges y sont nombreux, et ce à cause de la considérable étendue de notre ignorance, aussi bien au niveau qualitatif (quelles unités quelles propriétés, dans quelles structures, pour quelles fonctions ?) qu'au niveau quantitatif (comment "accorder" les paramètres, en nombre souvent considérable). Mais comme néanmoins les enjeux nous semblent essentiels, et le terrain susceptible de nombreuses récoltes, nous avons voulu tenter de décrire quelques grandes lignes de son état actuel, en nous concentrant exclusivement sur les caractéristiques de traitement de l'information actuellement dégagées, et sur les projets que l'on peut proposer pour espérer mettre en oeuvre de nouveaux formalismes.

SYSTEME AUDITIF PERIPHERIQUE (S.A.P.)

Propriétés principales

1. Renforcement des fréquences moyennes (1000-4000 Hz) assimilable aux préaccentuations classiques.

2. Analyse spectrale avec ligne à retard (codage géographique).

3. Phénomènes actifs non linéaires.

4. Suppression à deux tons.

5. Adaptation nerveuse.

6. Eléments à seuil et saturation - réponse au logarithmique de l'excitation.

7. Existence de plusieurs types de fibres, différenciées notamment par leur point de fonctionnement (seuils).

8. Synchronisation des décharges nerveuses sur les temps forts de l'excitation (codage temporel).

Les deux premières caractéristiques sont d'origine essentiellement mécanique, les cinq dernières d'origine essentiellement nerveuse, l'existence de phénomènes actifs étant l'élément nouveau apparu ces dernières années, et d'origine non encore totalement éclaircie.

Conséquences sur les modèles du S.A.P.

Quelle que soit leur philosophie (modèles "physiques" ou modèles "fonctionnels"), quelles que soient les hypothèses et les équations sur lesquelles ils s'appuient, tous les modèles qui se revendiquent explicitement du S.A.P. peuvent être entièrement décrits, au niveau des traitements qu'ils impliquent, par un sous-ensemble de ces propriétés. Ainsi, le modèle de Dolmazon-Bastet [1] contient les propriétés **1-2-4-6-8** le modèle de Delgutte [2] les propriétés **1-2-5-6-7** celui de Seneff [3] les propriétés **2-5-6-8** celui de Lyon [4] les propriétés **2-5-6-7-8**. Le modèle de Caelen [5], qui pourrait etre caractérisé ici par les propriétés **1-2-4-6-8** contient d'autres potentialités de traitement (adaptations multiples qui correspondraient à des voies efférentes) qui en font plus, actuellement, un système performant inspiré du S.A.P. qu'un modèle réaliste, jusqu'à plus ample information expérimentale.

Conséquences en psychoacoustique

L'essentiel des phénomènes de psychoacoustique "élémentaires" peuvent se comprendre entièrement à l'aide des propriétés du S.A.P., sans intervention de quelque autre mécanisme. Ainsi, nous citons certains de ces phénomènes en donnant entre parenthèses les éléments nécessaires à leur explication, définis par leur numéro donné ci-dessus.

- Seuils d'audibilité définis par les audiogrammes (**1-6**).
- Seuils différentiels en intensité, en fréquence (**1-2-6**).
- Bandes critiques (**2**)
- TTS, adaptation induite homolatérale (**3-5**)
- Masquage : courbes d'accord (2), suppression ou "unmasking" (**4**), masquage avant (**5**), "masking period pattern" (**5-6-8**).

Implications pour un modèle perceptif

Le fait qu'un nombre important de faits de perception puissent être attribués pour l'essentiel au fonctionnement du S.A.P., fait qui a eu des retombées décisives sur les rapprochements entre psychoacousticiens et physiologistes, mérite que l'on s'y attarde quelques instants. Le S.A.P. ne constitue que le point d'entrée d'un système complet dont le "coeur", le système nerveux central (S.N.C.) semble d'une puissance et d'une complexité qui suscite le respect et la frayeur des chercheurs qui s'y attaquent. Ainsi, il est clair (nous y reviendrons au chapitre suivant) qu'il s'y effectue un ensemble de traitements spectraux qui transforme considérablement la première image fournie par le S.A.P. Il faut donc se persuader qu'il existe également une voie directe du S.A.P. au système de décision, voie utilisée lors de ces tâches "élémentaires" dont nous avons parlé ci-dessus. Ce fait peut conduire à modifier l'idée classique qu'on se fait du S.A.P., celle d'une structure de traitement très hiérarchique, au profit d'un modèle à représentations multiples, non entièrement hiérarchique, tel que celui que l'on commence à dégager en perception visuelle [6] : le traitement se ferait au travers des multiples "cartes de représentation", toutes disponibles à tout instant, dont disposerait, à divers niveaux, le système de décision.

Conséquences en traitement de la parole

Les propriétés **1** et **2** sont très classiques, et utilisées dans quasiment tous les systèmes d'analyse du signal de parole. Le point **3**, au contraire, est le seul qui n'a encore suscité aucune application pour le codage de la parole dans le S.A.P., ni dans les modèles du S.A.P., comme le montre la liste ci-dessus, ni dans l'interprétation des résultats de neurophysiologie.

L'existence de phénomènes actifs suggère pourtant une hypothèse intéressante, à savoir que les transformations opérées sur le signal au niveau périphérique tiendraient compte d'informations délivrées dynamiquement par les niveaux supérieurs de traitement. Ce schéma fonctionnel qui ne correspond pas à notre conception habituelle d'un étage de traitement "à sens unique", pourrait avoir des répercussions sur le mode de sélection des informations extraites du signal, voire apporter des éléments de réponse au problème non résolu de la reconnaissance en milieu bruité.

Les propriétés **4**, **6**, **7** et **8** sont au coeur de tous les travaux réalisés sur le codage du spectre dans le S.A.P., et principalement sur le codage des voyelles et des fricatives sourdes ([2],[7]). Le point **5**, qui est à la base des principaux résultats de Delgutte sur le codage des particularités dynamiques ([2]), trouve un analogue fonctionnel dans certaines méthodes de compression temporelle fondées sur une mesure de variabilité spectrale [8]. Cependant, cette mesure concerne l'ensemble du spectre instantané, alors que la réponse d'une fibre nerveuse peut privilégier les variations d'énergie se produisant à l'intérieur d'une plage de fréquence particulière.

En définitive, tous ces résultats peuvent donc se comprendre dans les grandes lignes sur la base des caractéristiques des traitement détaillées ci-dessus, et donc se reproduire sur l'un ou l'autre des modèles existants.

TRAITEMENT DU SIGNAL DANS LES NIVEAUX SUPERIEURS

La plupart des éléments fonctionnels subsistent dans les niveaux supérieurs ; en fait, tous les éléments d'origine nerveuse, c'est-à-dire : seuils, saturations - suppression fréquentielle, réalisée à ces niveaux par des procédés totalement différents, mais sous une forme de traitement très semblable, l'inhibition latérale - adaptation nerveuse, concentrant le traitement sur tout ce qui "bouge", au détriment de ce qui n'évolue pas - phénomènes actifs généralisés, c'est-à-dire efférences multiples. Les propriétés de filtrage (principalement mécanique, dans le S.A.P.) sont réalisées et généralisées par les fonctions synaptiques classiques (excitations - inhibitions) où les facteurs de synchronisation sont très importants.

L'explosion combinatoire liée au cablage fait qu'on doit chercher non plus des propriétés unitaires mais des propriétés de réseaux, ou des fonctions nerveuses (fonctions au sens de traitement de l'information). Ainsi, on peut imaginer des modèles neuronaux simples simulant, en psychoacoustique, les systèmes proposés pour le traitement du pitch, de la sonie, de la rugosité, du timbre (voir par exemple [9], [10], [11]) ; ou, en traitement de la parole, des systèmes de mesure de VOT, de reconnaissance des plosives [2], de reconnaissance phonémique ([12]), ... Nous allons nous attacher au problème le plus étudié, celui du traitement spectral.

Traitement spectral

La plupart des travaux en modélisation du système auditif sont actuellement concentrés sur ce problème. L'objectif est de construire une fonction (de la forme spatio-temporelle de sortie du S.A.P.) qui mette en évidence les éléments spectraux primordiaux du spectre, les "particularités" au sens de Chistovich. Le cahier des charges concernant ce que doivent être ces éléments est rarement précisé, mais son contenu implicite est clair : on désire récupérer certains paramètres tels que F0 et les formants (ou parfois les intégrer dans une représentation plus globale), se défaire de la structure harmonique, des pics "parasites", diminuer le poids de l'énergie totale et des rapports d'amplitude entre maxima spectraux, le tout en espérant respecter plusieurs contraintes qualitatives (lissage temporel et fréquentiel, normalisation par exemple).

Tous les modèles proposés, à l'exception du dernier en date ([13], dont nous ne parlerons pas ici), peuvent être compris à partir de la Figure 1, proposée par Delgutte [2]. Cette figure montre schématiquement comment se répartit l'énergie dans la distribution spatio-temporelle de l'excitation des fibres du nerf auditif, stimulées par un spectre stationnaire de type voyelle, de fondamental F0, avec deux formants F1 et F2. L'un des axes de cette représentation est l'axe des fréquences caractéristiques des fibres, soit l'axe porteur de l'information géographique. L'autre fournit les valeurs des fréquences d'analyse par FFT de l'excitation d'une fibre de fréquence caractéristique donnée, c'est donc l'axe porteur de l'information temporelle. On peut alors définir trois types de modèles.

A. Modèles "géographiques"

Ils partent tous d'une information exclusivement géographique, c'est-à-dire d'une intégration de l'énergie selon l'axe des fréquences de FFT. La fonction de transformation initiale transforme donc $R(CF,F)$, répartition spatio-temporelle de l'énergie (CF=fréquence caractéristique d'une fibre, F=fréquence d'analyse FFT de la réponse de la fibre) en un spectre $R'(f)$ selon la formule:

$$R(CF,F) \quad \rightarrow \quad R'(f) = \int_0^{+\infty} R(f,F)dF$$

De là dérivent les types suivants :

1. Modèle sans hypothèse supplémentaire ([11], [12] avec deux régimes de fibres à points de fonctionnement différents, ce qui permet de simuler une dynamique élargie).

2. Modèles à inhibition latérale, pour renforcer l'énergie des maxima (voir [14], [15], [5]). Notons que la distance proposée par Klatt en reconnaissance de parole ([16]) constitue implicitement un modèle de ce type.

3. Modèles à intégration large bande, visant à simuler le phénomène de centre de gravité (voir [17] par exemple).

B. Modèles temporels

Ils partent tous d'une information exclusivement temporelle, c'est-à-dire d'une intégration le long de l'axe horizontal, soit:

$$R(CF,F) \quad \rightarrow \quad R'(f) = \int_0^{+\infty} R(CF,f)dCF$$

Ces modèles ne respectent donc pas le principe de localisation (voir [4], [18]).

C. Modèles mixtes

Compromis des deux types précédents, ces modèles utilisent à la fois les deux types d'information, en intégrant l'énergie dans des pavés, selon la formule :

$$R(CF,F) \quad \rightarrow \quad R'(f) = \int_{f-a}^{f+a} \int_{f-b}^{f+b} R(CF,F)dCFdF$$

Plusieurs exemples ont été proposés, voir [2], [3], [7], [19].

Tous ces modèles pourraient fournir un ou plusieurs niveaux de représentation, plus élaborés que le niveau initial. Néanmoins, comme nous l'avons noté au chapitre précédent, ces représentations doivent cohabiter avec celle fournie par le S.A.P., et c'est sur l'ensemble de ces images du signal que peut travailler le système de décodage.

La manière dont nous avons organisé ces travaux nous conduit à penser que l'essentiel de la substance a été tiré de la représentation portée sur la figure 1, et qu'il n'y a donc plus actuellement à attendre de progrès spectaculaires à ce niveau, du moins tant que le rôle des phénomènes actifs n'aura pas été éclairci.

Ainsi, les connaissances sur le S.A.P. d'une part, sur les premiers traitements spectraux susceptibles d'être appliqués au-delà du S.A.P. d'autre part, nous semblent maintenant relativement bien formalisées. Le poids peut donc être mis sur les traitements plus complexes, aussi bien au niveau du décodage acoustico-phonétique que des traitements cognitifs. C'est de tels traitements au niveau cognitif que nous allons aborder dans le dernier point.

TRAITEMENT DE L'INFORMATION AU NIVEAU CENTRAL

Au niveau central, hormis quelques résultats ponctuels sur la présence de détecteurs spécialisés [20] et l'existence de *lames iso-fréquentielles* dans le système auditif [21], on ignore l'organisation du réseau nerveux qui sous-tend les différentes composantes du traitement perceptif. Les seules données ayant fait l'objet de modélisations sont des résultats généraux de psychophysiologie qui ne concernent pas spécifiquement le système auditif, mais qui présentent néanmoins l'intérêt de suggérer des modes d'acquisition, de représentation et d'utilisation de l'information fort différents de ceux mis en oeuvre dans les systèmes actuels de Reconnaissance Automatique de la Parole (RAP).

C'est cette conception "non informaticienne" du traitement de l'information que nous allons chercher à préciser, à travers la présentation d'un échantillon de modèles connexionnistes appliqués à la Reconnaissance des Formes.

Modèles connexionnistes

Les modèles connexionnistes se définissent par la mise en oeuvre un réseau homogène de neurones formels, définition suffisamment générale pour admettre de nombreuses variantes, que ce soit au niveau du traitement effectué localement par chaque neurone formel, au niveau de la structure du réseau, ou encore au niveau des lois qui définissent son comportement [22]. Malgré leur diversité, il est possible de distinguer deux générations de modèles: les plus anciens, du type *Perceptron* [23] (référencé *(1)* dans la suite du texte), et la nouvelle génération, née de l'idée d'appliquer un formalisme bien rôdé (celui de la mécanique statistique), à l'étude du comportement de tels réseaux [24] (référencé *(2)*). Dans l'exposé des propriétés de traitement de l'information de ces modèles, nous faisons également référence *(3)* au système que nous développons actuellement [25], ses caractéristiques différant assez nettement de celles de ses prédécesseurs.

• Représentation des connaissances

Les modèles connexionnistes proposent un mode de représentation essentiellement *topographique* et *dynamique*, puisque c'est l'entrée en activité d'un sous-ensemble plus ou moins restreint du réseau qui répond à la présence d'une forme donnée, sous-ensemble incluant notamment les capteurs situés à l'entrée du réseau. Dans cette *mémoire associative*, la population de neurones formels

représentant un fait donné admet une partie commune avec la représentation interne d'autres faits.

Dans le modèle *(1)*, chaque neurone formel réagit sélectivement à un fait particulier, ce qui n'est pas le cas avec *(2)*, dont la représentation est plus distribuée. Chaque neurone formel du modèle *(3)* réagit à l'association d'un fait et d'un contexte particulier d'apparition de ce fait.

• **Accès aux représentations**

L'accès à l'information figurant dans le réseau s'effectue soit par activation directe des zones du réseau concernées par une forme donnée *(1)*, soit par stabilisation de l'état du réseau dans une configuration globale d'activité satisfaisant un certain critère *(2)*, soit par orientation de la propagation d'activité qui siège dans le réseau *(3)*.

l'accès est en théorie immédiat pour *(1)* et *(3)*, contrairement à *(2)* qui nécessite de nombreuses itérations du même processus de convergence vers un état stable.

• **Acquisition des représentations**

Dans les systèmes *(1)* et *(2)*, l'acquisition d'information demeure circonscrite à une phase préalable d'apprentissage qui concerne uniquement la pondération des liens entre neurones formels d'un réseau figé. L'ajustement de ces pondérations permet d'operer une classification discriminante *(1)*, ou de rendre compte statistiquement des interactions entre éléments du réseau *(2)*.

Dans le modèle *(3)*, le réseau est élaboré en cours de traitement, ce qui ne correspond pas à la distinction habituelle entre phase d'acquisition et phase d'utilisation des connaissances. Les seuils, délais de propagation le long des liaisons et la durée des signaux générés par les éléments du réseau sont également ajustés dynamiquement; quant à la valeur des pondérations, elle constitue une mesure de la fréquence des faits représentés.

Conséquences en traitement de la parole

La quasi-totalité des modèles connexionnistes traitent des formes spatialement étendues et stationnaires (forme visuelle, image spectrale d'une forme acoustique), ce qui les rend apparemment peu adapté à la RAP. Certains de ces réseaux sont néanmoins utilisés pour la segmentation du signal [26], ou bien en reconnaissance phonétique [27]. A noter également l'utilisation inatendue d'un modèle de type *(2)* dans une tâche de désambiguation lexicale impliquant des traitement syntaxiques et sémantiques [28], ainsi que l'existence d'un modèle de perception de la parole compatible avec des résultats de psycholinguistique [29]. Le système *(3)* est adapté au traitement de séquences d'événements instantanés spatialement localisés, et prend notamment en compte le problème des distorsions temporelles.

Parmi les propriétés de ces modèles potentiellement intéressantes pour la RAP, citons:
- Une certaine robustesse au bruit ou à l'absence d'une partie des composants d'une forme.
- Un fonctionnement intrasèquement parallèle qui pourrait être mis à profit pour mobiliser instantanément les multiples représentations associées aux informations plus ou moins abstraites du message vocal.
- Un traitement distribué qui pourrait permettre d'assigner à chaque fait représenté un jeu de paramètres tenant compte de ses particularités.

CONCLUSION

L'application de données de physiologie à la RAP s'est jusqu'à présent limité à la conception de modèles d'oreille utilisés comme organe de pré-traitement du signal dans un système de reconnaissance algorithmique [30], ou inversement, à la juxtaposition d'un réseau de neurones formels et d'un banc de filtres conventionnel [26]. Les performances de tels systèmes hybrides ne montrent pas de progrès notables par rapport aux systèmes classiques, ce qui semble mettre en cause l'intérêt de telles recherches pour la RAP. Selon nous, cela montre plutôt que si l'on choisit cette voie de recherche, il est nécessaire de s'y engager complètement, en faisant porter son effort conjointement sur les niveaux inférieurs et supérieurs de traitement, selon la démarche suivante:
- Organiser les données de psychologie, en y associant nos propres résultats expérimentaux (psychoacoustique), pour valider des modèles existants ou en donner les limites, afin d'élaborer de nouveaux modèles.
- Concevoir des outils dont le mode de traitement de l'information soit compatible avec la physiologie, moins pour en faire dans l'immédiat un système de RAP que pour pouvoir exploiter plus directement des donnés de physiologie.

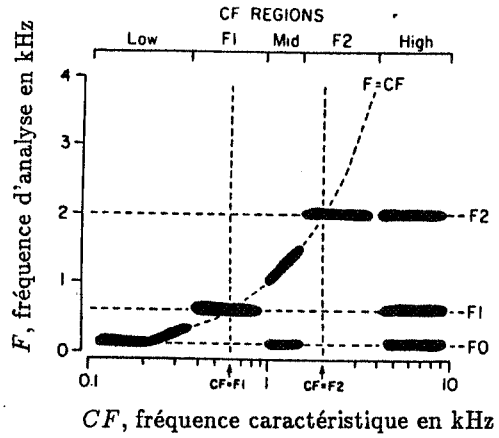


Figure 1 (d'après Delgutte [2]):

Répartition spatio-temporelle de l'énergie dans les fibres, $R(CF,F)$ (voir texte)

REFERENCES

[1] J.M. Dòlmazon, "Contribution aux recherches sur l'appareil auditif. Elaboration et exploitation d'un modèle de fonctionnement du système auditif périphérique", Thèse d'Etat, INP Grenoble, 1980.

[2] B. Delgutte, "Codage de la parole dans le nerf auditif", Thèse d'Etat, Université Paris 6, 1984.

[3] S. Seneff, "Pitch ans spectral analysis of speech based on an auditory synchrony model", M.I.T., Technical Report 504, 1985.

[4] R.F. Lyon, "Computational models of neural auditory processing", Proc. IEEE, 36.1.1, 1984.

[5] J. Caelen, "Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique", Thèse d'Etat, Université de Toulouse, 1979.

[6] J. Bullier, "Les cartes du cerveau", Séminaire de sciences cognitives, Université des Sciences Sociales de Grenoble, 1986.

[7] M.B. Sachs, E.D. Young, M.I. Miller, "Encoding of speech features in the auditory nerve", dans : The representation of speech in the peripheral auditory system, édité par R. Carlson et B. Granstrom, pp. 115-130, Elsevier Biomedical, Amsterdam, 1982.

[8] Gauvain, J.L. "Reconnaissance de mots enchaînés et détection de mots dans la parole continue", Thèse 3è Cycle, Orsay, Juin 1982.

[9] E. Terhardt, "On the perception of spectral information in speech", dans : The representation of speech in the peripheral auditory system, édité par R. Carlson et B. Granstrom, pp. 281-291, Elsevier Biomedical, Amsterdam, 1982.

[10] E. de Boer, "Pitch theories unified", dans : Psychophysics and Physiology of hearing, édité par E.F. Evans et J.P. Wilson, Academic Press, 1977.

[11] E. Zwicker, R. Feldtkeller, "Das Ohr als Nachrichtenempfanger", Hirzel éditeur, Stuttgart, 1967.

[12] E. Zwicker, E. Terhardt, E. Paulus, "Automatic speech recognition using psychoacoustic models", J. Acoust. Soc. Am. 65, pp. 487-498, 1979.

[13] S.A. Shamma, "Speech processing in the auditory system II : lateral inhibition and the central processing of speech evoked activity in the auditory nerve", J. Acoust. Soc. Am. 78, pp. 1622-1632, 1985.

[14] E.G. Karnickaya, V.N. Mushnikov, N.A.Slepokurova, S.J.Zhukov, "Auditory processing of steady-state vowels", Symposium : Auditory analysis and perception of speech, Leningrad, 1973.

[15] V.C. Dang, R. Carré, D. Tuffelli, "Research on the preprocessing by lateral inhibition", ICA 1986.

[16] D. Klatt, "Prediction of perceived phonetic distance from critical- band spectra : a first step", Proc. IEEE, pp. 1278-1281, 1982.

[17] P. Escudier, J.L. Schwartz, M. Boulogne, "Perception of spationary vowels : internal representation of the formants in the auditory system and two-formantmodels", Symposium Franco-Suédois Grenoble 1985.

[18] A. de Cheveigné, "A pitch perception model", IEEE 1986.

[19] P. Srulovicz, J.L Goldstein, "A central spectrum model : a synthesis of auditory-nerve timing and place cues im monaural communication of frequency spectrum", J. Acoust. Soc. Am. 73, pp. 1266-1276, 1983.

[20] W.D. Keidel "Information processing in the higher parts of the auditory pathway", Facts and Models in Hearing, Springer-Verlag, 1974

[21] W.R. Webster, J. Servieres, M. Brown "Inhibitory contours in the inferior colliculus as revealed by the 2-dioxyglucose method, Exp. Brain Res., 56, pp 577-581, 1985

[22] J.L. Schwartz, D. Béroule "Réalisation, à partir de données de physiologie et de psychoacoustique, d'un modèle de perception d'entités linguistiques", rapport d'ATP CNRS, 1985

[23] F. Rosenblatt "The Perceptron: A probabilistic model for information storage and organization in the Brain", Psychological Review, Vol.65, pp 386-407, 1958

[24] G.E. Hinton, T.J. Sejnowsky, D.H. Ackley "Boltzmann Machines: Constraints Satisfaction Networks that Learn", Rapport CMU-cs-119, 1984.

[25] D. Béroule "Un modèle de mémoire adaptative, dynamique et associative pour le traitement automatique de la parole", Thèse 3ème cycle, Orsay, 1985

[26] T. Kohonen, Self-Organization and Associative Memory Springer-Verlag, 1984.

[27] J.S. Bridle, R.K. Moore "Boltzmann Machines for Speech Pattern Processing, Proc. Inst. Acoust. Autumn Meeting, Novembre 1984

[28] D. Waltz, J. Pollack "Massively Parallel Parsing: A strongly interactive model of natural language interpretation" Cognitive Science No 9, 1985

[29] J.L. Mc Clelland, J.L. Elman "The TRACE Model of Speech perception, Cognitive Psychology", Vol. 18, No 1, 1986

[30] Y. Dologlou "Evaluation des performances d'un modèle du système auditif en reconnaissance de la parole (comparaison avec la prédiction linéaire)", Thèse DI, Grenoble, 1984

SEGMENTATION

Président

Jacqueline VAISSIERE

C.N.E.T. de Lannion

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

C. Benoît

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

ABSTRACT

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

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

INTRODUCTION

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

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

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

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

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

---

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

FORMULES GENERALES DE CORRELATION

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

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

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

et

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

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

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

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

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

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

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

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

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

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

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

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

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

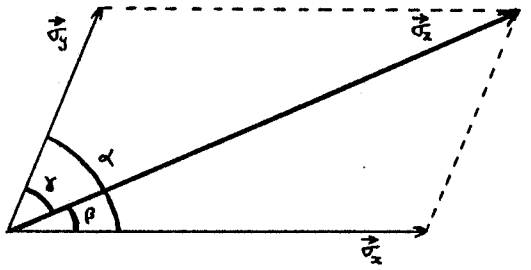


Fig. 1

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

PROPRIETES DES CORRELATIONS ENTRE VARIABLES "CHEVAUCHANTES"

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

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

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

On démontre :

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

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

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

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

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

---

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

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

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

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

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

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

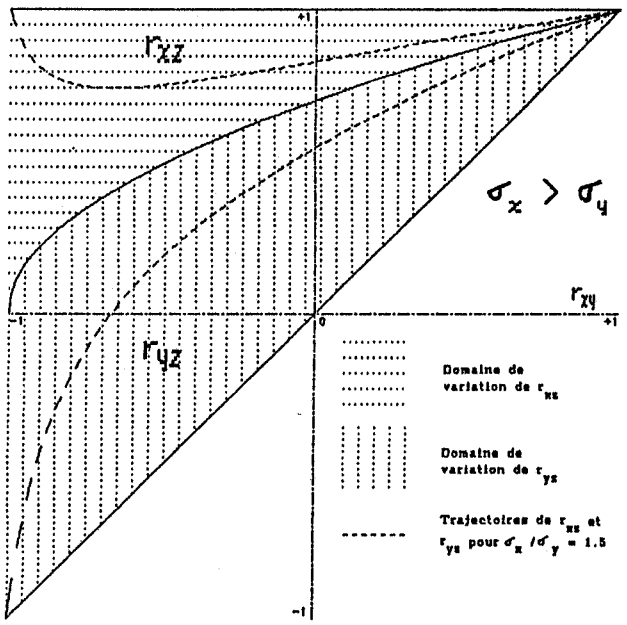


Fig. 2
Domaines de variation des corrélations de chacune des variables x et y avec leur somme z=x+y, en fonction du coefficient r_{xy}, obtenus en faisant varier le rapport $k = \sigma_x / \sigma_y$ de 1 (courbe de séparation en traits pleins) à l'infini (droite sup. horizontale pour r_{xz} et inf. oblique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

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

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

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

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

Deux relations remarquables en découlent :

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

et :

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

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

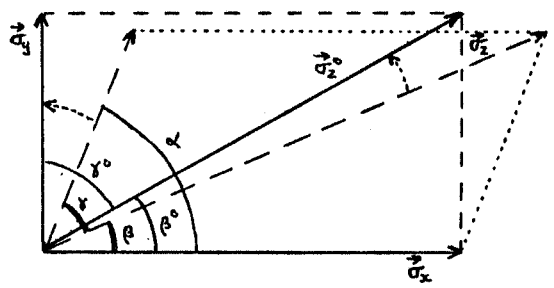


Fig. 3
Représentation géométrique analogue à la Fig. 1 pour les traits tiretés (corrélations observées : α, β et γ) et construction en traits pleins obtenue à partir des corrélations "espérées" : $\alpha^{o}=90^{o}$, β^{o} et γ^{o} telles que

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

CONCLUSION

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

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

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

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

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

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

REMERCIEMENTS

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

LOCALISATION ET REPRESENTATION TEMPORELLE D'EVENEMENTS PHONETIQUES

applications en étiquetage, en segmentation et en synthèse

F.BIMBOT* , S.M.MARCUS** , G.CHOLLET*,**

* C.N.R.S., UA 820, E.N.S.T. dpt SYC, 46 rue Barrault, 75634 PARIS cedex 13, France.

** Institute for Perception Research, IPO, P.O. box 513, 5600 MB EINDHOVEN, Pays-Bas.

ABSTRACT

Speech is a complex code, for which information units are highly context-dependent. Spectrally stable regions can however often be localized. Non-stationnary segments (polysons) involving interaction between two or more units are found between these regions.

Atal proposed a technique for decomposing speech into temporally overlapping events. Originally developped for speech coding, its relationship with phonetic labelling was only suggested. Marcus modified Atal's technique in order to obtain robust functions which may be more plausibly related to phones.

This paper relates Atal's technique to a class of time dependent models. It demonstrates how the combination of this technique with a spectral stability criteria can be used as a tool for phonetic decoding of speech segments. Emphasis is placed on applications to speech synthesis and a framework for developping acoustic rules is proposed.

I-INTRODUCTION.

Le signal de parole est le support d'une information linguistique répartie dans le temps. Au niveau segmental, chaque unité signifiante (phone) est caractérisée par un certain nombre d'indices acoustiques. Toutefois, ces indices sont sensiblement influencés par le contexte. Inversement, la zone d'influence d'un phone recouvre fortement celles de ses voisins. En ce sens, la parole est un code complexe, les éléments informatifs s'avèrant étroitement imbriqués.

Cependant, l'examen de spectres à court terme ou de l'évolution de modèles paramétriques met en évidence des zones stables ou quasi-stables, par opposition à des zones transitoires, douces ou brutales.

Cet article présente une technique de localisation, d'étiquetage et de description paramétrique d'événements phonétiques.

Une première partie décrit un algorithme de localisation de zones stables, en s'appuyant sur un modèle LPC. Ceci permet de définir une nouvelle unité segmentale: le polyson. La seconde partie est consacrée à la description d'une technique de décomposition temporelle de la parole, technique proposée par Atal, et améliorée par Marcus et qui vise à obtenir une représentation de la parole, sous forme d'unités discrètes consécutives dont les zones d'influence se recouvrent partiellement. La dernière partie tentera enfin de donner un aperçu des perspectives offertes par ces techniques, plus particulièrement en synthèse.

II-STABILITE SPECTRALE.

Un modèle simplifié de production de la parole suppose l'existence de cibles articulatoires associées à chaque phone, et réparties de manière non uniforme au cours du temps. A ces cibles correspondent en principe une zone de stabilité spectrale au cours de laquelle la variation des paramètres acoustiques connaît un ralentissement plus ou moins net. Cependant, à débit normal, l'inertie des articulateurs empêche parfois ces cibles d'être atteintes; la nature du phone, ainsi que son contexte sont déterminants. Pour ces phonèmes, la détection de longs segments stables devient hasardeuse, voire impossible.

Des techniques ont été mise au point afin de mettre en évidence les zones spectralement stables, en s'appuyant sur l'étude de l'évolution de modèles paramétriques [1]. la méthode que nous allons détailler opère à partir des log area ratios.

1) les log area ratios:

Dans le but d'aboutir à une quantification optimale des résultats de prédiction linéaire, Viswanathan & Makhoul [2] ont défini un jeu de paramètres (log area ratios), obtenus à partir d'une linéarisation de la courbe de sensibilité spectrale des k_i (coefficients de réflexion).

$$g_i = \ln \frac{1 + k_i}{1 - k_i}$$

L' espace E des log area ratios présente de nombreuses propriétés:

- il est **stable** par combinaison linéaire (toute combinaison linéaire de g_i donne naissance à un g_i),
- il est **homogène** (une variation "dg" sur un g_i a la même influence spectrale, quelle que soit la valeur de g_i),
- il est **isotrope** (une variation "dg" sur un g_i a la même influence spectrale, quelle que soit la valeur de i),
- il est **adapté à la parole**, (la répartition statistique des g_i réels est à peu près uniforme, du moins dans l'intervalle [-20,+20]).

2) mesure de stabilité spectrale:

Les propriétés des log area ratios permettent de munir E de la distance euclidienne.

A une fenêtre de signal indiciée n, on peut associer un jeu de log area ratios déduit d'un modèle LPC sous-jacent. A ces paramètres correspond un point représentatif dans E.

Pour 2p+1 fenêtres consécutives autour de la fenêtre n (court terme), on obtient un nuage de points $\{G_{n-p}, \ldots, G_n, \ldots, G_{n+p}\}$, dont on peut déterminer le centre de gravité $\overline{G_n}$. L'inertie moyenne de ce nuage par rapport à $\overline{G_n}$ (chaque point étant affecté du poids 1), est alors donnée par la formule:

$$I_p = \frac{1}{2p+1} \sum_{k=-p}^{k=+p} d\,(G_{n+k}, \overline{G_n})^2$$

Pour 2q+1 fenêtres consécutives, toujours centrées sur n, et avec q > p (long terme), on peut calculer l'inertie moyenne du nuage de points agrandi $\{G_{n-q}, \ldots, G_n, \ldots, G_{n+q}\}$, par rapport au centre de gravité du court terme:

$$I_q = \frac{1}{2q+1} \sum_{k=-q}^{k=+q} d\,(G_{n+k}, \overline{G_n})^2$$

Si l'on se trouve dans une zone stable, I_p et I_q seront relativement faibles, et du même ordre de grandeur. En revanche, en zone transitoire, I_p aura tendance à augmenter, et I_q encore davantage (cf figure 1).

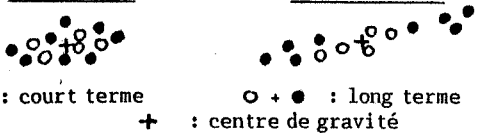
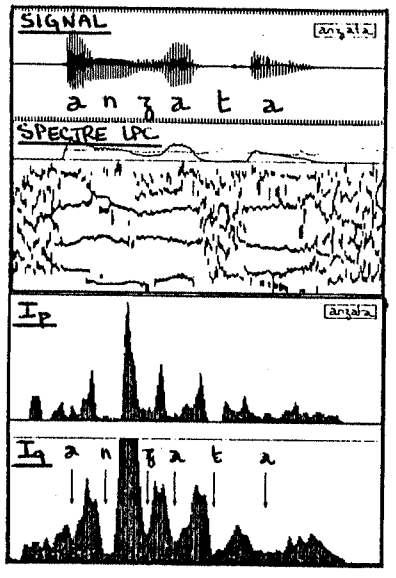


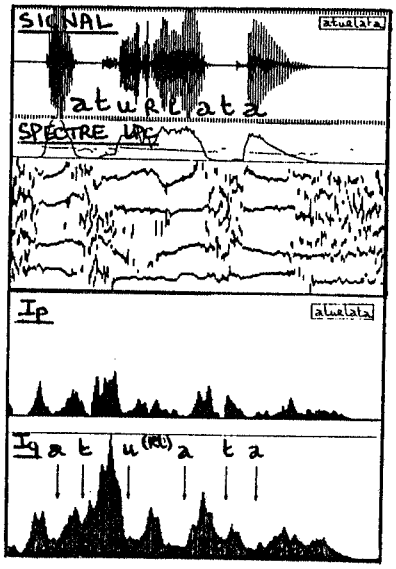
fig 1

Si p et q sont convenablement choisis, on peut espérer caractériser les zones de stabilité grâce à I_p et les maximums de stabilité grâce à I_q. Le recoupement des informations fournies par les deux critères parait d'ailleurs indispensable à toute décision. La figure 2 donne un aperçu des résultats obtenus.



court terme: 5 cs, long terme: 10 cs,
LPC 12 toutes les 1/2 cs.

fig 2-1



mêmes paramètres que fig 2-1.
R et l ne sont pas nettement détectés.

fig 2-2

3) localisation de zones stables:

On se propose d'appliquer ce critère au traitement de logatomes dont on connaît la transcription phonétique. L'expérience montre qu'un simple seuillage des inerties est mal adapté à la localisation des zones stables. A partir d'une représentation hiérarchique des pics et des vallées du critère de stabilité spectrale (Ehrich [3] par exemple), et avec l'aide de renseignements auxiliaires (énergie, passages par zéro, etc...), un logiciel d'étiquetage et de segmentation (au maximum de stabilité spectrale) va être développé.

4) le polyson:

L' étude du comportement du critère de stabilité spectrale recoupe des constatations antérieures sur certains phénomènes phonétiques: une majorité de phones offre une zone de stabilité plutôt franche, alors que d'autres laissent le critère inopérant. Dans cette optique, une nouvelle unité segmentale a été conçue: le polyson. Il est défini en tant que segment de parole aux frontières duquel apparaissent 2 zones de nette stabilité, correspondant par ailleurs à des phonèmes peu sensibles au contexte. Cette définition est légèrement différente de celle proposée par Laferrière [4].

En première approximation, un polyson peut donc être défini formellement comme une unité segmentale polyphonémique dont les frontières ne sont constituées ni par des liquides, ni par des semi-voyelles; on sait en effet que ces phonèmes sont très sensibles au contexte, et que leurs réalisations ne mettent en général pas de zones de stabilité spectrale en évidence.

Par exemple, dans le cadre de cette définition (somme toute provisoire), pi, as, pt, ao, ..., twa, tra, alt, ila, uji, ...,trwa, ajri, ..., arlwi, ..., sont des polysons, il, kl, pw, jp, rj, rl, ... n'en sont pas. A l'heure actuelle, en tenant compte d'un certains nombre de règles combinatoires spécifiques au français, on dénombre 7725 polysons. Toutefois, cette liste a priori sera nécessairement remaniée, à la lumière de l'expérience.

III-DECOMPOSITION TEMPORELLE.

Hormis les zones stables ou quasi-stables, l'essentiel du signal de parole est constitué de transitions entre ces états stationnaires. C'est la modélisation de ces transitions que ce paragraphe va maintenant aborder.

1) principe:

Soit $\vec{g}(n)$, le jeu de paramètres acoustiques (log area ratios, en l'occurence), observé pour une fenêtre indiciée n, au sein d'un segment de parole [0,N]. Soit $(\phi_k)_{1\leq k\leq p}$, une base de fonctions, chacune d'entre elles étant définie sur l'intervalle [0,N].

Une approximation $\vec{g}^*(n)$ de $\vec{g}(n)$ est fournie par l'expression:

$$\vec{g}^*(n) = \sum_{k=1}^{k=p} \phi_k(n) \cdot \vec{g}_k$$

La valeur approchée de $\vec{g}(n)$ est donc obtenue en tant que combinaison linéaire d'un nombre limité de valeurs de référence (p en l'occurence), les coefficients de cette combinaison linéaire étant des variables du temps (les fonctions ϕ_k).

Les deux méthodes suivantes diffèrent par le choix des fonctions ϕ_k:

2) méthode de Grenier:

Grenier [5,6] choisit une base **a priori**. Les fonctions de base sont **orthogonales** 2 à 2. Cette décomposition ne préjuge pas des propriétés particulières de la parole, et peut être appliquée à tous les signaux non-stationnaires.

3) méthode d'Atal:

Atal [7] adopte un autre point de vue. Pour cette méthode, les fonctions ϕ_k sont déterminées **a posteriori**, en tenant compte de propriétés spécifiques à la parole.

Une partie préliminaire de l'algorithme consiste donc à calculer les fonctions ϕ_k les mieux adaptées au segment traité; la seule condition imposée à ces fonctions est d'être **compactes**, c'est-à-dire partout nulles, sauf sur un intervalle. Cette contrainte s'étaye sur le principe que la zone d'influence d'un phone est localisée autour de sa cible, et par conséquent elle-même compacte. Une telle condition fait donc directement référence à des notions de phonétique acoustique et articulatoire.

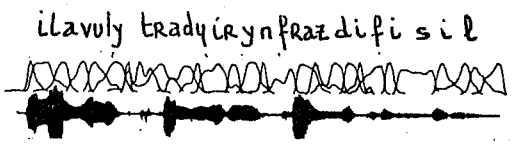
Les fonctions ϕ_k sont appelées "fonctions d'interpolation".

Marcus [8,9] a apporté des modifications en ce qui concerne le procédé de détermination des fonctions ϕ_k, mais la philosophie sous-jacente reste la même.

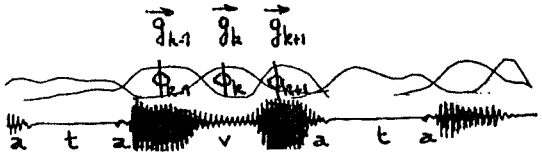
4) résultats:

Les figures 3 et 4 donnent un aperçu des résultats obtenus par l'algorithme d'Atal, modifié par Marcus.

On imagine facilement la considérable réduction de débit qu'une telle représentation temporelle peut apporter en codage. La restitution du signal a donc lieu en 2 étapes. D'abord une synthèse des paramètres LPC à partir des cibles ($\vec{g}_k$) et des fonctions d'interpolation (ϕ_k). Ensuite, une synthèse LPC classique à partir des paramètres obtenus eux-mêmes de façon synthétique.

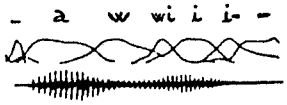


décomposition temporelle (Marcus).
"il a voulu traduire une phrase difficile".
signal et fonctions d'interpolation normalisées.



décomposition temporelle (Marcus): "atavata".

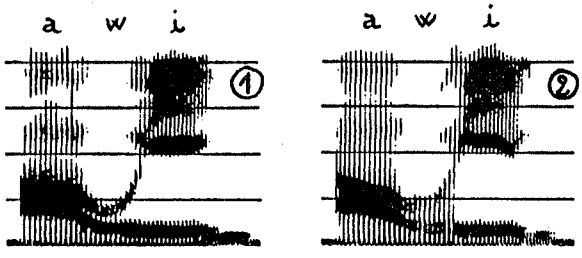
fig 3



décomposition temporelle (Marcus): "awi".

fig 4

La réduction de débit est de l'ordre d'un facteur 10. La différence de qualité entre les synthèses LPC à partir des paramètres originaux ou synthétiques est à peine audible. La comparaison des spectrogrammes respectifs vient appuyer cette constatation (fig 5).



1) spectrogramme signal synthétique, paramètres LPC originaux, 2) spectrogramme signal synthétique, paramètres LPC synthétiques.

fig 5

D'un autre point de vue, cette technique de décomposition temporelle fournit une description structurelle du signal de parole. Elle fait directement référence aux notions d' indice acoustique (cibles) et d'inertie articulatoire (fonctions d'interpolation). Le phénomène de coarticulation est donc pris en compte, modélisé, et même chiffré.

Dans un futur assez proche, on envisage de faire intervenir le critère de stabilité spectrale au sein de l'algorithme de décomposition temporelle. Une connaissance a priori des zones franchement stables permettrait en effet de guider la recherche des fonctions de base principales, le reste des fonctions étant obtenu lors d'une deuxième passe. Outre une possible accélération de l'algorithme, un tel procédé permettrait d'éviter un certain nombre de décompositions artificielles, peut être plus performantes du point de vue du codage, mais moins conformes à la réalité phonétique du signal.

IV-REPRESENTATION DE L'INFORMATION PHONETIQUE.

La technique de décomposition temporelle d'Atal offre des perspectives nouvelles en représentation de l'information inhérente au signal de parole. Un certain nombre d'ouvertures sont possibles, plus particulièrement en synthèse.

1) synthèse segmentale:

Un tel codage possède de nombreuses qualités. Il est plus économique que le codage LPC classique, et permet donc de stocker beaucoup plus de segments à encombrement égal. Par ailleurs, il offre une bien plus grande maniabilité en particulier pour l'application des règles prosodiques (allongement, accentuation, débit, ...).

Le polyson, dans cette optique, paraît être une unité segmentale particulièrement bien adaptée. Ses frontières sont des zones de stabilité spectrale (une seule fonction d'interpolation), qui correspondent à des phonèmes peu sensibles au contexte (cibles quasiment identiques pour un même phonème). La concaténation pourrait donc s'opérer sans difficultés majeures, au moyen d'un simple lissage (fig 6).

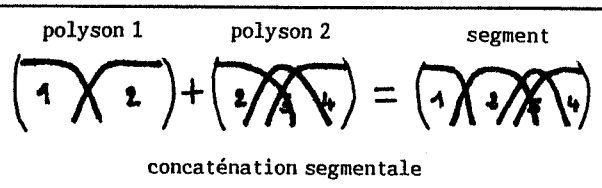


fig 6

Enfin, une description valable d'unités segmentales peut conduire à des applications en codage ou en reconnaissance [10].

2) synthèse par règles:

Par application de techniques de quantification vectorielle, on peut imaginer créer un dictionnaire de cibles. Il est utopique d'espérer obtenir une seule cible par phonème; il apparait plus probable de déboucher sur quelques centaines d'unités, auxquelles pourraient être associées une étiquette. Certaines seraient monophonétiques (avec parfois plusieurs étiquettes pour des réalisations différentes d'un même phonème), d'autres seraient diphonétiques (losqu'une cible et une fonction d'interpolation supplémentaires sont nécessaires pour modéliser une transition). Les applications seraient nombreuses, tant en codage qu'en synthèse, voire en reconnaissance.

Par ailleurs, on peut espérer modéliser les fonctions d'interpolation elles-mêmes. Une approximation trapézoidale (3 segments de droite par fonctions) suffira peut-être, mais si cela s'avère nécessaire, on aura recours à d'autres fonctions (trigonométriques surélevées, par exemple). Il est par ailleurs probable que le modèle de fonction proposé tiendra compte du phone considéré, de son contexte, du débit du segment dans lequel il intervient, etc...

A plus long terme, on peut donc envisager de développer une nouvelle technique de synthèse par règles, ces règles intervenant au niveau du choix des cibles, et de la génération des fonctions d'interpolation. Une étude prospective devrait être menée dans ce sens.

V-CONCLUSION.

Outre le domaine de la synthèse, la méthode d'Atal peut donner lieu à de nombreuses applications:

- en codage, bien entendu, puisqu'elle permet une considérable réduction de débit, à qualité similaire. Dans ce cadre, il n'est pas important qu'il y ait étroite correspondance entre les résultats du codage et la structure acoustico-phonétique du signal.

- en reconnaissance éventuellement. Sous cet angle, la parfaite cohérence entre décomposition temporelle et événements phonétiques paraît primordiale.

L'application de cette technique de décomposition temporelle dans le cadre de l'étiquetage et de la segmentation de logatomes connus sera l'occasion de la mettre à l'épreuve.

BIBLIOGRAPHIE

[1] D VICARD, L MICLET (1986), Steady part recognition of continuous speech for acoustic-phonetic decoding, in proc. ICASSP, Tokyo (à paraître).

[2] R VISWANATHAN, J MAKHOUL (1975), Quantization properties of transmission parameters in linear predictive systems, IEEE trans. ASSP, vol 23, p 309-321, juin 1975.

[3] R W EHRICH, J P FOITH, (1976), Representation of random waveforms by relational trees, IEEE trans. COMP 25:7, p 725-736.

[4] P LAFERRIERE, G CHOLLET, L MICLET, J.P TUBACH (1985), Segmentation d'une base de données de "polysons", application à la synthèse de la parole, XIVe JEP, GALF, Paris.

[5] Y GRENIER, (1983), Time-dependant ARMA modeling of non-stationary signals, IEEE trans. ASSP 31:4, p 899-911.

[6] M.C OMNES, Y GRENIER, G CHOLLET (1985), Synthèse de parole à l'aide des fonctions d'aire logarithmiques évolutives, XIVe JEP, GALF Paris.

[7] B.S ATAL (1983), Efficient coding of LPC parameters by temporal decomposition, IEEE-ICASSP, p 81-84, Boston.

[8] S.M MARCUS, R.A.J.M VAN LIEHOUST (1984), Temporal decomposition of speech, IPO annual progress report 19.

[9] S.M MARCUS, Decoding the speech code (1986), soumis à JASA.

[10] G CHOLLET, J.F GALLIANO, J.P LEFEVRE, E VIARA (1983), On the generation and use of a segment dictionnary for speech coding, synthesis and recognition, IEEE-ICASSP, p 1328-1331, Boston.

LE DECODAGE ACOUSTICO-PHONETIQUE A L'AIDE SYSTEME EXPERT SERAC-IROISE

A. Bonneau*, G.Mercier*, M. Gérard*, M.Rossi**

* CNET Lannion - ** Institut de Phonétique d'Aix-en-Provence

ABSTRACT

The implementation of acoustic-phonetic decoding rules into the SERAC-IROISE Expert System is presented. This phonetic recognition is decomposed into a sequence of step activations : initialization, labelling, segmentation and primary phonetic feature recognition. The main phases of these step activations are summarized in Fig. 1.

A great advantage of this last approach is the explanation component which proves to be very useful for understanding the errors of recognition and consequently for improving rules and cues.

Preliminary results of recognition of phonetic classes automatically extracted from numbers uttered by eight speakers are given.

INTRODUCTION

Le décodage acoustico-phonétique correspond à cette étape de la connaissance automatique de la parole qui cherche à transformer le signal acoustique en une suite d'unités pseudo-phonétiques discrètes (syllabes, phonèmes, phones). C'est la première étape de l'interprétation linguistique du signal.

Pour prendre en considération la complexité et la variabilité de celui-ci, on définit des unités intermédiaires (généralement infra-phonémiques) appelées phones qui correspondront soit à des évènements acoustiques facilement localisables soit à des variantes allophoniques d'un phonème (segments homogènes).

On sait d'autre part que les unités linguistiques et en particulier les phonèmes peuvent être décrits par des traits phonétiques, articulatoires ou perceptifs (JAKOBSON et Al, 1951 ; AUTESSERRE et ROSSI, 1985). Ces traits possèdent l'avantage d'être plus facilement reliés à l'onde acoustique ou à la position et aux mouvements des organes articulatoires. Dans ce cas on peut mettre en oeuvre un processus de reconnaissance de traits qui cherchera à localiser sur une fenêtre de signal des indices ou des primitives de base dont l'utilisation dans les règles d'interprétation facilitera l'identification du trait correspondant.

Les expériences de lecture automatique de spectrogrammes (CARBONNELL et Al, 1983 ; ZUE, 1979) ont montré que les humains étaient capables de segmenter correctement et ensuite (ou conjointement) d'interpréter la suite de ces évènements acoustiques. Il semble que ces experts utilisent à la fois des règles, des procédures et tout un ensemble de connaissances acoustiques, phonétiques, perceptives, articulatoires, phonotactiques et prosodiques, pas toujours faciles à formaliser, pour élaborer progressivement les différentes hypothèses phonétiques.

Le développement récent de l'intelligence artificielle a favorisé l'éclosion d'outils facilitant les problèmes de représentation, de formalisation, d'acquisition et de manipulation de connaissances, d'où l'apparition, ces dernières années de systèmes experts utilisant le formalisme des règles de production pour la modélisation des connaissances acoustiques, phonétiques, pour le décodage (STERN et Al, 1986 ; CARBONNEL et Al, 1983), et pour la mise au point de nouveaux modèles de reconnaissance de la parole (MELONI, 1983).

ORGANISATION DU SYSTEME EXPERT SERAC-IROISE

Les trois composantes essentielles du système IROISE sont : la base de données, le moteur d'inférences et l'interface utilisateur (GILLOUX, 84).

La base de connaissances se compose de :

- une base des règles dont les subdivisions correspondent à chaque problème d'un module ;
- une base des objets représentant des faits ou hypothèses (connaissance incertaine), divisée en fonction des classes d'objets distincts ;
- une pile des problèmes ou agenda : à un instant donné, l'ordre de priorité des problèmes à résoudre ;
- un ensemble de fonctions Lisp, de paramètres et de fichiers de données.

ORGANISATION GENERALE DU MODULE DE DECODAGE

L'interface phonétique d'une phrase se fait

séquentiellement de gauche à droite et par étapes. Chacune de ces étapes correspond à une phase importante de l'analyse phonétique :

1. Lecture des échantillons spectraux et calcul de leurs paramètres : centres de gravité, dérivées, énergies hautes et basses fréquences, ...
2. Détection du début de phrase.
3. Première interprétation phonétique à partir des spectres centisecondes : voyelle-consonne, nasal, fricative, liquide, sourd, voisé.
4. Segmentation en pseudo-syllabes.
5. Détection du noyau vocalique, détection des frontières de syllabes et calcul des attributs de syllabes.
6. Détection des zones stationnaires et transitoires entre les noyaux vocaliques.
7. Calcul des attributs phonétiques de ces évènements acoustiques : silence, barre de voisement, mode et lieu d'articulation...

Comme on le voit chacune de ces étapes est elle-même une suite de problèmes qu'un ensemble limité de règles peut résoudre. Seul cet ensemble est activé lors de la résolution d'un problème. Chaque règle d'interprétation procède à partir des données et des hypothèses construites durant les étapes précédentes, les enrichit, les modifie, construit de nouvelles hypothèses et met à jour la pile des problèmes. Ces différentes étapes sont itérées au fur et à mesure que la reconnaissance de la phrase progresse jusqu'à ce que la fin de la phrase soit détectée.

L'enchainement suivant un ordre prédéterminé des problèmes met en évidence dans l'analyse phonétique un premier niveau d'interprétation des macro-classes basé sur des règles mettant en jeu des indices spectraux, un second niveau où les indices temporels sont les plus importants et un troisième niveau où les indices spectraux, temporels et contextuels entrent en action pour une interprétation phonétique de plus en plus fine (figure 1).

ETIQUETAGE CENTISECONDE ET SEGMENTATION

Détection du début de phrase

Le problème est résolu par le déclenchement de règles comparant les niveaux d'énergie dans diverses bandes de fréquence à des seuils normalisés (afin de détecter la présence de parole) et aussi au maximum d'énergie calculé sur une fenêtre temporelle supposée contenir la première voyelle.

Etiquetage centisecondes
Etiquetage voyelle-consonne

L'attribution de l'étiquette voyelle ou consonne à un spectre met en jeu les indices suivants :

$E_{[\alpha-\beta]}$: énergie dans la bande de fréquence $[\alpha-\beta]$

$M_{[\alpha-\beta]}$: amplitude du maximum d'énergie dans la bande $[\alpha-\beta]$

CDG : centre de gravité spectrale du spectre.

Formax : amplitude du plus grand formant sur toute la partie de phrase déjà analysée.

$Min_{[\alpha-\beta]}$: amplitude du minimum d'énergie dans la bande $[\alpha-\beta]$

$F_{[\alpha-\beta]}$: fréquence du maximum d'énergie dans la bande $[\alpha-\beta]$

Une vingtaine de règles comparant ces indices entre eux ou à des seuils normalisés permet la distinction voyelle-consonne.

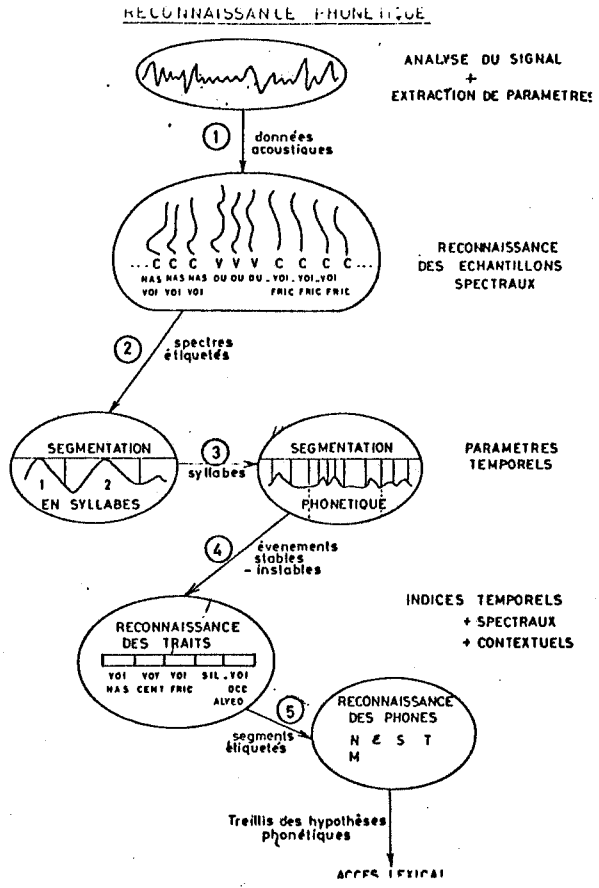


Figure 1

La détection du trait de voisement

La distinction consonne voisée-consonne sourde est basée sur les indices suivants : valeur du fondamental (fo > o), rapport de l'énergie basse fréquence à l'énergie haute fréquence et rapport entre l'énergie basse fréquence et l'énergie totale.

Détection des fricatives

Les principaux indices facilitant la reconnaissance de cette classe de consonnes sont le nombre de passages par zéro de la dérivée du signal, les rapports entre l'énergie haute et basse fréquence et la valeur du centre de gravité spectrale. Les consonnes [f, s, ʃ] et [ʒ, z] sont bien détectées par ces indices. La fricative v plus difficile à détecter est caractérisée en plus par un spectre diffus et plat au-dessus de 300 Hz et sera détecté par des indices mesurant cette propriété (différence min-max haute fréquence faible).

La détection des consonnes nasales

Ces consonnes sont caractérisées par une forte résonnance entre 200 et 300 Hz. La comparai-

son de cette énergie aux énergies supérieures constitue la règle principale de cette classification.

La détection des consonnes liquides

Cette détection en est encore au stade du balbutiement ; pour le moment la position du maximum d'énergie spectrale et la comparaison des énergies basse, moyenne et haute fréquence sont les indices utilisés.

Ces traits sont pour le moment les seuls traits que l'on recherche au niveau du spectre centiseconde.

Segmentation en pseudo-syllabes

On utilise comme indices principaux l'évolution temporelle de trois courbes d'énergie : l'énergie du signal, l'énergie dans la bande 250-4200 Hz et l'énergie dans la bande 250-850 Hz, on recherche les maxima et minima de cette courbe. Dans une première étape le signal est coupé aux minima à condition qu'entre deux minima il y ait un pic d'une durée et d'une amplitude suffisantes.

Dans une deuxième étape on vérifie si ces segments possèdent un noyau vocalique. Pour cette vérification, on calcule les indices suivants :
1. différences des énergies totales et basses fréquences et des amplitudes du premier formant calculées aux points culminants du segment et de la syllabe précédente et aux frontières (creux des vallées),
2. le nombre d'échantillons étiquetés "voyelle" du segment,
3. le nombre d'échantillons voisés consécutifs entourés d'échantillons sourds.

Une combinaison linéaire de ces indices nous permet d'attribuer une plausibilité à la présence ou non d'un noyau vocalique. Si ce segment ne contient pas de noyau vocalique il est regroupé avec le segment potentiellement syllabique suivant.

Segmentation en phones

Les voyelles sont localisées par la zone stable entourant le maximum d'énergie.

Entre ces zones vocaliques on repère les suites de zones de stabilité et d'instabilité dont les frontières sont détectées par la dérivée du spectre. Ces segments sont le cadre à l'intérieur duquel une identification plus précise des traits phonétiques est effectuée.

RECONNAISSANCE DES VOYELLES

Elle intervient après le module de détection des noyaux vocaliques et s'effectue selon un mode binaire dans une arborescence où tous les indices sont hiérarchisés. Au total, 37 indices qui peuvent généralement être considérés comme des indices d'acuité, d'ouverture, de bémolisation, de nasalité et périphériques sont utilisés. Ils sont fondés sur les variations d'énergie dans le spectre et calculés soit sur toute la durée de la voyelle, soit sur une partie de celle-ci - la partie stable généralement-.

Voici deux des principaux indices utilisés :

Si (E 250-450 Hz > = E 650-1050 Hz) sur plus de 50 % d'ouverture nommé "ouv1" est négatif.

Si (Max-E 1600-3400 Hz > = Max-E 650-1600 Hz) sur plus de 50 % des échantillons de la voyelle, l'indice d'acuité nommé "aigu1" est positif.

Ei-J représente l'énergie calculée dans la bande de fréquences limitée par les valeurs de i et j.
Max-Ei-j représente l'énergie maximale dans la bande de fréquences limitée par les valeurs de i et j.

Dans Sérac, les indices et les règles de reconnaissance sont écrits séparément. Chaque règle représente un chemin de l'arbre.

Voici, comme exemples, la règle Sérac chargée de reconnaître l'indice de bémolisation "mbl" et une règle de reconnaissance de la voyelle e ouvert.

```
(defregle ro-mbl
          échantillon
          mbl
Si
    (phone-voy ?pv (début ?d) (fin ?f))
Et
    (ou ( > (énergie-bande-temps 1600 2800 ?d ?f)
            (énergie-bande-temps 2800 4300 ?d ?f))
        ( > (énergie-bande-temps 1600 1900 ?d ?f)
            (* 1/2 (énergie-bande-temps 3100 3900
                                        ?d ?f))))
Alors
    (modifier (phone-voy ?pv) (mbl vrai))
    (nouveaupb (reconnaissance-voyelle)))
```

Explication : l'indice "mbl" de la voyelle PV, délimitée par son début "d" et la fin "f" est vrai quand $\sum_{i=d}^{f}$ Ei 1600-2800 Hz > $\sum_{i=d}^{f}$ Ei 2800-4300 Hz ou $\sum_{i=d}^{f}$ Ei 1600-1900 Hz > 1/2 ($\sum_{i=d}^{f}$ Ei 3100-3900 Hz).

```
(defregle r0-ouv-aig
          échantillon
          ouv-aig
Si
  (phone-voy ?pv (ouv1 vrai) (ouv4 vrai)
                 (aigu1 vrai) (aigu4 vrai)
                 (aigu5 vrai))
Et
Alors
  (modifier (phone-voy ?pv) (voyelle 'ai))
  (nouveaupb (ouv-grave)))
```

Explication : S'il existe une voyelle représentée par la variable ?pv (voyelle courante) et dont les indices "ouv1", "ouv4", "aigu1" sont vrais, alors la voyelle est probablement "ai" (e ouvert).

Une première évaluation du système de reconnaissance des voyelles a donné les résultats suivants pour deux locuteurs masculins :
-79 % de reconnaissance correcte pour les phrases phonétiquement équilibrées de Combescure,
- 76 % pour les nombres connectés.
Le système fournit en moyenne 1,6 réponses pour chaque voyelle à reconnaître.

Une nouvelle évaluation a été effectuée sur un corpus de nombres allant de 0 à 999 et prononcés par 6 locuteurs masculins et trois locutrices et des matrices de confusion entre les voyelles, mais aussi entre les grandes classes vocaliques ont été établies systématiquement.

Il convient de séparer les résultats obtenus pour les locuteurs masculins de ceux obtenus pour les locutrices, les indices ayant été créés à l'aide d'une base de données ne comportant que des voix masculines.

Pour les locuteurs masculins, le % de reconnaissance correcte s'élève à 72 %, toutes réponses comprises, et à 50 % en première réponse.

Scores obtenus pour chaque voyelle :

Voyelle	i	ei	ai	a	y	eu	oe	u	of	oo	on	in	an
%		77	54	71	73	42	72	58	74	66	56	72	75

ei : e fermé, ai : e ouvert, of : o fermé, oo : o ouvert.

Scores obtenus pour la reconnaissance des classes vocaliques :
Voyelles aigues : 96 % (in et oe ne sont pas
..... graves : 97 % classées)
..... ouvertes : 99 % (les voyelles d'ouverture moyenne dont la réalisation est délicate à juger ne sont pas classées sur l'axe ouvert-fermé).
Voyelles fermées : 97 %
..... orales : 82 %
..... nasales : 63 %
..... bémolisées : 43 %
.. .. diésées : 87 %.

Pour les locutrices, le % de reconnaissance des voyelles ne dépasse pas 53 %, en revanche les classes vocaliques sont relativement bien reconnues.

Ces résultats confirment l'existence pour les grandes classes vocaliques de caractéristiques multilocuteurs très solides et relativement invariantes alors que, à l'intérieur de ces classes, les distinctions fines entre les voyelles sont davantage tributaires du locuteur. Dans certains cas, l'ambiguité peut être levée grâce à la connaissance du contexte : c'est l'objet des futurs travaux qui vont être menés sur les voyelles.

RECONNAISSANCES DES CONSONNES - RESULTATS

Une première estimation de la reconnaissance des principaux traits effectués sur des nombres aléatoires compris entre 0 et 999 et prononcés par 4 locuteurs masculins et 4 locuteurs féminins donne les résultats suivants :

Le pourcentage de détection du trait fricatif pour les consonnes [f, s, v, z] les seules présentes dans le corpus, est supérieur à 94 % (10 erreurs sur 168 phonèmes).

Le pourcentage de reconnaissance du trait occlusif est de 91 % (10 erreurs sur 115).

Le pourcentage de reconnaissance du trait de voisement (voisé ou sourd) est supérieur à 90 % (5 % d'erreur et 5 % de cas d'indétermination où aucune décision n'est prise).

Enfin le pourcentage de reconnaissance du lieu d'articulation des occlusives sourdes et des fricatives présentes dans le corpus n'est que de 76 % et demande à être amélioré.

CONCLUSIONS

Le système SERAC écrit en Common Lisp et en de Lisp est opérationnel soit sous VMS soit sous UNIX. Grâce à ce système, nous avons déjà pu améliorer un certain nombre de règles du système complet de reconnaissance de la parole KEAL et y introduire de nouvelles connaissances.

Malgré tout, comme les résultats nous l'ont montré, il reste encore à introduire dans le système Serac tout un ensemble de règles, à les mettre au point et à les évaluer.

REFERENCES

[1] D. Autesserre, M. Rossi, "Propositions pour une segmentation et un étiquetage hiérachisé ; application à la base de données acoustiques du GRECO communication parlée, 1985.

[2] A. Bonneau, M. Rossi, G. Mercier, "Hierarchical Recognition of French Vowels by the SERAC-IROISE Expert System" Actes du Séminaire Franco-Suédois, Grenoble, 1986.

[3] N. Carbonnel, J.P. Haton, J.M. Pierrel et F. Lonchamp, "Elaboration d'un système expert pour le décodage phonétique automatique ; Speech Com. Vol 2 n° 2-3 pages 231-233, 1983

[4] M. Gilloux, G. Mercier, C. Tarridec, "Un Système Expert pour la reconnaissance de la parole, Congrès AFCET, R.F et IA, pages 99-111, 1984.

[5] R. Jakobson, G. Fant et M. Halle, "Preliminaries to Speech Analysis", M.I.T. Press, Cambridge, MA, 1951.

[6] H. Meloni, J. Guizol, "Identification d'évènements pseudo-phonétiques pour la reconnaissance automatique de la parole ; Speech Com. Vol 2, n° 2-3, p. 211-214, 1983.

[7] G. Mercier, M. Gilloux, C. Tarridec, J. Vaissière, "A New Rule-Based Expert System for Speech Recognition" in New System and Architectures for Automatic Speech Recognition and Synthesis, De Mori and Suen Nato ASI Series F, Vol 16, 1985.

[8] P.E. Stern, M. Eskenazi, D. Memmi, "An Expert System for Speech Spectrogram reading, Proc. IEEE ICASSP, 1986 TOKYO.

[9] V.W. Zue, R. Cole, "Experiments on Spectrogram reading ; Conference Record, IEEE-ICASSP, p. 116-119, 1979.

SEGMENTATION ET ETIQUETAGE POUR UN SYSTEME DE RECONNAISSANCE AUTOMATIQUE MULTILOCUTEUR

F. Delaire & M. Rossi

Institut de Phonétique d'Aix-en-Provence, UA CNRS 261 'Parole et Langage'

ABSTRACT

The aim of the present study is to describe some devices for automatic multispeaker phonemic segmentation and labelling.
For segmentation, the value of the system lies in its ability to determine stable borders; the strategy requires both marking the acoustical events in the signal and using parameters computed at different levels of analysis.
For labelling, macroclasses are first recognized for both the centre and the sides of the segment. Then, contextual and coarticulatory rules are applied to define a phonetic interpretation and a macroclass interpretation if possible.

INTRODUCTION

La segmentation est considérée comme la première opération prioritaire pour le décodage acoustico-phonétique. Le problème se pose de savoir si cette étape est justifiée quand on sait que le signal est un continuum qui ne reflète pas le caractère discret des unités des unités phonologiques. Pourtant si les matrices de traits du niveau des représentations phonétiques sont bien des ordres aux effecteurs du système d'exécution, ces dernières doivent laisser des traces dans le signal, c'est-à-dire des discontinuités acoustiques et phonétiques (Abry et al. 1985.).
Ces événements doivent donc être pris en compte, mais nous savons qu'ils ne se situent pas forcément aux frontières des unités phonétiques. Cependant, si nous créons des frontières aux seuls événements significatifs qui ne créent pas de sursegmentation au sein des unités, nous sommes nécessairement conduits à sous-segmenter le continuum. Comment donc dans ce cas identifier toutes les unités reflétées dans le signal ?
La méthode qui permet d'apporter une solution à ce problème implique une stratégie fondée sur un multiplexage d'informations de différents niveaux (fig. 1) analysées par l'homologue d'un système expert.

I. PRESENTATION DES DONNEES

Le signal de parole est numérisé à une fréquence de 16 kHz, avec un pas de 10 ms. Pour chaque tranche de signal d'une durée de 16 ms, nous évaluons :
1. Le nombre de passages par zéro,
2. L'énergie totale,
3. Les énergies contenues dans 14 bandes de fréquences à l'aide d'un banc de filtres simulé à résolution spectrale de 65 Hz, comportant un fenètrage de Hamming, une préaccentuation à 500 Hz.

II. SEGMENTATION

On peut être conduit à sursegmenter le signal à l'aide d'une fonction d'instabilité dans laquelle sont inclus des paramètres acoustiques et phonétiques, on identifie alors des segments subphonémiques qui sont considérés comme des phones homogènes (Caelen et al. 1981, 1985).
La plupart des systèmes existants choisissent de sursegmenter le signal de parole.
Quant à nous, au lieu de discrétiser intempestivement le signal, nous sommes amenés à définir des indicateurs de frontières dénués d'ambiguïté sur l'organisation interne ou sur les frontières des unités phonétiques. Ils résultent d'un traitement d'informations hiérarchisées, de nature phonétique ou acoustique.

La recherche de ces informations s'effectue en plusieurs étapes :
1. Evaluation de paramètres,
2. Description de l'évolution de l'énergie,
3. Définition d'un précodage en macroclasses,
4. Etude de la stabilité des échantillons,
5. Présegmentation,
6. Reconstitution des segments phonétiques.

Paramètres

Nous caractérisons l'énergie à l'aide de plusieurs types de calculs : l'énergie totale : IRMS, la somme des énergies contenues dans les quatorze canaux : EN, l'énergie des basses fréquences : ENB, une fonction qui amplifie les variations sur les canaux : EN53.
Nous calculons, en outre, trois types d'instabilité : INSTA, DERIV, DERIVK.
Des indices acoustiques définissent des événements phonétiques tels que la tenue, la détente, la détente longue, l'explosion, la frication, le silence, la transition.

Certains de ces indices apportent déjà une information sur la nature consonantique ou vocalique du segment. Cette information sera validée par l'application de règles phonétiques qui reconstituent les unités phonétiques (cf. plus loin); elle est indépendante et parallèle à l'information apportée par MAX12 et MAXEN (cf. ci-dessous).
Six des indices phonétiques (2 d'ouverture, 4 d'acuité) définis par Rossi et al. (1986) sont évalués sur chaque échantillon. La fluctuation de ces indices fournit une instabilité de nature phonétique : HOMO.

Suivi de l'énergie

a. Nous procédons à un repérage des extrema sur la fonction d'énergie IRMS, nous codons un vecteur de variations à l'aide de INSTA, enfin, ces traitements effectués, nous définissons une proposition de codage reflétant la position sur minimum ou maximum d'énergie.
b. Nous vérifions les résultats obtenus en confrontant les extrema observés sur EN et validons les résultats antérieurs si nous observons une corrélation; sinon nous affectons sur le segment erroné un indicateur d'erreur; nous obtenons ainsi MAXEN.
c. Nous définissons ensuite un vecteur, image du comportement de l'énergie des basses fréquences soit MAX12.

Définition des macroclasses

A l'aide des vecteurs MAX12 et MAXEN, nous définissons la macroclasse syntagmatique de chaque tranche de signal soit une classification en suites :
. vocaliques,
. consonantiques,
. consonnes vocaliques,
. transitoires.

Stabilité des échantillons

Elle s'évalue en tenant compte, d'une part, de la variation de la quantification des indices phonétiques calculés, d'autre part, des changements de polarités observés sur les vecteurs MAXEN et MAX12 et permet de constituer les éléments de INSTA4 qui reflètent les discontinuités relevées et leur provenance.

Présegmentation

A l'aide des vecteurs associés à INSTA4, DERIVK, DERIV et INSTA, nous sommes en mesure de définir des frontières de différents niveaux :
1,2,3,4 pour les frontières sûres
5,6 pour les frontières potentielles.
Ces frontières définies correspondent à une corrélation de plusieurs marques de discontinuité; suivant leur importance, elle affecte un degré de fiabilité variable.

Reconstitution des segments phonétiques

La reconstitution des segments phonétiques se subdivise en deux traitements distincts :
- Traitement de reconstructions consonantiques,
- Traitement de repérage de segments vocaliques.
Le premier traitement s'effectue sur les échantillons marqués : transition, consonantique. Il résulte de l'application de règles phonétiques relatives à la succession temporelle des événements acoustiques (ex. : tenue, explosion, ...).
Le second porte sur les échantillons marqués : transition, vocalique ou consonne vocalique. Il procède à une recherche de frontières éventuellement omises à l'aide de la fonction d'énergie EN53. Si nécessaire, il procède à la validation d'une frontière supplémentaire avec réattribution adéquate des classifications en macroclasses. Puis, il réexamine toutes les frontières. Si des incohérences sont relevées à partir de la stabilité du segment considéré HOMO ou de EN53, la frontière proposée sera effacée ou plutôt réduite à un simple événement intraphonémique.

III. RECONNAISSANCE EN MACROCLASSES

La segmentation précédente fournit un ensemble de segments phonétiques préétiquetés en macroclasses syntagmatiques : voyelle, consonne vocalique, consonne, transition.
Selon la nature des segments, un ensemble de vingt-deux règles phonétiques seront appliquées pour la reconnaissance des macroclasses paradigmatiques (Rossi et al. 1986). Les règles mises en œuvre pour la reconnaissance des unités vocaliques sont appliquées, d'une part sur une zone centrée homogène du segment, d'autre part sur les parties stables, si elles existent, des zones excentrées. Elles permettent ainsi d'apporter une information :
1 - sur la classe du segment vocalique,
2 - sur les transitions des segments en amont et en aval.
Les classes sont définies par un faisceau d'indices phonétiques (cf. plus haut); au nombre de dix-neuf, elles renvoient en moyenne à 2,5 candidats éventuellement accompagnés d'information contextuelle.
Exemple :
Classe 16: u,eO,o/dental.
En ce qui concerne la reconnaissance des macroclasses consonantiques, des règles concatènent les événements phonétiques définis sur les indices acoustiques (SIL : silence, DET : détente, ...) et leur affectent les classifications :
. consonne occlusive sourde
. consonne occlusive sonore
. consonne constrictive sourde
. consonne constrictive sonore
. consonne vocalique.
Cette partie est en cours d'élaboration. La reconnaissance paradigmatique n'est pas encore mise en œuvre; mais elle utilisera en particulier les indices définis par A. Bonneau pour les occlusives (A. Bonneau 1984).

CONCLUSIONS

La méthode que nous venons de présenter fournit de bons résultats. Pour les macroclasses syntagmatiques et paradigmatiques sur un corpus de 120 mots, pour quatre locuteurs distincts (hommes ou femmes), ne comportant pas de groupes consonantiques, aucune frontière erronée n'a été introduite. Les unités omises sont prises en compte par le module d'interprétation phonétique. Toutefois nous ne donnerons

pas de taux d'erreurs : le corpus étant trop restreint, il n'aurait pas de signification. Quant aux macroclasses paradigmatiques, elles contiennent toutes les voyelles à reconnaître, mais les candidats ne sont pas hiérarchisés.

BIBLIOGRAPHIE

Abry, C., Benoit, C., Boe, L.J. & Sock, R. (1985), "Un choix d'événements pour l'organisation temporelle du signal de parole", **Actes des 14èmes J.E.P.**, Paris, pp. 133-138.

Bonneau, A. (1984), **Indices de reconnaissance des consonnes occlusives sourdes du français**, thèse de 3e cycle, Aix.

De Mori, Renato (1983), **Computer Models of Speech using Fuzzy Algorithms**, Plenum Press, New York.

Mercier, G. (1981), "Acoustic processing and phonetic analysis in continuous speech recognition", **Proceedings of the Fourth FASE Symposium on Acoustics and Speech**, Venise, pp. 87-109.

Mermelstein, P. (1975), "A phonetic-context controlled strategy for segmentation and phonetic labelling of speech", **IEEE Trans. A.S.S.P.**, 23, pp. 79-82.

Perennou, G. & De Calmes, M. (1985), "Segmentation en événements phonétiques et en unités syllabiques", **Actes des 14èmes J.E.P.**, Paris, pp. 142-147.

Rossi, M., Bonneau, A. & Grenié, M. (1986), **Automatic Recognition of Vowels and Stops using Acoustic Cues** (à paraître).

Vigouroux, N. & Caelen, J. (1985), "Segmentations en vue d'une base de données acoustiques et phonétiques", **Actes des 14èmes J.E.P.**, Paris, pp. 152-156.

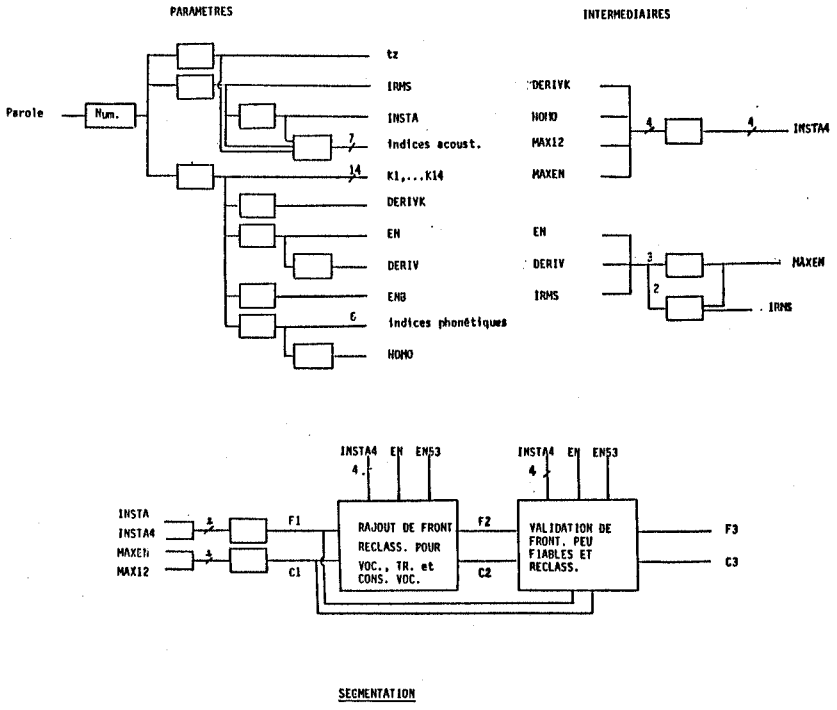


Figure 1
Présentation des paramètres et segmentation mise en œuvre

Nota : Les rectangles représentent des modules de traitement.

tz	Nombre de passages par zéro
IRMS	Energie totale
INSTA	Dérivée de l'énergie totale
K1,...K14	Energie contenue dans chaque bande de fréquences
DERIVK	Dérivée de l'énergie des 14 canaux
DERIV	Dérivée de EN
ENB	Energie des basses fréquences, fournit MAX12
EN53	Fonction de EN et de K1,...K14.

HL51 M1 bloiiiiidi

dur/flobase/M1/bloiiiiidi

n	mMe53	tz	i	53	ns	nd	f1	c1	f2	c2	c3	f3	o1	o4	a1	a2	a3	a5
1	-1110	2	0	1	0	0	0	27	0	27	0	27	-1	-2	0	1	0	0
2	-100	4	1	2	0	0	0	27	0	27	0	27	-6	-101		1	-1	-1
3	-10	2	1	3	0	0	0	27	0	27	0	27	-7	-124		-1	-4	-4
4	0	1	0	4	0	0	0	27	0	27	0	27	-4	-7	1	-1	-4	-6
5	0	1	0	5	0	0	0	27	0	27	0	27	-3	-6	0	-1	-4	-7
6	0	1	0	9	0	0	0	27	0	27	0	27	-3	-7	1	-1	-5	-8
7	0	2	2	27	0	0	0	27	0	27	0	27	0	-3	-1	-1	-5	-9
8	0	1	2	46	0	0	0	10	0	10	0	10	4	3	1	-1	-4	-2
9	0	3	0	64	0	0	1	3	1	3	1	3	8	7	-1	-1	-7	-9
10	0	5	0	69	0	0	0	3	0	3	0	3	7	7	2	-1	-7	-7
11	0	5	0	71	0	0	0	3	0	3	0	3	9	9	0	-1	-7	-9
12	0	5	0	71	0	3	0	3	0	3	0	3	11	12	0	-1	-5	-5
13	10	5	1	71	0	3	0	3	0	3	0	3	11	12	0	-1	-4	-3
14	100	5	1	71	0	3	0	3	0	3	0	3	11	12	-1	-1	-5	-4
15	1000	5	0	72	0	3	0	3	0	3	0	3	13	16	-2	-1	-5	-4
16	100	5	0	70	0	3	0	3	0	3	0	3	13	16	-2	-1	-4	-4
17	10	5	0	68	0	3	0	3	0	3	0	3	13	16	-3	-1	-4	-5
18	0	5	0	66	0	3	0	3	0	3	0	3	11	14	-2	-1	-3	-4
19	0	5	2	63	0	3,	0	3	0	3	0	3	10	13	0	-1	-4	-1
20	0	4	2	57	0	0	0	3	0	3	0	3	7	10	3	-1	-5	2
21	0	2	0	50	0	0	0	3	0	3	0	3	5	6	3	-1	-4	5
22	0	3	2	44	0	0	0	3	0	3	0	3	2	0	4	-1	-5	7
23	0	2	3	39	0	0	0	3	0	3	0	3	2	0	-1	1	4	6
24	0	3	2	34	0	0	0	3	0	3	0	3	1	-1	-1	1	4	5
25	-10	2	1	29	0	0	5	2	5	2	5	2	-1	-4	1	1	3	5
26	-100	2	0	25	0	0	0	2	0	2	0	2	-4	-8	0	1	4	5
27	-1000	2	0	23	0	0	0	2	0	2	0	2	-4	-7	-1	1	7	9
28	-90	2	0	28	0	0	0	2	0	2	0	2	-3	-7	1	1	4	8
29	90	2	0	35	0	0	0	3	0	3	0	3	-1	-4	1	1	2	1
30	1000	2	1	41	0	0	2	3	2	3	2	3	-4	-5	3	1	3	4
31	100	3	1	41	0	0	0	3	0	3	0	3	-2	-3	-2	1	5	2
32	10	2	0	37	0	0	0	3	0	3	0	3	-3	-5	-1	1	6	1
33	0	2	0	35	0	0	0	3	0	3	0	3	-4	-6	-1	1	5	1
34	0	2	1	34	0	0	0	3	0	3	0	3	-3	-5	-1	1	5	3
35	0	2,	1	35	0	0	0	3	0	3	0	3,	-5	-7	1	1	5	2
36	0	1	0	35	0	0	0	3	0	3	0	3	-5	-7	1	1	6	3
37	0	2	0	33	0	0	0	3	0	3	0	3	-5	-8	2	1	5	2
38	0	1	0	31	0	0	0	3	0	3	0	3	-5	-8	2	1	4	2
39	0	2	0	28	0	0	0	3	0	3	0	3	-5	-9	2	1	4	2
40	0	1	0	23	0	0	0	3	0	3	0	3	-5	-102		1	4	2
41	0	1	0	17	0	0	0	3	0	3	0	3	-6	-130		1	7	5
42	0	2	1	9	0	0	3	3	3	3	3	3	-6	-111		1	5	0
43	0	1	1	3	0	0	0	42	0	42	0	42	-5	-9	-1	1	4	4
44	0	1	0	1	0	0	0	27	0	27	0	27	-6	-100		1	1	1
45	0	1	0	1	0	0	0	27	0	27	0	27	-6	-9	0	1	0	0
46	0	1	0	1	0	0	0	27	0	27	0	27	-6	-9	0	1	1	1
47	-10	1	0	1	0	0	0	27	0	27	0	27	-6	-9	0	1	0	0
48	-100	1	0	1	0	0	0	27	0	27	0	27	-5	-8	0	1	0	0
49	-1000	1	0	0	0	0	0	27	0	27	0	27	-5	-8	0	1	0	0
50	-100	1	0	2	0	0	0	27	0	27	0	27	-4	-6	0	1	0	0
51	-10	1	2	4	0	0	0	10	0	10	0	10	-2	-4	0	1	0	0
52	0	1	2	4	0	0	2	10	2	10	2	10	3	4	0	1	0	0
53	0	1	0		0	0	0	27	0	27	0	27	3	4	-1	1	2	1
54	0	4	0		0	0	0	1	2	2							54	0

n	ccr	rc	chr	rb
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	11215	
10	0	0	11215	
11	0	0	11215	
12	0	0	11215	
13	0	0	0	0
14	0	0	0	0
15	1033	0	0	
16	1033	0	0	
17	1033	0	0	
18	1033	0	0	
19	0	0	0	0
20	0	0	0	0
21	0	0	0	0
22	0	0	0	0
23	0	0	0	0
24	0	0	0	0
25	0	0	0	0
26	11922	0	0	
27	11922	0	0	
28	11922	0	0	
29	0	0	0	0
30	0	0	0	0
31	0	0	0	0
32	0	0	11922	
33	0	0	11922	
34	0	0	0	0
35	0	0	0	0
36	11922	0	0	
37	11922	0	0	
38	11922	0	0	
39	11922	0	0	
40	11922	0	0	
41	0	0	0	0
42	0	0	0	0
43	0	0	0	0
44	0	0	0	0
45	0	0	0	0
46	0	0	0	0
47	0	0	0	0
48	0	0	0	0
49	0	0	0	0
50	0	0	0	0
51	0	0	0	0
52	11720	0	0	
53	0	0	0	0
54	0	0	0	0	0	0	2	0	0	0	0	1	2	3

n															
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	6	2	0	0	0	1	0	0	0	1	0	1	0	1	1
3	7	2	0	0	4	0	0	0	0	0	0	0	1	1	0
4	7	4	1	2	4	0	0	0	0	0	0	0	1	0	2
5	7	4	1	3	4	0	0	0	1	0	1	2	1	1	
6	7	3	1	3	5	0	0	0	0	0	0	0	0	1	
7	7	A	2	5	5	1	0	0	1	0	0	2	0	0	
8	7	6	5	6	6	6	4	2	5	2	3	3	2	2	
9	12	11	10	10	11	8	4	5	8	6	4	7	5	5	
10	12	12	10	9	12	9	5	6	9	8	6	8	6	7	
11	12	12	11	10	13	8	6	6	9	8	7	8	6	6	
12	11	12	12	10	12	10	7	8	9	7	8	6	6		
13	11	12	12	10	11	11	7	7	9	10	7	9	6	6	
14	11	12	12	10	11	10	7	6	9	8	7	9	6	6	
15	10	13	12	11	10	8	6	9	10	7	8	6	6		
16	10	13	12	11	11	10	8	7	8	10	7	8	7	7	
17	10	13	12	11	11	9	8	7	8	8	7	8	6	6	
18	10	13	11	10	10	9	7	7	9	7	7	10	6	7	
19	10	13	11	9	10	10	8	6	10	8	8	10	6	7	
20	10	13	9	9	11	8	6	9	9	8	10	6	5		
21	10	11	8	7	7	11	8	7	10	7	9	7	5	5	
22	11	9	7	6	6	11	8	8	9	7	7	7	5	5	
23	11	9	7	6	4	8	8	8	8	7	7	7	4	6	
24	10	8	6	5	4	6	8	8	7	6	7	8	5	6	
25	10	7	5	4	4	6	7	4	5	9	7	8	4	6	
26	10	6	3	3	2	4	6	3	3	8	8	7	3	7	
27	9	6	3	2	0	4	7	4	6	8	7	6	2	4	
28	8	4	4	1	1	5	5	5	6	8	6	5	3	4	
29	8	5	3	4	3	5	3	5	5	7	8	6	5	5	
30	9	8	4	1	4	4	5	7	7	9	9	8	7	8	
31	10	9	5	3	3	3	5	8	6	6	10	9	8		
32	10	8	4	3	3	3	4	9	6	7	10	10	8	7	
33	10	8	4	2	2	3	2	7	6	6	8	10	8	7	
34	9	7	4	2	1	4	2	6	5	5	8	9	8	8	
35	9	7	3	1	2	3	2	7	5	5	8	10	9	8	
36	9	7	3	1	2	3	3	8	5	5	8	10	9	8	
37	9	6	3	1	3	3	3	8	6	5	9	9	9	8	
38	9	6	3	1	3	3	3	7	5	4	8	9	9	8	
39	9	5	3	1	3	3	3	7	3	4	7	9	9	8	
40	9	4	2	2	3	3	4	7	4	4	6	8	8	7	
41	9	2	2	1	1	2	5	8	3	3	6	8	7	7	
42	8	3	0	2	2	1	3	6	2	2	6	6	5	7	
43	7	3	1	1	0	1	4	3	2	1	5	4	3	4	
44	6	2	0	0	0	0	1	0	0	1	1	1	0	1	
45	6	3	0	0	0	0	0	0	0	0	1	1	0	0	
46	6	3	0	0	0	0	0	1	0	0	0	1	1	1	2
47	6	3	0	0	0	0	0	0	0	0	0	1	0	0	
48	5	2	0	0	0	0	0	0	0	0	1	1	1	1	
49	5	2	0	0	0	0	0	0	0	0	0	0	1	1	
50	4	2	0	0	0	0	0	0	0	0	0	1	0	1	
51	2	0	0	0	0	0	0	0	0	0	0	1	0	0	
52	3	4	3	3	3	3	3	3	3	3	3	4	3	3	
53	2	3	3	2	2	2	3	1	2	2	2	2	1	4	

Figure 2
Illustration : segmentation et reconnaissance en macroclasse pour le mot "bolide".

tz	Nombre de passages par zéro
i	Instabilité des indices
53	EN53
12	MAX12
Me	MAXEN
mMe53	Extrema sur EN53
fx,cx	Respectivement frontière et classe successivement établies
o1,o4	Indices d'ouverture
a1,a2,a3,a5	Indices d'acuité.

BASES DE DONNEES
BASES DE CONNAISSANCES

Président

Raymond DESCOUT

C.N.E.T. de Lannion

UN SYSTEME A BASE DE CONNAISSANCES POUR LA RECONNAISSANCE AUTOMATIQUE DES TONS DU CHINOIS

Yifan GONG Jean-Paul HATON

Equipe Reconnaissance des Formes et Intelligence Artificielle
Centre de Recherche en Informatique de Nancy
Université de Nancy 1
Campus Scientifique BP 239
54506 Vandoeuvre-les-Nancy

ABSTRACT - Tones (fundamental frequency variation) classify Chinese words into five categories and are indispensable for speech understanding. Context dependancy of tones in real speech makes statistical methods insufficient for tone recognition. We propose a knowledge-based system that interprets partially recognized data using production rules. A morphological analysis on the intensity contour of speech is used for segmentation. From the pitch contours a multi-reference classifier creates primitive patterns with certainty factors. An inference engine capable of reasoning on the uncertain data then decides the final tone type of each segment using the context information.

1. INTRODUCTION

Le Problème de la Reconnaissance de Tons

Dans la langue chinoise, les tons portent une quantité très importante d'information et il est indispensable de les reconnaître pour pouvoir comprendre un discours en Chinois. Les tons correspondent à des variations de la fréquence fondamentale de la parole, et linguistiquement, ils sont classés en cinq catégories: T1 (haut et plat), T2 (montant), T3 (descendant puis montant), T4 (descendant) et T0 (prolongement du ton précédent). Il existe différentes réalisations d'un ton linguistique suivant les tons qui le précèdent et qui le suivent. L'exemple le plus connu est qu'une suite de T3 se transforme souvent en une suite de T2 à l'exception du dernier T3: {T3} → {T2}T3. La dépendance contextuelle est due en grande partie à la tendance de faciliter la prononciation. La difficulté de la reconnaissance des tons réside principalement dans l'influence du contexte.

Motivations

Des études ont été menées sur le comportement des tons et des travaux fondés sur des méthodes purement statistiques ont permis de déterminer des paramètres relativement discriminants et ont donné des résultats intéressants [1]. Cependant, comme dans d'autres domaines d'application, l'utilisation de modèles statistiques est inadéquate pour la reconnaissance de structures complexes [2]. Les méthodes purement statistiques sont en particulier insuffisantes en reconnaissance de la parole pour modéliser la dépendance contextuelle, les défauts de détection de fréquence fondamentale et la variation inter- et intra-locuteur. Elles sont ainsi totalement incapables de distinguer différents tons correspondant à des variations de courbes semblables. Le raffinement des techniques statistiques (recherche des invariants, définition de différentes distances, ...) n'améliorent pas sensiblement les résultats.

Pour arriver à un résultat de reconnaissance des tons utilisable dans un système de reconnaissance de parole, il est nécessaire d'exploiter des techniques plus fines, notamment l'introduction de connaissances contextuelles. Ces techniques consistent à donner une interprétation structurelle des formes primitives obtenues par une classification rudimentaire du signal. Dans la résolution des autres problèmes de reconnaissance de formes similaires au nôtre, des méthodes syntaxiques (analyse grammaticale) ou l'utilisation de systèmes experts ont permis d'améliorer considérablement les résultats.

Dans l'état actuel de la recherche sur les tons une certaine expertise a été acquise mais la connaissance reste morcelée et incomplète. On ne connaît pas, par exemple, le nombre de formes nécessaires pour représenter les variations de tons, ni les relations phonologiques entre eux, la dépendance contextuelle des formes, etc. Il est donc nécessaire d'utiliser un système souple, facile à modifier et permettant d'améliorer progressivement nos connaissances. Nous considérons donc qu'une approche d'intelligence artificielle par système à base de connaissances sous forme de règles de production est bien adaptée à une telle tâche.

Des approches analogues se sont révélées fruc-

tueuses pour améliorer des techniques algorithmiques de détection de pitch [6], mais aussi de segmentation d'images [7]. Plus généralement le travail que nous présentons ici s'inscrit dans la ligne des recherches menées par notre équipe dans l'utilisation des techniques à base de connaissances en reconnaissance automatique de la parole [8][9].

Présentation Générale de la Méthode

A partir du signal vocal continu fourni en entrée notre système fournit une suite de symboles correspondant à la variation linguistique de tons de la phrase, assortis de coefficients de vraisemblance. Il s'agit donc de reconnaître à laquelle des cinq classes disjointes précédentes (T1, T2, T3, T4, T0) appartient un segment de parole.

Notre système comporte un niveau d'interprétation (traitement symbolique) et un niveau d'identification de formes primitives qui fournit les symboles à partir du signal physique. Les grandes étapes du traitement sont: (1) Calcul et lissage des contours de pitch et d'intensité à partir du signal de parole; (2) Segmentation automatique du signal de parole en segments linguistiques par une analyse morphologique; (3) Modélisation des segments du contour de pitch par des coefficients de N-ième régression; (4) Classification des segments en formes primitives par une méthode statistique; (5) Interprétation de ces formes primitives pour reconnaître les tons à l'aide d'un moteur d'inférence raisonnant sur des faits incertains et disposant d'une base de connaissances sous forme de règles de production. Cette base contient l'information contextuelle (et l'information sur d'autres relations), les données fournies par les étapes précédentes constituant la base de faits initiale.

2. PRE-TRAITEMENTS

Estimation du Contours de Pitch

Le contour de pitch (valeur de fréquence fondamentale pour la parole voisée ou indication de non-voisement ou silence dans le cas contraire) est calculé par l'éditeur de signal "ASSIA" [5] par un algorithme de détection de pitch spécial développé dans notre équipe [3]. Bien que ce détecteur fournisse des résultats assez fiables, nous avons inclus une phase de traitement afin de corriger des points dus à l'irrégularité de prononciation et à d'autres causes. Ce traitement consiste d'abord à calculer l'histogramme du contour (fréquences de chaque valeur dans le contour), et ensuite à remplacer les points dont la valeur est à l'extérieur de l'histogramme, filtré par un filtre passe-bas, par une valeur obtenue en moyennant certains points précédents. Un opérateur médian est enfin appliqué sur le contour ainsi traité pour éliminer le bruit.

Estimation du Contour d'Intensité

Dans notre étude, la segmentation de contour de pitch est fondée sur le contour de l'intensité du signal de parole. L'intensité est la valeur maximum d'une période de signal de parole voisé estimée statistiquement. Naturellement, elle est différente de l'énergie moyenne d'un segment du signal. Notamment, le contour de l'intensité montre une nette variation pour le changement de voyelles pour lesquelles l'énergie est souvent constante. Ce paramètre n'a pas été utilisé dans le traitement de parole parce que sa détection est difficile. L'algorithme de détection de pitch que nous utilisons permet aussi l'estimation du contour d'intensité. Cet algorithme, utilisant un modèle temporel, est non seulement capable de fournir la valeur de période de l'excitation mais aussi d'indiquer la valeur et la position du maximum (ou minimum) dans chaque periode du signal. Le contour d'intensité est ensuite soumis à un filtre passe-bas. Pour ne pas introduire de déphasages entre les différents composantes spectrales et obtenir ainsi une bonne précision lors de la localisation des segments, un filtre à réponse finie est utilisé.

3. SEGMENTATION AUTOMATIQUE

Dans la parole continue, chaque transition de phonèmes entraîne un changement du signal et provoque en général une variation du contour d'intensité. Notre méthode de segmentation consiste à rechercher sur ce contour tous les creux correspondant aux transitions. Nous avons utilisé le formalisme de la morphologie mathématique pour l'analyse de contours. La morphologie mathématique est une théorie ensembliste initialement appliquée à l'image [4] (une image est une fonction $f: \mathbf{R}\times\mathbf{R}\rightarrow\mathbf{R}$, telle que $\exists M, \forall x, 0\leq f(x)\leq M$). Elle permet de systématiser des idées et des méthodes de traitement d'image et fournit une méthodologie de traitement avec des outils appropriés. L'utilisation d'éléments structurants (objets géométriques pouvant être déplacés sur toute image) et l'analyse à l'aide d'opérations ensembliste; constituent les concepts de base des algorithmes de la morphologie mathématique. Deux opérations élémentaires sont intéréssantes pour détecter les creux dans le contour de l'intensité:

- La dilatation: $(f \oplus B)(x) = \mathrm{Sup}\ \{f(y);\ y \in B_x\}$ (1)

Cette opération donne l'ensemble des plus grands éléments dans f(y) lorsque y parcourt l'ensemble B centré en x;

- L'érosion: $(f \ominus B)(x) = \mathrm{Inf}\ \{f(y);\ y \in B_x\}$ (2)

Cette opération donne l'ensemble des plus petits éléments dans f(y). Dans (1) et (2) B_x représente l'élément structurant symétrique centré en x.

En considérant que le contour d'intensité I est une

image particulière, nous composons ces opérations pour obtenir un signal C contenant des pics normalisés correspondant aux creux du contour, la largeur des creux pouvant être sélectionée:

$$C = (I \ominus B) \oplus B - I \qquad (3)$$

Nouns examinons ensuite les pics fournis par l'analyse morphologique. Si la hauteur d'un pic (égale à la profondeur d'un creux dans le contour d'intensité) est suffisamment grande, le point correspondant est considéré comme une limite de segment candidat. Certains creux sont produits par des petites variations à l'intérieur d'un phonème, par exemple entre la frontière de la zone instable (où le contour de l'énergie est montant ou déscendant) et la zone stable d'un phonème, et donc ne correspondent pas à des limites de segment. Ces points sont éliminés par un critère fondé sur les rapports de la valeur du premier pic à gauche et celui à droite du creux à la valeur du creux du contour d'intensité.

4. MODELISATION DU CONTOUR DE PITCH

L'objectif de la modélisation du contour est d'extraire un ensemble restreint de paramètres qui caractérisent la variation de fréquence fondamentale avec une perte d'information aussi faible que possible. Ces paramètres sont utilisés par le niveau d'interprétation si nécessaire. Chaque segment est modélisé par sa N-ième régression par rapport au temps. A partir de ce modèle, il est facile de calculer en chaque point (avec N multiplications) la valeur du contour, ainsi que les dérivées première et seconde. Nous avons synthétisé le contour à l'aide des coefficients de régression et constaté que la précision est suffisante lorsque l'ordre du polynôme N est supérieur à 2. Cependant, si les valeurs de pitch des premiers points ou des derniers points d'un segment ne sont pas correctement estimées, le contour du segment synthétisé est déformé de façon notable.

5. CLASSIFICATION

Le but de cette étape est de classer les courbes de variation tonale en formes primitives. Pour modéliser tous les phénomènes de variation d'un ton, nous avons utilisé multi-références pour chaque ton.

Soit $\mathbf{C} = \{C_1, C_2, C_3, \ldots\}$ l'ensemble de classes de tons où

$$C_i = \{C_{i,1}, C_{i,2}, C_{i,3}, \ldots\}$$

est l'ensemble des sous-classes pour un ton i donné. Ces ensembles sont disjoints:

$$C_{i,j} \cap C_{k,l} = \Phi \quad \forall i,j,k,l$$

A chaque sous-classe $C_{i,j}$ est associée un vecteur $R_{i,j}$, correspondant à la forme de référence, dans l'espace de représentation. Soit F le vecteur de test. La classification consiste à prendre une décision

$$d(F) = C_{i,j}$$

telle que $||F - R_{i,j}||$ soit minimum.

Les variations de durée d'élocution entraînent des segments de longueur différente et nécessitent une normalisation temporelle avant la comparaison avec des formes de références. Les frontières de segments étant détectées avec une grande précision, nous avons utilisé un simple réglage temporel linéaire pour que la longueur de la forme de test et de la forme de référence soient égales.

6. INTERPRETATION

Introduction

Le phénomène de dépendance contextuelle des tons est très courant en Chinois. On observe ainsi des courbes correspondant au même ton et la même forme de courbe pour des tons différents. Citons par exemple: T4 précédé par T3 a tendance de devenir T2; T3 et T1 pour différents locuteurs peuvent avoir tendance à être identiques; T3 précédé par T1 a tendance à se présenter comme T4, etc. Nous avons utilisé le terme "tendance" pour dire que le phénomène est possible mais qu'il y a des exceptions. Un raisonnement approximatif est ainsi nécessaire. Pour la description des tons, nous n'avons pas un vocabulaire suffisant pour exprimer et formuler les connaissances (comme dans l'analyse de phonèmes, les termes "formant", "burst", etc). Ceci a soulevé des difficultés dans la construction de la base de connaissances déclaratives et dans la définition des fonctions fournissant des informations symboliques nécessaires pour le raisonnement.

Représentation des Connaissances

Dans notre système, le niveau d'interprétation est chargé de corriger les erreurs de reconnaissance de tons introduites par la classification statistique. Cet interprète est implanté sous forme d'un système expert. Pour chaque segment du signal de parole, les données fournies par les étapes précédentes sont:

(1) Les coefficients de régresion. Ces coefficients contiennent toute l'information (sous forme très compacte) sur la forme de courbe du contour de pitch et peuvent être utilisés éventuellement pour calculer des paramètres nécessaires lors de l'interprétation;

(2) Les noms des tons primitifs reconnus par la classification statistique avec les scores de reconnaissance;

(3) L'emplacement et la longueur du segment;

(4) La moyenne de fréquence sur une phrase.

Ces informations sont représentées sous forme d'un objet et la donnée de l'interprète est donc une liste de ces objets. Les connaissances sont

représentées sous forme de règles de production. Nous distinguons deux types de règles: les règles locales qui décrivent un segment et les règles contextuelles qui portent sur les segments situés à gauche et à droite du segment courant. La syntaxe d'une règle est la suivante:

```
REGLE ::= si (CONDI) alors (CONSEQUENCE) cv (COEFF)
CONDI ::= <(FAIT) | ((PREDICAT))> CONDI |
CONSEQUENCE ::= concl (CONCLUSION) act (ACTION)
CONCLUSION ::= (FAIT) CONCLUSION |
ACTION ::= S-expression en LISP ACTION |
FAIT ::= chaîne de caractères
PREDICAT ::= S-expression en LISP
COEFF ::= nombre réel compris entre (0,1]
```

La condition d'application d'une règle peut se composer de faits ou de prédicats assortis d'un coefficient de vraisemblance. A chaque règle est également associé un facteur de certitude. Quand tous les faits dans la partie condition sont prouvés, les faits dans la partie conclusion sont établis par le moteur d'inférence. Les ACTIONs dans la partie CONSEQUENCE d'une règle consistent en un ensemble d'actions que le moteur d'inférence doit effectuer quand la règle est déclenchée. Elles consistent à modifier la base de faits, à prendre une décision sur le résultat d'interprétation ou à calculer des faits non déductibles directement.

Mécanisme de Raisonnement

Nous avons réalisé un moteur d'inférence qui invoque des règles en chaînage-arrière et produit des sous-problèmes en profondeur d'abord avec possibilité de raisonnement approximatif. Son rôle essentiel est de prouver toutes les hypothèses sur le type de ton et de prendre comme décision celle qui est la plus vraisemblable. Le coefficient de vraisemblance CV d'un fait déduit d'une règle dont le facteur de certitude est CFR est déterminé par:

$$CV = CFR \times \min \{CV_i\}$$

où CV_i est le CV de la i-ième prémesse dans la partie condition de la règle. Si plusieurs règles conduisent à la même conclusion, la vraisemblance est renforcée:

$$CV = \max \{CV_i\}$$

où CV_i est le CV obtenu en appliquant la i-ième règle qui conduit au fait. Seuls les faits dont le CV est suffisamment élevé sont considérés comme établis.
Pour faciliter la mise au point du système, nous avons implanté aussi des mécanismes de base permettant de suivre la trace d'exécution: possibilité d'explication (pourquoi, comment, règles utilisées, etc).

7. RESULTATS PRELIMINAIRES

Le premier corpus sur lequel nous avons testé notre méthode contient 156 chiffres (de 0 à 10). Certains chiffres sont prononcés différemment (le même phénomène qu'en Anglais où "zero" se prononce "zero" ou "o") et le corpus contient 15 sons. Notre méthode de segmentation a localisé correctement tous les segments du corpus. Le résultat de la classification de formes primitives est présenté dans la table suivante (où T_i et TR_i sont des données et des résultats respectivement):

	TR1	TR2	TR3	TR4	TOTAL
T1	50(83%)	1	0	9	60
T2	0	7(100%)	0	0	7
T3	0	0	42(98%)	1	43
T4	0	0	0	46(100%)	46
TOTAL					156(93%)

Les erreurs de classification sont dues aux problèmes suivants: - La longueur des formes à identifier est normalisée avant la comparaison avec les références et n'est pas utilisée en classification. - L'effet de contexte dont nous avons cité quelques exemples au par. 6.1; - Les erreurs d'estimation de pitch. Toutes ces erreurs ont été corrigées par le système expert d'interprétation.

8. CONCLUSIONS

L'approche règles de production est puissante et souple; elle nous a permis de modifier, d'augmenter et de compléter facilement le système. L'utilisation d'une base de connaissances nous a permis aussi d'introduire l'interaction entre le traitement du signal et le traitement des symboles, les deux étant traditionellement séparés. La méthode est efficace pour inclure des connaissances contextuelles et perceptuelles et des connaissances pour la correction des erreurs de détection de formes primitives. Ces connaissances sont difficiles à modéliser mathématiquement. Les premiers résultats obtenus montrent que la méthode est prometteuse et on obtient facilement (malgré la manque de connaissances phonétiques et phonologiques) d'assez bons résultats (sur un tout petit corpus) qui peuvent être utilisés comme l'entrée d'un système de reconnaissance du langage parlé. Nous pensons que l'évaluation doit continuer pour compléter la base de règles et pour obtenir une statistique fiable sur la performance.

REFERENCES

[1] P. Hallé, "Les Tons du Chinois de Pekin, Leur Comportement en Parole Continue", Actes des 14-ièmes JEP, GALF, PARIS, Juin 1985.
[2] N. Nandhakumar, J.K. Aggarwal, "The Artificial Intelligence Approach to Pattern Recognition - A Perspective and an Overview", Pattern Recognition, Vol.18, No.16. 383-389, 1985.
[3] Y. Gong, J.P. Haton, "Time Domain Harmonic Matching Pitch Estimation Using Time-Dependent Speech Modeling", soumis à publication.
[4] J. Serra, "Image Analysis and Mathematical Morphology", Academic Press, 1982.
[5] Y. Gong, J.P. Haton, "Manipulation et Analyse de Signaux sous UNIX: l'Editeur ASSIA", Actes du congrès Sm90, Versailles, PARIS, 1985.
[6] W. Dove et al., "Knowledge-based Pitch Detection", Proc. IEEE ICASSP, Boston, 1983.
[7] A.M. Nazif and M.D. Levine, "Low Level Image Segmentation: An Expert System", IEEE Tr. PAMI, 6, No.5, 555-577, 1984.
[8] J.P. Haton and J.P. Damestoy, "A Frame Language for the Control of Phonetic Decoding in Continuous Speech Recognition", Proc. IEEE ICASSP, Tampa, 1985.
[9] N. Carbonell et al., "APHODEX, Design and Implementation of an Acoustic-Phonetic Decoding Expert System", Proc. IEEE ICASSP, Tokyo, 1986.

ANALYSE LINGUISTIQUE DE CORPUS D'ORAL FINALISE

F.Neel*, D.Luzzati*, M.A.Morel, D.Delomier, C.Leroy, E.Thévenon

LIMSI-CNRS* PARIS III

ABSTRACT

An experimentation was prepared by the GRECO CP dialogue group and realized with the CNET scientific and financial support. The aim was to record and analyse real task-oriented dialogues in information centers. Three phases was recorded: real dialogue, operator's voice modified by a vocoder, operator's comprehension limited whenever the user's production did not respect syntaxic constraints. The linguistic analysis deals with lexical, syntaxic and interactive aspects and shows to what extent the user, thinking he is in presence of a machine, inconsciously adopts a particular language behaviour.

1 - RAPPEL DE L'EXPERIMENTATION

Il s'agissait d'enregistrer et d'analyser des dialogues naturels, produits d'une part face à un interlocuteur humain, et de l'autre en situation de communication homme/ machine (4). Deux corpus ont été réalisés (l'analyse du second, en cours, ne sera pas abordée ici):
- Renseignement SNCF: les enregistrements, effectués par téléphone, concernent les horaires et, dans une moindre mesure, les tarifs.
- Centre d'Information et d'Orientation de l'UER de psychologie du travail de l'Université de Paris V: les enregistrements portent sur l'orientation d'étudiants de licence et de maîtrise.

Chacune de ces expérimentations comportait trois phases:
- Phase I: phase de référence, avec voix naturelle.
- Phase II: l'opératrice, dont la voix était mécanisée, observait quelques contraintes de production simples.
- Phase III: à celles-ci s'ajoutent des contraintes d'incompréhension lorsque l'utilisateur avait recours à des structures syntaxiques trop complexes ou lorsque ses requêtes n'étaient pas explicitement libellées sous forme d'interrogations directes ou indirectes.

Dans les phases II et III, les utilisateurs étaient prévenus qu'ils allaient s'adresser à une machine, et l'utilisation du processeur de signal SCH fixant le fondamental à une valeur constante rendait crédible cette impression.

L'analyse linguistique a porté sur plusieurs aspects (7), notamment:
- Lexique (mots de l'oral, mots propres à la tâche).
- Syntaxe.
- Phénomènes interlocutionnels.

L'objectif était multiple:
- Extraire les éléments caractéristiques de la tâche et du type de dialogue induit par la réalisation de cette tâche, ceci afin de permettre l'élaboration d'un modèle informatique relativement simple permettant de simuler une ce type de dialogue finalisé.
- Retrouver, notamment dans la phase I, les marques de l'oral spontané (étudiés dans les conversations et débats radiophoniques ou télévisés)(6) (1) (8) et mettre en évidence l'influence de la tâche.
- Mesurer l'influence du mode de production de la machine simulée en phases II et III, notamment sur la syntaxe adoptée par l'utilisateur et sur son comportement langagier en général.
- Observer la capacité d'adaptation de l'utilisateur en cas de marques d'incompréhension de la part de la machine.

Ces deux derniers points étaient difficiles à mettre en évidence, du fait de la brièveté des communications en phases II et III (de l'ordre de 70 à 90 sec)(9).

2 - MODELE DE LA TACHE

On constate que, pour la tâche SNCF, les utilisateurs savent avec exactitude quels renseignements ils désirent obtenir et connaissent, en général, les éléments nécessaires à l'obtention de ce renseignement. Cela différencie fondamentalement cette tâche de celle qui s'est déroulée dans un CIO, où les étudiants n'avaient pas une idée très claire de ce qu'ils recherchaient. Dans le corpus recueilli apparaissent deux sous-tâches principales: les demandes d'horaire et, plus accessoirement, les demandes de tarificcation. Les analyses montrent qu'on peut aisément extraire une structure-type des requêtes, comportant un nombre fini d'éléments: catégorie du train (CA), gare de départ (L1), gare d'arrivée (L2), jour (T1) et plage horaire (T2). Ces éléments, qui ne sont pas fournis systématiquement dans le même ordre, sont en général précédés d'une locution verbale introductrice exprimant une requête: "*j'aimerais/ je voudrais + savoir/ connaître...*". Pour les demandes de tarification, qui viennent souvent en complément des demandes d'horaire, il faut ajouter le type de billet (aller/ retour, premières/ secondes) et, en général, des conditions de réduction.

Le modèle de la tâche peut ainsi être représenté par un tableau comportant ces différents éléments, auxquels seront associées des valeurs préétablies au dialogue ou statiques: *optionnel, valeur par défaut, obligatoire*, et des valeurs dynamiques *satisfait/ non satisfait*... Un modèle de ce type s'apparente à des modèles développés précedemment pour des places d'avion par exemple (3).

On ne saurait cependant se contenter d'une simple analyse par mots-clés, et une syntaxe locale est à prendre en compte, faisant intervenir les syntagmes prépositionnels qui précèdent certains éléments ("*au départ de*" + L1 par exemple). Dans le cours du dialogue, les requêtes sont rarement complètes. Une série d'échanges mettant en jeu des formules anaphoriques est alors nécessaire, ce qui suppose un historique du dialogue pour en garder une trace et les traiter (10).

3 - CARACTERISTIQUE DE L'ORAL

Les enregistrements de la phase I ont surtout servi de référence. Ils ont été peu étudiés pour eux-mêmes.

On y observe cependant diverses propriétés des discours oraux. L'interaction notamment est très importante, avec les multiples interruptions et les chevauchements de propos qu'elle suppose. Ceci entraîne des échanges brefs avec des phrases nominales ou tronquées. Hésitations et auto-corrections sont nombreuses, ainsi que les ajouts, sous forme de formules métadiscursives appréciant le renseignement fourni ou justifiant la requête: *"si tôt c'est pas possible", "vous comprenez on m'attend à l'hôtel"...*

La redondance est permanente car les propos informatifs antérieurs sont presque constamment repris, par l'opératrice comme par l'utilisateur:

O13 BON *là vous avez deux trains à la même heure,* HEIN
C13 oui
O14 e:: ça vous va comme horaire ou::
C14 BEN *17H 10 e: je pense pour 21H 40 parce que, e: l'hotel est à l'extérieur de: de la chapelle* DONC *e,, e:: oui` ou alors je sais pas qu'est qu'est-c'qu'y a qu'est-ce qu'il y a après 1'17H 10*
(1.60.9)

09 arrivée 17H 59, j'en ai pas entre
C10 BON BIEN ECOUTEZ *j'vais leur noter ça, sinon après e: c'est 18H 25* HEIN *(...)*
O14 voilà, c'est ça il ne circule, c'est bien celui d'18H 28 ah mais attendez, attendez parce que y en a deux,, OKAY *non c'est bon vous en avez un deuxième qui circule lui tous les jours sauf les dimanches,,* DONC *c'est bien ça 18H 28(...)*
C17 d'accord oui c'est c'qu'on m'avait dit
O17 HEIN
C18 BON BEN ECOUTEZ *j'pense que ça dvrait aller, merci beaucoup madame*
(1.56.9/18)

On peut supposer que ces reprises assument plusieurs fonctions dont il faudrait tenir compte dans la réalisation de dialogues homme-machine: cela permet de maintenir le contact (rôle phatique), de mémoriser un renseignement ou de formuler une confirmation implicite (rôle métadiscursif), afin de provoquer une rectification en cas de mauvaise audition. Comme le montrent des expériences menées à l'IPO d'Eindöven sur la comparaison de

deux modes (clavier/ téléphone), ce procédé semble en partie dû au canal utilisé, en l'occurrence le téléphone, et il permet d'assurer une plus grande sécurité des échanges.

On peut enfin constater un comportement particulièrement coopératif de l'opératrice, qui propose spontanément divers horaires, anticipant ainsi sur les requêtes du correspondant, ou qui s'enquiert d'elle-même si le train proposé n'est pas "*trop tôt*" ou "*trop tard*".

4 - COMPORTEMENT LANGAGIER INDUIT PAR LA MACHINE (PHASES II ET III)

Ces deux phases ne présentent pas de différence fondamentale. Les tendances qu'elles dénotent sont simplement parfois plus marquées en phase III qu'en phase II.

Les propos sont dans l'ensemble plus complets et également plus concis avec voix mécanisée qu'avec voix naturelle. Dans ces deux phases, on peut notamment comparer la requête initiale, formulée face à l'opératrice, à sa reformulation face à la machine présumée(5). Sur l'ensemble des communications, les locuteurs s'efforcent de conférer un maximum d'efficacité au dialogue, en fournissant d'eux-mêmes une information aussi exhaustive que possible. Cette complétude sémantique ne va pas toujours de pair avec une complétude syntaxique, et on peut trouver une simple juxtaposition de mots-clés:" horaires Paris Nonencourt...".

Simultanément, lexique et syntaxe s'appauvrissent. Les appuis du discours (*"bon", "ben", "alors"*...) subsistent à peine. Les marques formelles de politesse comme *"s'il vous plaît"* diminuent de moitié. Les variantes du lexique propre à la tâche sont moins nombreuses... Cela est dû au mode d'interaction induit par la présence de la machine: les différentes marques de personnalisation des énoncés (jugements ou témoignages personnels) n'ont plus cours, ce qui entraîne la disparition de conjonctions de subordination (*parce que, si*). Les locuteurs n'interrompent pas la machine et attendent qu'elle leur cède explicitement la parole pour intervenir après: *"ce renseignement vous convient-il"*. La situation de communication homme/ machine les conduit en somme à délaisser une forme d'interactivité dans le dialogue, en procédant comme s'ils se trouvaient face à une console, c'est-à-dire à attendre que la machine présumée leur *cède la main* avant de parler.

Le recours fréquent à l'ellipse et à l'anaphore produit des énoncés beaucoup plus stéréotypées qu'en phase I. Le locuteur semble rechercher une plus grande économie de propos, parallèlement à une meilleure efficacité.

Les déîctiques se référant au présent ou au *moi* des locuteurs (*"à partir de maintenant", "ce soir"*) s'estompent ou sont remplacés par des expressions dépersonnalisées, dans lesquelles le repérage spatio-temporel est absolu: *"demain après 6H"* a par exemple tendance à devenir *"vendredi 26 décembre 1985 à partir de 18H"*. Les formules présentatives de l'opératrice (phase I) *"j'ai un train qui..."* n'apparaissent plus dans le langage de la machine (phases II et III).

En phase III, dans les cas de demande de reformulation, on constate que la structure d'un tiers des énoncés demeure identique, les modifications ne concernant que la prosodie (articulation plus marquée, débit plus lent). Dans plus de la moitié des reformulations cependant (66%), on peut relever un apport d'information, sans que soit adoptée la formulation souhaitée, alors qu'on voulait inciter les locuteurs à recourir à des marques d'interrogation explicites.

5 - CONCLUSION

Un mimétisme certain peut être observé, mais il porte sans doute plus sur la prosodie de l'utilisateur que sur les structures syntaxiques adoptées, ce qui conforte les conclusions des expériences menées par Chapanis (2).

La comparaison des différentes phases montre que les jeux d'influence ont tendance, en fait, à s'exercer en sens contraire. En phase I, lorsque l'opératrice utilise des énoncés brefs, à caractère anaphorique ou elliptique, l'utilisateur a parfois tendance à insérer dans ses réponses de longs commentaires parfaitement construits. En phases II et III, alors que la syntaxe des énoncés de la machine est toujours rigoureuse, les témoins ont souvent recours à des formulations abruptes, avec ellipses et anaphores, sans utiliser les procédés du dialogue: reprises de propos, formules d'acquiescement...

Les dialogues obtenus sont ainsi peu naturels. Ils sont en tout cas ressentis comme pesants. Même si les personnes interrogées affirment avoir confiance en la machine, nombreux sont ceux qui, souvent de façon insidieuse, sollicitent une confirmation du renseignement fourni. Cela tendrait à prouver que les reprises de propos sont indispensables à plus d'un titre: ils allongent certains échanges tout en allégeant l'effet du dialogue; ils rendent ce dernier plus fiable, dans la mesure où l'information n'hésite pas à être redondante.

Il faut enfin noter l'importance des constructions utilisées par la machine. On a ainsi remarqué que la formule "*énoncez votre demande s'il vous plaît*" était assez mal ressentie, et avait un effet beaucoup plus déstabilisateur sur le langage des témoins qu'un énoncé interrogatif comme: "*quels renseignements désirez-vous obtenir*". Les énoncés injonctifs sont en effet peu courants dans le dialogue naturel où abondent formules de politesse et actes de parole indirects.

BIBLIOGRAPHIE

(1) **C. Blanche-Benveniste** et al.: *Des grilles pour le français parlé,* GARS No 2, février 1979.
(2) **A. Chapanis**: *Interactive communication: a few research answers for a technological explosion,* Nouvelles tendances de la communication homme-machine, Cours CCE J.S. Lienard éd. INRIA, 17-27 avril 1979.
(3) **S.E. Levinson** and **K.L. Shipley** : *A conversational mode arline information and reservation system using speech input and output,* The Bell system technical journal, Vol 59 No 1, january 1980.
(4) **D. Luzzati**: *ORSO projet pour la constitution et l'étude de dialogues homme/ machine No 1,* note scientifique LIMSI, septembre 1984.
(5) **D. Luzzati, M.A. Morel**: *ORSO No 3,* note scientifique LIMSI, mars 1986.
(6) **M.A. Morel** et al.: *L'oral du débat,* langue française No 65, février 1985.
(7) **M.A. Morel** et al.: *Analyse linguistique d'un corpus d'oral finalisé,* GRECO CP, septembre 1985.
(8) **E. Roulet**: *Echanges, interventions et actes de langage dans la structure de la conversation,* Etudes de linguistique appliquée No 44, oct-déc 1981, p. 7-39.
(9) **J.C. Sperandio, C. Letang-Figeac**: *simulation expérimentale de la synthèse vocale en dialogues oraux de communication homme-machine. Etude ergonomique,* GRECO CP, février 1986.
(10) **GRECO CP groupe dialogue homme-machine**: *Dialogue oral homme-machine orienté par l'action,* 5ème congrès Reconnaissance des formes et Intelligence Artificielle, AFCET-INRIA, Grenoble, 27-29 novembre 1985.

REPRESENTATION D'UN LEXIQUE A L'AIDE DE CONNAISSANCES DE PHONOLOGIE GENERATIVE EN R.A.P.C

Jacques Gispert

G.I.A., Faculté de Luminy
70 Route Léon Lachamp 13288 Marseille Cedex 09

ABSTRACT

The system we present processes Phonological and Lexical data constituing a coherent sub-set of a complete System for Automatic Recognition of Continuous Speech entirely realized in PROLOG II.

Every lexical unit include a root morpheme, the possible grouping with other morphemes, and syntactico-semantical features. The rules of Generative Phonology, declaratively stated, authorize the derivation of the phonetic form of words from their underlying form. Morpho-syntactical rules allows the deduction of informations from morpheme grouping. Access to words are separately encoded using phonetic and syntactic informations.

Encoding of Lexicon is automatic ; however, some ambiguities need to be resolved manually. A set of predicates authorize the extraction of informations represented in the Lexicon, by analysis or derivation.

INTRODUCTION

Ce travail traite des informations phonologiques et lexicales constituant un sous-ensemble des connaissances utilisées par un système complet de R.A.P.C. entièrement réalisé en PROLOG II.

Chaque unité lexicale comprend un morphème racine caractérisé par sa description phonologique, une structure représentant les combinaisons possibles avec d'autres morphèmes ainsi que des informations résumant certains caractères syntactico-sémantiques. Les règles de Phonologie Générative, énoncées déclarativement, permettent la dérivation des formes phonétiques des mots à partir de leur forme sous-jacente construite par combinaison d'éléments lexicaux. Des règles morpho-syntaxiques permettent de déduire les informations provenant du regroupement des morphèmes. Les accès aux mots sont codés séparément au moyen d'informations syntaxiques et phonétiques.

Le codage du lexique est automatique à partir de la forme phonétique ; seules certaines ambiguïtés doivent être levées manuellement. Un ensemble de prédicats permet d'extraire toutes les informations représentées dans le lexique, par analyse (obtention de la structure phonologique sous-jacente) ou par dérivation.

I. POSITION DU PROBLEME

Le système présenté fait suite aux travaux de Meloni [12,13]. Son but est la reconnaissance de mots à partir d'un treillis de phonèmes produit par le décodage acoustico-phonétique. Les prédicats décrivant l'analyse en morphèmes de la forme phonétique servent à la constitution du lexique, à l'analyse directe d'un mot à partir du clavier, et également à la reconnaissance. Dans ce dernier domaine, les stratégies n'ont pas été programmées car le module de décodage acoustico-phonétique est en cours de développement.

Le système est limité à la composante phonétique des mots codés dans le Lexique ; c'est la partie pertinente pour la reconnaissance. La forme graphique nécessaire à la sortie des résultats sera ajoutée ultérieurement sous forme d'un nouveau module.

Les difficultés à résoudre sont de deux ordres :
- linguistique : nécessité de trouver une description complète et cohérente de la langue qui soit naturelle pour les spécialistes, en utilisant leur langage ;
- informatique : le système doit autoriser une représentation quasi directe des connaissances à traiter, et tenir compte des contraintes de place et de temps habituelles sur machine.

II - CADRE THEORIQUE DU SYSTEME

Le codage du Lexique ne doit pas être limité *a priori* par des considérations d'encombrement de sa représentation. On élimine donc le codage explicite de la forme phonétique de chaque mot individuellement. On cherche à exploiter les propriétés linguistiques dérivées de la parenté entre mots, permettant de ne coder qu'une fois les parties communes.

Bescherelle [1] présente un catalogue descriptif des conjugaisons qui sépare les verbes en de nombreuses classes de particularités. J. Pinchon [15] opère quelques regroupements et aboutit à un ensemble de formes un peu plus restreint, mais de même principe

La Phonologie Générative [2,3] offre un cadre analytique dans lequel les parentés entre mots trouvent des explications par le biais de variations phonologiques. L'ensemble du Français est décrit par Schane [20]. Des théories ont été élaborées pour d'autres langues, et d'abord pour l'Anglais.

Une étude des variantes de la Phonologie Générative se trouve dans [6]. Une analyse des alternances vocaliques du système de Schane en relation avec la position de l'accent est faite par F. Dell [5]. Le même auteur [4] nous propose une étude du e muet. M. Plenat a publié des travaux sur les liquides et les liaisons en Français [16], sur le phénomène de nasalisation [18], et sur la conjugaison des verbes irréguliers [17].

Ces auteurs et d'autres proposent donc des théories élaborées sur des bases plus générales que celle de Schane. Ces nouveaux développements assurent qu'un système les utilisant pourra rester pertinent, à condition de suivre l'évolution de la théorie par mise à jour des règles et des stratégies. Ceci nous a fait choisir ce cadre pour notre codage du Lexique.

III - CONNAISSANCES UTILISEES

Le système expérimental actuel traite les connaissances phonologiques données par Schane basées sur les traits articulatoires suivants : *consonantique, vocalique, haut, bas, avant, rond, nasal, tendu.* On utilise également des connaissances lexicales :

- découpage des mots en morphèmes (radicaux, préfixes et suffixes, désinences) qui permettent le regroupement en familles ;
- définition des groupements de morphèmes possibles ou interdits ;

et des connaissances grammaticales :

- catégories et attributs syntactico-sémantiques.

IV - CODAGE

IV - 1 des données

Chaque phonème dérivé ou sous-jacent est décrit par une clause associant son identificateur à son vecteur de traits :

Exemple :

<aa,vocalique.bas.tendu> → ;
<Aa,vocalique.bas> → ;

Aa représente le phonème /a/ sous-jacent lâche, et *aa* le phonème /a/ tendu, sous-jacent ou dérivé.

Des prédicats permettent :

- d'atteindre un trait dans la liste, par exemple trait-cons(l,c) donne la valeur positive ou negative du trait consonantique de l,
- de caractériser des classes de phonèmes (voyelle-orale, consonne...),
- de modifier la liste de traits selon les besoins des règles (négation, affirmation, échange...).

IV - 2 des règles

Les règles sont codées à l'aide de clauses Prolog sous une forme très proche de celle proposée par les linguistes. La tête de clause présente la séquence de phonèmes initiale et celle qui en dérive. La queue de la clause détermine les contraintes contextuelles d'application de la règle et les liens existant entre les deux formes.

Certaines règles peuvent contenir une information supplémentaire (domaine spécifique d'application en particulier pour les verbes, exclusion par rapport à d'autres règles etc.).

Exemple :

```
nasalization(v.c.q,v'.q)  →
    voyelle-orale(v)
    consonne-nasale(c)
    nasaliser(v,v')
    non-vocalique(q) ;

pretonic-adjust(v.q,v'.q,pretonic-adjust)  →
    voyelle-pretonique(v)
    non-tendu(v)
    inverser(trait-rond(v),trait-bas(v')) ;
```

V - INTERPRETATION DES REGLES

Les règles de phonologie que nous utilisons sont partiellement ordonnées. Leur interprétation se fait selon un ordre total qui est une particularisation arbitraire de l'ordre théorique ; des méta-règles décrivent cet ordre de manière déclarative.

Chaque règle, à son tour, est prise en compte. Lorsque son environnement est satisfait, ceci correspond à la possibilité d'un phénomène phonologique. Son application est par conséquent obligatoire ; la stratégie de synthèse force son application sur la chaîne à transformer en toutes les positions où son contexte est vérifé.

Par contre, en analyse, on ne peut appliquer systématiquement toutes les règles dont la partie droite est contenue dans la chaîne à analyser. En effet :

plusieurs règles peuvent produire un même phonème à partir de formes sous-jacentes distinctes, il y a donc indéterminisme sur le choix de la règle ;

certains phonèmes sont à la fois sous-jacents et dérivés ; ils peuvent figurer tels quels dans la forme phonologique ou bien être obtenus par dérivation. L'indéterminisme porte ici sur l'application ou non de la règle.

Il existe donc plusieurs formes sous-jacentes différentes conduisant à la même forme phonétique. Les connaissances phonologiques ne peuvent résoudre cette ambiguïté, qui sera levée par la présence dans le lexique de l'une des formes proposées.

La commande d'analyse ou de synthèse se fait par le prédicat *derive* qui assure la correspondance entre les formes phonologique et phonétique.

Le système doit être capable de reconnaître une phrase, même prononcée en violation de certains phénomènes comme par exemple les liaisons [16]. Les règles qui décrivent ces phénomènes pourront être facultatives. On admettra ainsi des formes phonétiques fausses sur le plan théorique, mais d'emploi assez fréquent.

VI - CODAGE DU LEXIQUE

Les éléments lexicaux sont des morphèmes. Ils sont représentés par des clauses indiquant la nature du morphème (prefixe, radical, marques de genre et de nombre...) et des données dépendant de cette information.

- L'identificateur du morphème est construit avec les phonèmes qui le constituent,
- les suffixes portent une information indiquant la catégorie syntaxique qu'ils produisent,
- les radicaux décrivent l'ensemble des mots construits autour d'eux, grâce à un terme qui indique les préfixes et suffixes possibles. Ce terme est constitué de doublets <préfixe,suffixe>, des opérateurs *et* et *ou* et de la fonction *event* (éventuellement),
- tous les morphèmes peuvent comporter des données particulières sur l'emploi des règles (exceptions sur les formes soumises ou non à la règle pretonic-adjust, formes apprises ...).

Exemple :

ddOoll(radical,ou(<vide,ou(vide,et(Oorr,event(Oozz)))>,
<Aann,et(Oorr,iirr)>))) → ;
Oorr(suffixe,nom) → ;
Oozz(suffixe,adjectif) → ;
Aann(prefixe,appris) → ;
iirr(suffixe,<verbe(3),appris>) → ;

Le radical permet d'accéder à tous les morphèmes constituant les mots qui lui correspondent. Par contre, les préfixes et suffixes n'offrent pas cette possibilité, le nombre de radicaux associés à un même préfixe étant trop grand.

Pour la décomposition en morphèmes d'un mot, la forme sous-jacente étant produite par l'analyse phonologique, il est préférable de disposer d'accès indépendants aux autres constituants du mot pour éviter de rechercher le radical n'importe où. L'analyse se fait de gauche à droite par recherche de préfixes, puis du radical et enfin de suffixes. Cette recherche est bien entendu non déterministe.

Le phonème le plus à gauche dans un morphème donne un accès naturel pour une analyse de gauche à droite.

Exemple : *acces-morpheme(Oo,Oorr,Oo.rr.nil) → ;*
acces-morpheme(dd,ddOoll,dd.Oo.ll.nil) → ;
acces-morpheme(Aa,Aann,Aa.nn.nil) → ;
acces-morpheme(ii,iirr,ii.rr.nil) → ;
acces-morpheme(Oo,Oozz,Oo.zz.nil) → ;

VII - TRANSFORMATION DES REGLES

L'interprétation directe des règles de phonologie aboutit à une solution linguistiquement correcte du problème. Elle l'est malheureusement moins sur le plan informatique, car l'analyse d'un mot requiers plusieurs minutes de temps machine sur VAX750. La synthèse ne nécessite que quelques secondes, la différence étant liée à l'indéterminisme d'application 'à l'envers' de règles conçues pour la synthèse.

En se fixant une règle qui détermine les formes gauche et droite les plus générales qui la concernent, et en appliquant à ces formes toutes les règles possibles de façon non déterministe, jusqu'à obtenir d'un côté une forme phonologique et de l'autre une forme phonétique, on définit tous les usages qu'il est possible de faire de cette règle. A chaque solution, on fait correspondre une macro-règle qui représente l'enchaînement des règles qui l'ont produite.

Exemple de Macro-Règle :

<macro-regle11,<Oo.c-n.c.q,c.q>,<un.q',q'>,
nasalization.vowel-fronting-a.nil,tonique,x) →
geler(c-n,consonne-nasale(c-n))
geler(c,consonne(c)) ;

Cette macro-règle peut se traduire par :

$$Oo\ C_n \rightarrow un\ /\ C$$

où C_n est une consonne nasale et C une consonne. Elle indique l'enchainement des règles auquel elle correspond, et la position relative à l'accentuation.

les deux prédicats figurant en partie droite de clause sont des contraintes d'application : elles seront déclenchées dès que le phonème correspondant aura été déterminé.

L'usage de ces macro-règles est possible grâce à des accès définis sur les phonèmes gauches des deux formes concernées.

Exemple :

<acces-descendant-Oo,macro-regle11> → ;
<acces-remontant-un,macro-regle11> → ;

L'analyse d'un mot se fait maintenant de gauche à droite, on accède à une macro-règle par le phonème de gauche. Les variables q et q' servent à récupérer la partie droite du mot non encore analysée, et dont le premier phonème permet l'accès à la règle suivante.

Cette technique d'analyse est presque déterministe. Les temps de calcul sur VAX 750 sont maintenant de l'ordre de la seconde.

Le remplacement des règles phonologiques par des macro-règles revient à déduire un catalogue des différents

cas particuliers. Cependant, ce catalogue est obtenu automatiquement à partir des connaissances que les linguistes souhaitent manipuler. Il ressort donc que ce système peut convenir à la fois à la mise au point d'un jeu de règles et à son exploitation en situation de reconnaissance. Les macro-règles apportent une solution informatique compatible avec la formulation linguistique choisie.

VIII - CONSTRUCTION AUTOMATIQUE DU LEXIQUE

Un mot nouveau est proposé au système sous sa forme phonétique, avec ses attributs syntaxiques (catégorie, type d'emploi...). Le système en fait d'abord l'analyse phonologique, qui propose une forme sous-jacente dont il pourrait dériver.

L'analyse en morphèmes de cette forme ne peut être envisagée de toutes les manières possibles sans référence à des morphèmes connus. On imposera donc que les préfixes, suffixes et désinences soient tous répertoriés à priori dans le lexique. Seul le radical pourra donc être inconnu. On utilise un procédé voisin de la technique des rectifications successives [9].

Ainsi ne peuvent subsister que certaines ambiguïtés d'analyse résultant de la confusion d'une partie du radical avec un préfixe ou un suffixe existant. Le choix de l'une ou l'autre forme est déterminé par d'autres mots de la même famille, qui possèdent le même radical. Ces ambiguïtés sont levées par l'utilisateur qui doit fournir au système un mot de même famille.

IX - CONCLUSION

Cette étude met en évidence certains aspects de l'approche choisie par rapport au traitement automatique du problème :

- l'interprétation des connaissances proposées est délicate, certains aspects restant implicites,
- il est difficile de mélanger des règles provenant de plusieurs auteurs, celles-ci utilisant souvent des jeux de traits différents, dont la correspondance n'est pas évidente,
- le système a permis de valider l'hypothèse de faisabilité sous Prolog (moyennant la compilation des règles),
- il constitue un outil de test pour de nouvelles théories phonologiques que l'on pourrait appliquer de la même manière,
- cette approche, dans les applications de reconnaissance, permet de suivre l'évolution du matériel théorique disponible.

BIBLIOGRAPHIE

[1] Bescherelle, Le Nouveau Bescherelle Hatier 1980

[2] Chomsky N., Halle M. The Sound Pattern of English Cambridge, Mass. MIT Press 1968

[3] Dell F. Les règles et les Sons, introduction à la Phonologie Générative Hermann Paris 1973

[4] Dell F. E muet : fiction graphique ou réalité linguistique ? in Andeson Kiparsky " a festchrift for Morris Halle" 1975

[5] Dell F., On French Phonology and Morphology and some vowel alternations in French, Studies in French Linguistics, Vol. 1, N°3, January 1979

[6] Dell F., Vergnaud J.-R. Les développements récents en Phonologie : quelques idées centrales Forme Sonore du Langage Hermann 1984

[7] Dell F. Deux nasalisations en Français Séminaire Lexique et Traitement Automatique des Langages Toulouse Janvier 1986

[8] Gispert J. Représentation d'un lexique pour la R.A.P.C. à l'aide de connaissances phonologiques Séminaire Lexique et Traitement Automatique des Langages Toulouse Janvier 1986

[9] Gruaz C. La dérivation suffixale en français contemporain dans les familles des monosyllabes de haute fréquence. Thèse d'état Juin 1984

[10] Dictionnaire de la prononciation Larousse 1980

[11] Martinet Syntaxe Générale Armand Colin 1985

[12] Meloni H. Etude et réalisation d'un système de Reconnaissance Automatique de la Parole Continue Thèse d'Etat Université d'Aix-Marseille II 1982

[13] H.Meloni, J.Gispert, J.Guizol Traitement de Connaissances Déclaratives pour l'identification analytique de mots dans le Discours Continu. 14èmes JEP Paris 1985

[14] Perennou G. Base de Données Lexicale Rapport scientifique GRECO communication parlée Juin 1984 CRIN Nancy

[15] J.Pinchon, B.Couté le Système Verbal du Français Nathan 1981

[16] Plenat M. La loi de Littré Cahiers de Grammaire n°2, Toulouse 1980

[17] Plenat M. L'autre conjugaison, ou de la régularité des verbes irréguliers Cahiers de Grammaire n°3 Avril 81 Centre de Sociologie et de Dialectique Sociale Toulouse 1981

[18] Plenat M. Sur quelques aspects de la nasalisation en Français standard Cahiers de Grammaire n°9, Toulouse 1985

[19] Lexique et Phonologie : observations sur la liaison, la nasalisation et le comportement des liquides en français standard Séminaire Lexique et Traitement Automatique des Langages Toulouse Janvier 1986

[20] Schane S.A. French Phonology and Morphology Cambridge, Mass. MIT Press 1968

PROPOSITIONS POUR UN MODELE SEMANTIQUE SIMPLIFIE DE LA COMPLEXITE DES SIGNIFIES

G. Caelen-Haumont

Laboratoire C.E.R.F.I.A. - Toulouse

ABSTRACT

This study fits within the scope of cognitive sciences and more precisely the natural understanding of texts. This paper aims to purpose a model stemmed from the reflection about the formal nature of signification in the area of the semantic complexity of the lexical items in the texts. Otherwise this model tries to quantify this semantic complexity by weighting each category of concept defined in the model.

This method is suitable to be applied to text analysis (relation writer/text) and text reading experiments (relation reader/text). In this latter case, the aim is to verify the hypothesis according to which there exist actual traces of the construction or the stucturation of the signification within the prosodic continuum of speech. This idea states 1/ the whole theme of research and 2/ the particular frame of this paper where we try to apply this method to the analysis of two texts.

INTRODUCTION

Selon G. NOIZET [1], "Construire la signification d'une phrase, c'est dégager, à l'aide de procédures de traitement de nature diverse (identification et catégorisation des unités, mise en relation des unités, etc.), un réseau complexe de relations, une structure qu'on pourrait qualifier de structure cognitivo-sémantique pour bien montrer d'une part que cette structure résulte d'opérations formellement définissables (prédication, implication, ordination temporelle, etc.) d'autre part que les termes sur lesquelles portent ces opérations sont analysables en unités de sens (LE NY [2])".

Cette définition de la construction de la signification précise le cadre général de l'étude que nous présentons. Cette dernière s'inscrit en effet dans le domaine de la compréhension naturelle des textes. Plus précisément, elle se propose de décrire un modèle d'analyse de la complexité des signifiés d'un texte en vue de déterminer ultérieurement si les réalisations prosodiques des locuteurs (F_0, énergie, durée, pauses) ne portent pas les indices d'une élaboration ou d'une structuration de la signification.

MODELE D'ANALYSE DE LA COMPLEXITE DES SIGNIFIES

Principe général

Ce modèle original (cf tableau 1) se présente sous la forme d'un graphe procédural constitué de différents niveaux de notions catégorielles assorties d'une quantification. Il se compose donc d'une réflexion de type fondamentaliste sur la nature formelle de la signification dans le domaine de la complexité des signifiés et propose en outre une quantification relative au degré de complexité dégagé. Il s'agit en effet de faire correspondre à chaque item lexical saisi dans son contexte (item simple ou composé comme dans le cas des expressions lexicalisées ou "lexies" selon B. POTTIER [3]) un poids qui donne une image simplifiée et la moins subjective possible de sa complexité. Cette complexité est appréciée au sein d'une échelle que fournit le graphe (de 1 à 25) : le poids total de l'item est calculé par addition successive des poids attachés aux catégories, ces dernières étant sélectionnées en fonction de la spécificité de l'item dans son contexte. Les flèches parallèles indiquent l'existence de macrocatégories en concurrence (cf substance et attribut) dominant l'une et l'autre des catégories de niveau inférieur pondérées (cf état... / intrinsèque...).

Bien entendu, nous ne prétendons pas définir un modèle absolu de la complexité sémantique, ni même avoir résolu le problème de la subjectivité de l'analyse : ce modèle se définit comme une approche simplificatrice et perfectible d'une analyse de la complexité des signifiés et propose simplement une quantification subjective et relative à ce type de modèle .

Par ailleurs, le modèle s'est attaché à décrire l'effet sémantique produit par la contextualisation des items lexicaux auxquels sont conviés les divers procédés syntaxiques. Cependant nous avons adopté le parti de les considérer comme implicites : nous ne chercherons donc pas à les analyser.

Il est important d'attirer l'attention sur le fait que la méthode d'analyse est textuelle et non lexicale, la perspective se situant à l'intérieur des relations lexicales et non pas à un niveau intralexical. En effet pour analyser cette complexité, 3 relations sont explorées :

- relation entre signification interlexicale et sens intralexical
- relations de signification dans le syntagme (caractérisation sémique ou lexicale)
- relations de signification dans le texte (caractérisation sémique ou lexicale).

Lorsque le modèle est appliqué à un texte lu, il nous a semblé interessant d'adjoindre un module spécifique, à caractère psycholinguistique, faisant intervenir l'adhésion du locuteur. Cette composante n'est pas en relation avec la complexité sémantique dans son contenu mais en rapport externe avec elle, puisque l'absence d'adhésion de la part du lecteur correspond en fait à un foyer de résistance sémantique et une irréduction de la complexité sémantique.

En effet il semble que tout texte suppose dans sa phase d'écriture (auteur) comme de lecture ou d'écoute, une adhésion de l'individu, considérée d'ailleurs le plus souvent comme implicite. Cette adhésion (qui inclut la compréhension) requiert une confrontation entre le sens intralexical aux niveaux dénotatif et connotatif et les réseaux interlexicaux de sigification, dépendant tous deux de la "compétence" spécifique de l'individu. Si les compétences interindividuelles (au sens le plus général) ne coïncident pas, l'adhésion est rendue impossible. Les virtualités de signification sont donc des composantes effectives du texte dans son extension maximale et la production de texte suppose donc, à sa création comme à sa réception, une actualisation restrictive des potentialités de sens intralexical.

Les modules et leurs catégories

Outre la référence au sens lexical et la mention de l'adhésion, le modèle comprend deux modules, le premier ayant pour champ d'application les relations intrasyntagmatiques, le deuxième, les relations lexicales intersyntagmatiques.

Module intrasyntagmatique

Au sein du syntagme et en référence au sens lexical, le lexème (et lexie) est analysé en 5 niveaux :

1/ registres fondamental, standard ou spécialisé banalisé, et spécialisé (poids respectifs +1, +4, +7).

Le registre fondamental fait référence au "Français Fondamental" (1er et 2ème degré) [3], le registre spécialisé aux termes spécifiques non tombés dans l'usage commun et dans le cas de l'application de ce modèle à la lecture, il est conçu comme n'appartenant pas à la spécialité du locuteur; dans le cas contraire, le modèle renvoie au registre standard ou spécialisé banalisé. Le poids le plus fort du modèle est attribué au registre spécialisé.

2/ référents concret, concret/abstrait et abstrait ou imaginaire (poids 0, +2, +4). Seule la catégorie concret/abstrait demande à être explicitée : elle recouvre en effet des termes qui lexicalement concrets s'appliquent cependant par l'effet de textualisation à des notions abstraites ou imaginaires, ou bien désignent des termes qui présentent un référent concret sous une catégorisation abstraite (ex: "biologiste" : être humain concret + catégorisation abstraite).

3/ notion d'essence

Aux deux catégories de substance et d'attribut (qui n'entraînent une quantification que par les catégories qu'ils dominent) correspondent :

- pour la substance, 2 niveaux d'analyse :

(a) Le premier distingue les sous-catégories d'état et de procès qui s'opposent comme les notions de structure et d'évolution (aspect statique vs. dynamique) avec la possibilité d'une combinaison de ces deux notions (poids respectifs 0, +1, +2). Des termes comme "addition" par exemple peuvent contextuellement tour à tour désigner un état, un procès, ou la combinaison des deux (état résultatif d'un procès).

(b) Le deuxième envisage quatre types d'opposition. Les étiquettes discontinu/continu renvoient à la fois aux notions spatiales (état) et temporelles (procès). En effet il apparaît que la notion statique est réintroduite de manière secondaire dans la catégorie du procès (et corrélativement la notion dynamique dans la catégorie de l'état) par la notion de continuité (et corrélativement par la notion de "discontinuité" pour la catégorie de l'état). Selon qu'il s'agit donc de l'une ou l'autre catégorie, les poids associés aux étiquettes discontinu et continu s'inversent (0, +1 pour l'état et +1, 0 pour le procès). Le modèle prévoit une neutralisation possible de ces catégories dans le cas par exemple des termes abstraits, et d'autre part inversement une combinaison de ces notions (poids 0, +1). Lorsque les deux catégories état et procès sont combinées, il s'opère une neutralisation des notions de continuité /discontinuité (poids 0).

- pour l'attribut, un seul niveau d'analyse. La notion d'attribut est complexe dans la mesure ou elle renvoie aux qualités d'un objet considéré comme état ou procès : même si les contenus de ces deux types d'attribut ne sont pas comparables, le processus de caractérisation est dans notre perspective analysable par les mêmes structures. En effet, cette notion d'attribut peut renvoyer à l'analyse des adjectifs qualificatifs ou des adverbes, les uns et les autres précisant de manière comparable des qualités ou des modalités attachées aux noms, adjectifs, adverbes et verbes. Le modèle ne prévoit par ailleurs aucune exclusion de classe adverbiale (en particulier les modificateurs et les modalisateurs).

Dans cette catégorie de l'attribut, on distingue entre les attributs intrinsèques (redondance des sèmes de l'attribut par rapport à la notion désignée), les attributs de complémentarité (pas de redondance mais concordance des sèmes), et les attributs extrinsèques (addition de sèmes).

Ces attributs extrinsèques correspondent à une qualification maximale, parfois à une désignation notionnelle spécifique : dans ce cas les frontières syntaxiques entre les catégories du nom et de l'adjectif sont les plus floues, la pertinence de l'information pouvant reposer essentiellement sur l'adjectif. Les conditions du glissement de l'adjectif vers la catégorie du substantif sont alors réunies (ex : l'océan atlantique --> l'atlantique; le chemin de fer métropolitain --> le métro).

Module intersyntagmatique

Dans ce module, les relations entre lèxèmes sont saisies à un niveau supérieur par l'intermédiaire des syntagmes.

Le module intersyntagmatique prévoit trois niveaux d'analyse, opérant sur le plan de la forme ou du contenu. Au niveau formel, une seule catégorie concernant le réemploi d'un terme de registre standard ou spécialisé, réemploi conçu comme une diminution de la complexité dans ce registre (-3), et neutralisé dans le cas du registre fondamental.

Le deuxième niveau est celui de la figuration envisagée selon 3 degrés d'expressivité croissante : pas de figuration ou cliché, figuration lexicalisée (ex: "à brûle pourpoint"), figuration originale (poids respectifs : 0, +3, +5).

Le troisième niveau envisage l'appartenance d'un terme à un champ lexical comme propre à réduire la complexité de ce terme : appartenance / pas d'appartenance, initialisation du champ ou changement (poids respectifs 0, +2).

Module psycholinguistique

Comme on l'a précisé ci-dessus, ce module n'est à considérer que dans le cas de l'actualisation orale d'un texte, lecture par exemple. Le phénomène de l'adhésion est pris en considération dans le modèle dans la mesure où il est susceptible surtout lors d'une première lecture, de modifier considérablement le rapport pragmatique de l'individu au texte, surtout lors d'une première lecture. L'absence d'adhésion est assortie d'un poids +5.

APPLICATION

Nous avons appliqué le modèle à l'analyse de deux textes, choisis pour leurs styles et leurs instances d'énonciation différents : le premier est un extrait tiré d'un ouvrage de R. BARTHES [4], le second un extrait de l'Amour Fou d'A. BRETON [5].

Extrait de S/Z de R. BARTHES

Le texte de 73 items lexicaux est le suivant : "La connotation est la voie d'accès à la polysémie du texte; il faut la garder comme la trace nommable, computable, d'un certain pluriel du texte. Qu'est-ce donc qu'une connotation ? C'est une détermination, une relation, un trait qui a le pouvoir de se rapporter à des mentions antérieures, ultérieures ou extérieures au texte, à d'autres lieux du texte ou d'un autre texte. Il ne faut surtout pas confondre connotation avec association d'idées: celle-ci renvoie au système d'un sujet, celle-là est une corrélation immanente au texte, aux textes; c'est si l'on veut, une association opérée par le texte-sujet à l'intérieur de son propre système ... Les connotations sont des sens qui ne sont ni dans le dictionnaire ni dans la grammaire de la langue dont est écrit un texte ... La connotation assure une dissémination des sens, répandue comme une poussière d'or sur la surface apparente du texte ... Fonctionnellement la connotation, engendrant par principe le double-sens, altère la pureté de la communication; c'est un "bruit" volontaire, soigneusement élaboré par l'auteur."

Extrait de l'Amour Fou d'A. BRETON

Le texte comportant 66 items lexicaux a été réduit à la partie descriptive des lieux, personnages et des coïncidences (en particulier celle d'une éclipse de lune et d'un bijou suggérant le phénomène) qui ont eu pour cadre un "petit restaurant" fréquenté par l'auteur : "Le soir le patron, "qui fait la cuisine", regagne son domicile à motocyclette. Des ouvriers semblent faire honneur à la nourriture. Le plongeur, vraiment très beau, d'aspect très intelligent, quitte quelquefois l'office pour discuter, le coude au comptoir, de choses apparemment sérieuses avec les clients. La servante est assez jolie : poétique plutôt. Le 10 avril au matin elle portait, sur un col blanc à pois espacés rouges fort en harmonie avec sa robe noire, une très fine chaîne retenant trois gouttes claires comme de pierre de lune, gouttes rondes sur lesquelles se détachait à la base un croissant de même substance, pareillement serti. J'appréciai, une fois de plus, infiniment, la coïncidence de ce bijou et de cette éclipse."

Analyse

Le tableau n°2 propose une répartition des signifiés du texte en fonction du poids total de l'item. Ces deux textes n'ont pas exploité toute la complexité théorique du modèle (valeur de quantification jusqu'à 20 pour un texte non soumis à une expérience de lecture) : la valeur maximale est de 12 pour l'Amour Fou et de 15 pour S/Z. L'opposition des deux textes est clairement rendue par la répartition des lèxèmes à travers la grille, le premier se situant dans des régions de complexité minimale ou moyenne, le second dans des zones moyennes ou au-delà. On retrouve cette opposition au niveau de la valeur de complexité moyenne des textes : 5.0 pour l'Amour Fou, 8.1 pour S/Z.

Il est nécessaire de préciser que la comparaison de la répartition terme à terme des items lexicaux n'est valide qu'à l'intérieur d'un même texte de même que seule reste valide pour plusieurs textes la comparaison des valeurs moyennes de leur complexité . En effet la grille de répartition ne présente pas des classes de lexèmes groupés par affinités de sens, mais des unités relationnelles actualisées par la textualisation et donc dépendantes du texte.

Remerciements

Nous tenons à remercier particulièrement A. BORILLO, J. COURTES et E. LHOTTE pour les suggestions et critiques fructueuses du modèle que nous leur avons soumis.

REFERENCES

[1] G. Noizet, L'activité opératoire lors de la lecture de prases, Bull. de Psychologie, XXXV, n°356, pp. 607-619.
[2] J-F Le Ny, La sémantique psychologique, Paris, Presses Universitaires de France, 1979, 257 p.
[3] Institut National de Recherche et de Documentation Pédagogiques, Le Français fondamental, I et II, ouvrage collectif, Paris, 1972-1973.
[4] R. Barthes, S/Z, Ed. du seuil, 1970.

[5] A. Breton, L'Amour Fou, Gallimard, Folio.
[6] G. Caelen-Haumont, Grammatical components and macro-prosody : quantitative analysis toward statistical correlations, Proceedings of 13TH ICA, Montréal, 1986 (to be published).

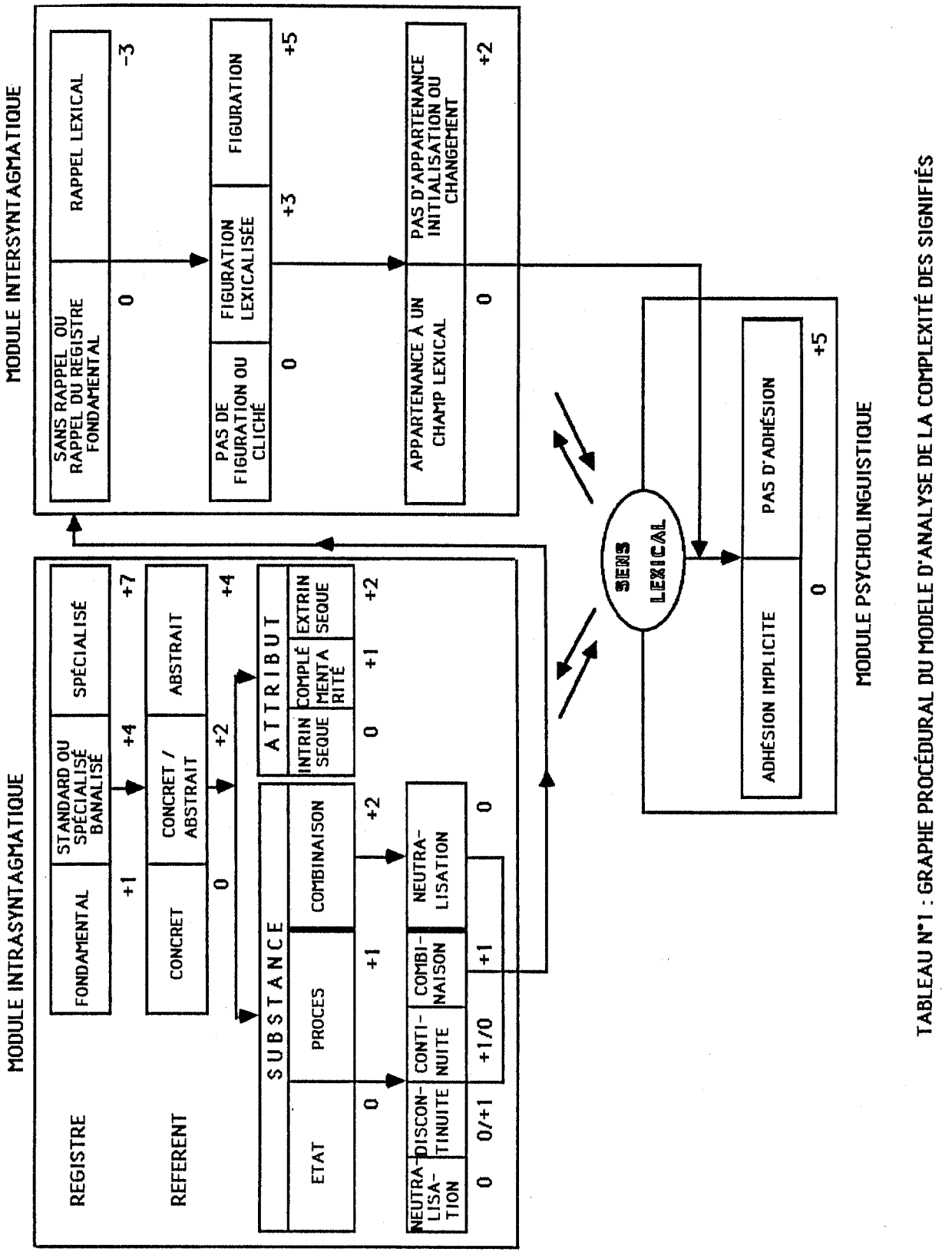


TABLEAU N°1 : GRAPHE PROCÉDURAL DU MODELE D'ANALYSE DE LA COMPLEXITÉ DES SIGNIFIÉS

Tableau n° 2 : Complexité sémantique des signifiés et Quantification.
Application à un extrait de l'Amour Fou d'A. Breton et de S/Z de R. Barthes

Poids	1	2	3	4	5	6	7	8
L'Amour Fou	motocyclette ouvrier client col blanc pois rouge robe noir fin chaîne rond	portrait goutte(2) détacher	patron ouvrier quitter coude matin clair base bijou	faire la cuisine plongeur office discuter comptoir servante retenir	domicile nourriture aspect très(2) chose assez joli fort espacer même	intelligent quelquefois sérieux sertir	soir regagner vraiment très beau plutôt avril substance	sembler harmonie apprécier éclipse
S/Z		texte(8)			texte renvoyer système sens(2) dictionnaire grammaire langue surface auteur	rapporter association opérer assurer	garder autre(2) sujet surtout texte-sujet écrire apparent	pouvoir lieu intérieur principe double-sens

Poids	9	10	11	12	13	14	15	16
L'Amour Fou	faire honneur pareillement infiniment			poétique coïncidence pierre de lune				
S/Z	certain antérieur ultérieur extérieur confondre association d'idées propre répandre poussière d'or élaborer	trace nommable connotation(5) relation trait système corrélation dissémination communication bruit volontaire	voie d'accès polysémie mention soigneusement	pluriel	fonctionnellement engendrer altérer pureté		connotation computable détermination immanent	

ARCANE : ACQUISITION ET RECHERCHE DE CONNAISSANCES ACOUSTICO-PHONETIQUES DANS UN NOYAU EVOLUTIF

J. Caelen, G. Caelen-Haumont, N. Vigouroux, C. Barrera, J. Malet

Laboratoire CERFIA UA-CNRS N°824
Université P. Sabatier, 118 Route de Narbonne
31062 Toulouse CEDEX

ABSTRACT

ARCANE is a system able to aid the acquisition and research of acoustico-phonetic knowledge from speech database. Its architecture has an iterative structure built around a production-management nucleus. Information are organized from data to knowledge and are split into four database. Their structure relationships give us a powerful tool for studying speech.

1. INTRODUCTION

A l'heure actuelle, l'utilisation des systèmes experts (SE) en reconnaissance automatique de la parole (RAP) pose clairement un double problème:

(a) celui de la constitution de bases de connaissances performantes

(b) celui de la modélisation d'un raisonnement approprié aux données disponibles et aux connaissances acquises.

Bien que ces deux problèmes soient liés, il est évident que la carence la plus nette, en matière de décodage acoustico-phonétique, concerne le point (a): on trouve en effet davantage de résultats formalisés au point (b) [Gillet et al, 83], [Memmi et al, 83], [De Mori, 83], [Carbonnel et al, 84], [Haton, 85]. La mise en forme des connaissances passe par l'abstraction des données brutes, par leur description conceptuelle --pour les plonger dans des classes plus vastes-- par le filtrage des résultats partiels et enfin par l'unification de ces résultats selon les buts à atteindre (tout en visant la cohérence et la stabilité de la base ainsi produite). Cette démarche reflète en fait, une certaine forme de raisonnement, qui part des données pour aboutir aux connaissances et peut se concevoir comme un système d'aide à l'expertise dans l'acquisition et la gestion des connaissances.

C'est autour de ces quelques points que nous nous proposons de centrer nos réflexions, plus précisément sur le problème de l'apprentissage des connaissances, en tentant de dégager quelques perspectives et quelques outils mis en oeuvre dans le projet ARCANE: Acquisition et Recherche de Connaissances Acoustico- phonétiques dans un Noyau Evolutif.

2. LA PROBLEMATIQUE DU DECODAGE ACOUSTICO-PHONETIQUE (DAP)

Le DAP peut être envisagé comme un problème plus général de mise en correspondance d'une *substance* et d'une *forme* [Rossi, 83], chacune organisée dans des plans successifs (Fig. 1). L'analyse se situe précisément à la jonction de ces plans et participe de l'*ascendant* comme du *descendant*. C'est à ce niveau que s'effectue le couplage entre le *concret observable* (substance acoustique) et l'*abstrait linguistique* (forme phonologique). Dans une telle vision, on ne considère plus le décodage comme une transformation du signal en unités phonétiques discrètes par analyses successives, mais, au contraire, comme une mise en relation avec ces unités: substance et forme coexistent et se correspondent --cela serait à rapprocher des hypothèses récentes sur l'agencement des étapes

de traitement de l'information dans le cerveau [Changeux, 84]. Cette opération de *couplage* ne peut se faire que si la substance est vidée de toute redondance vis-à-vis du *moule* concret de la forme. Ce moule, défini par un modèle phonético-phonologique, est généré dans le plan phonologique de surface sous forme d'un *code ordonné* [Benveniste, 66] qui contient la succession des faits (pile des faits et des contextes) filtrés dans le plan perceptif. La recherche de l'adéquation entre *contenant* et *contenu* est le but du DAP, à travers les unités phonétiques elles-mêmes, mais aussi à travers leurs relations et les connaissances qui les sous-tendent. Ces connaissances sont de nature diverse:

(a) perspective ascendante:

- *articulatoires,* c'est-à-dire de type mécanique (gestes articulatoires, cibles, etc), le plus souvent *implicites* en DAP. Elles doivent donc être corrélées avec des faits acoustiques pour la phase ultérieure d'interprétation.
- *acoustiques,* elles décrivent les faits acoustiques sur l'axe temporel de manière combinatoire.
- *perceptives,* pour tenter d'abstraire ces faits par empilement dans une mémoire cumulative. Bien que les mécanismes perceptifs humains ne soient pas encore bien connus à l'heure actuelle, il est important de prévoir dans les systèmes de DAP une couche de traitement, même informelle, permettant de modéliser ces phénomènes.

(b) perspective descendante:

- *phonologiques,* décrivant le système de contraintes de la langue et dérivant du lexique (BDLEX par exemple [Perennou et al, 81]). A ce niveau il faut pouvoir prédire les chaînes phonétiques admissibles [Shoup, 80].
- *phonétiques,* pour réduire et expliquer la variabilité contextuelle, individuelle et idio-syncrasique.

3. LES CONNAISSANCES

Le plus grand soin doit être apporté à la représentation cohérente des connaissances énumérées ci-dessus et, dans un deuxième temps, ces connaissances doivent être structurées et/ou hiérarchisées pour éclairer le *raisonnement* modélisé par exemple, dans un moteur d'inférences pour le DAP.

Ces connaissances peuvent être classées en cinq catégories:

(C1)- *descriptives,* quantifiables ou non, par exemple à travers un système d'indices instantanés et contrastifs (au niveau acoustique) et un système de traits (au niveau phonologique).

(C2)- *factuelles,* vis-à-vis des locuteurs, de l'environnement, etc., pour fixer les contraintes du problème et décrire les sources implicites d'informations utilisées (dialectes, particularités linguistiques ou articulatoires du locuteur).

(C3)- *relationnelles,* en tant qu'actualisation d'un système phonétique, sur les axes syntagmatique (contraintes phonotactiques) [Zue, 83] et paradigmatique (oppositions et combinatoire des unités phonétiques).

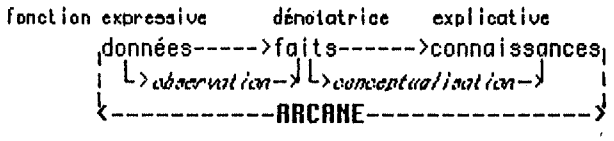
(C4)- *représentatives,* du modèle explicite (acoustique, phonétique, etc),

(C5)- *explicatives,* pour remonter des conséquences aux causes (méta-connaissance, savoir-faire).

4. ACQUISITION DES CONNAISSANCES

4.1. Méthodologie

L'acquisition et la recherche de connaissances à partir des données peuvent être comparées aux fonctions *expressive, dénotatrice et explicative* du signe selon le schéma ci-après:



Le système ARCANE, placé sous contrôle de l'expert, contient deux fonctions essentielles, l'*observation* et la *conceptualisation.* Il est à noter que ce système ne couvre pas toute l'étendue du problème visant à produire l'ensemble des règles après apprentissage (le filtrage de la connaissance, l'unification des règles, la cohérence et la convergence de la base ne sont pas pris en compte à ce niveau). Il permet d'accumuler les *connaissances* C1,C2,C3 énumérées au paragraphe 3 à partir des *données,* en passant par des *faits* (connaissances

descriptives) et ce, sous contrôle de l'expert, qui possède, lui, les connaissances C4 et C5. La fonction d'observation consiste alors à sélecter les données, les ordonner, les répartir en classes a priori avant de leur appliquer des analyses prédéfinies (analyse de données par exemple). Les faits obtenus décrivent l'état de l'univers des données traitées. Traduits en langage naturel, ces faits sont par exemple:

ƒ1: dans le corpus X, il y a 20% de phonèmes /a/ aigus,
ƒ2: l'opposition /i/-/y/ est maximale pour l'indice bémolisé-diésé,
ƒ3: le F1 moyen de /u/ est 320 Hz,
etc.

La fonction de conceptualisation consiste à regrouper ces faits en classes hiérarchisées et à les décrire à l'aide d'un système relationnel. Dans un langage plus formel, un fait peut s'écrire:

fait:=<Opérateur; Opérandes; Attributs>
ce qui donne pour les exemples ci-dessus:
ƒ1: 20:=<Fréquence;/a/;corpus_X,indice_aigu>
ƒ2: Max:=<Opposition;/i/_/y/;indice_bémolisé>
ƒ3: 320:=<Moyenne;F1;/u/>
etc.

4.2. Architecture du système

L'architecture du système ARCANE a une structure itérative (Fig. 2) construite autour d'un noyau Production-Gestion (Fig. 3). Les informations sont hiérarchisées des données aux connaissances et éclatées en quatre bases:

- BDAP, base de données AP (acoustico- phonétique)
- BOAP, base d'observations AP,
- BFAP, base des faits AP,
- BCAP, base de connaissances AP.

La base BDAP sert uniquement de tampon entre les données et les faits, on peut la considérer comme une base de travail.

La production des données fait intervenir l'expert à deux niveaux:

- pour spécifier le mode de production (nature des données, procédures, paramètres)
- pour modéliser et placer des informations quantitatives (pondérations, probabilités, etc.) ou qualitatives (étiquettes) en correspondance avec les données. Pour cela des interfaces spécialisées sont mises en oeuvre dans chaque base.

Les systèmes de gestion de ces bases sont construits sur le même modèle, à partir de la notion très générale d'objet. Des descripteurs, que l'utilisateur peut lui-même définir, permettent de modifier et de spécifier les objets manipulés [Vigouroux et al, 86]. Le système établit des relations entre les "objets" que constituent les faits ou entre les sous-objets (opérateurs, opérandes, etc.). Par exemple sous le concept OPPOSITION seront réunis tous les faits pour lequel "Opposition" est un opérateur. Les connaissances sont ainsi traitées à l'aide d'un système de gestion relationnel [Delobel et al, 83] [Zurfluh, 85] --elles ne peuvent dans cette méthode, être mises directement sous forme de règles.

5. CONCLUSION

Dans cet article nous avons voulu faire apparaître l'intérêt du projet ARCANE centré autour de la problématique de l'acquisition et de la représentation des connaissances acoustico-phonétiques. Pour ARCANE il ne peut s'agir encore que d'un système d'aide à l'expertise en cours de conception --qui nécessite en outre plusieurs types d'experts (phonéticien, statisticien, etc.). Actuellement les bases BDAP et BOAP sont opérationnelles, la base BFAP est en cours de mise au point et la base BCAP est en phase de spécification. Leur parenté de structure en fait un outil puissant pour l'étude de la parole, des données aux connaissances.

6. REFERENCES

[Benveniste, 66] E. Benveniste, *Les niveaux de l'analyse linguistique,* Problèmes de linguistique générale, 1966, pp. 119-131.
[Caelen et al, 86] J. Caelen, N. Vigouroux, *Producing and Organizing Phonetic Knowledge from Acoustic Facts in multi-level Data-Information,* IEEE-ICASSP, Tokyo, 1986.
[Carbonnel et al, 85] N. Carbonnel, J.P. Haton, J.M. Pierrel, F. Lonchamp, *Techniques d'intelligence artificielle en décodage*

acoustico-phonétique, 14èmes JEP, GALF-CNRS, Paris, 1985.

[Changeux, 84] J.P. Changeux, *L'homme neuronal*, Dunod éd., 1984.

[Cole et al, 81] R.A. Cole, R. Reddy, *Knowledge acquisition and knowledge engineering in speech understanding research*, CMU Report, 1981.

[Delobel et al, 83] C. Delobel, M. Adiba, *Bases de données: des modèles réseau et hiérarchique au modèle relationnel*, TSI, Vol. 2, N°1, 1983, pp. 43-62.

[De Mori, 83] R. De Mori, *Computer Models of Speech Using Fuzzy Algorithms*, Plenum Press, New York and London, 1983.

[Gallaire, 84] H. Gallaire, *Logic and Databases: a Deductive Approach*, ACM Computing Surveys, Vol. 16, N° 2, 1984.

[Gillet et al, 84] D. Gillet et al, *SERAC: un système expert en reconnaissance acoustico-phonétique*, 4ème congrès AFIA, AFCET-INRIA, Paris, 1984.

[Haton, 85] J.P. Haton, *Les systèmes à bases de connaissances dans la communication homme-machine*, Colloque COGNITIVA, 1985, pp. 211-224.

[Hjemslev, 47] L. Hjemslev, *Structural analysis of language*, Studia Phonetica, Vol. 1, 1947, pp. 69-78.

[Memmi et al, 83] D. Memmi, M. Eskenazi, J. Mariani, A. Nguyen-Xuan, *Un système expert pour la lecture de sonagrammes*, Speech. Com., Vol. 2, N° 2-3, 1983, pp. 234-236.

[Mercier, 83] G. Mercier, *Le décodage acoustico-phonétique de la parole*, Bulletin de l'INRIA, N° 84, 1983.

[Perennou et al, 81] G. Perennou, M. DeCalmès, *Le décodage au niveau phonologique dans ARIAL II*, Actes du séminaire "Processus d'encodage et de décodage phonétiques", 1981, pp. 191-203.

[Perennou et al, 83] G. Perennou, M. DeCalmès, *Spécifications pour un système générateur de modèles de décodage phonétique*, 14èmes JEP, GALF-CNRS, Paris, 1985.

[Rossi, 83] 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.

[Shoup, 80] J.E. Shoup, *Phonological Aspects of Speech Recognition*, in Trends in Speech Recognition, W. LEA ed., Prentice-Hall 1980.

[Vigouroux et al, 86] N. Vigouroux, J. Caelen, *Le système de traitement des informations dans ARCANE*, Journées BD, AFCET, Paris, 1986.

[Zue, 83] V.W. Zue, *The use of phonetic rules in automatic speech recognition*, Speech. Com., Vol. 2, N° 2-3, 1983, pp. 181-186.

[Zurfluh, 85] G. Zurfluh, *Bases d'Informations Généralisées: le modèle agrégatif et son langage de manipulation*, Thèse d'Etat, Toulouse, 1985.

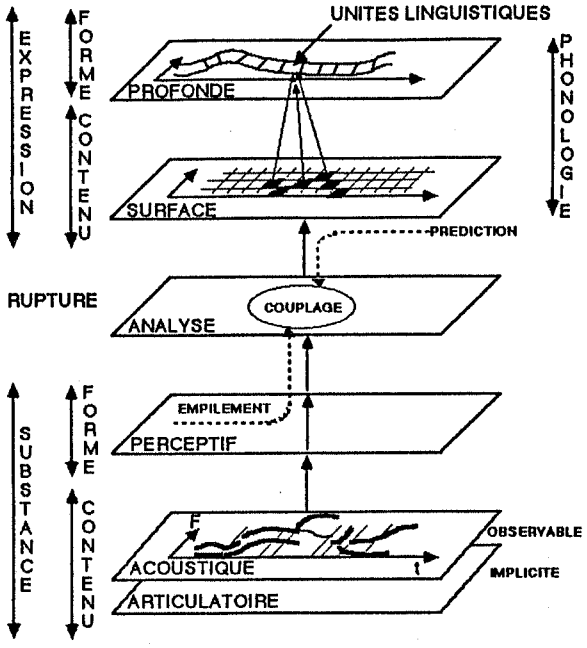


Fig. 1: Les plans du DAP

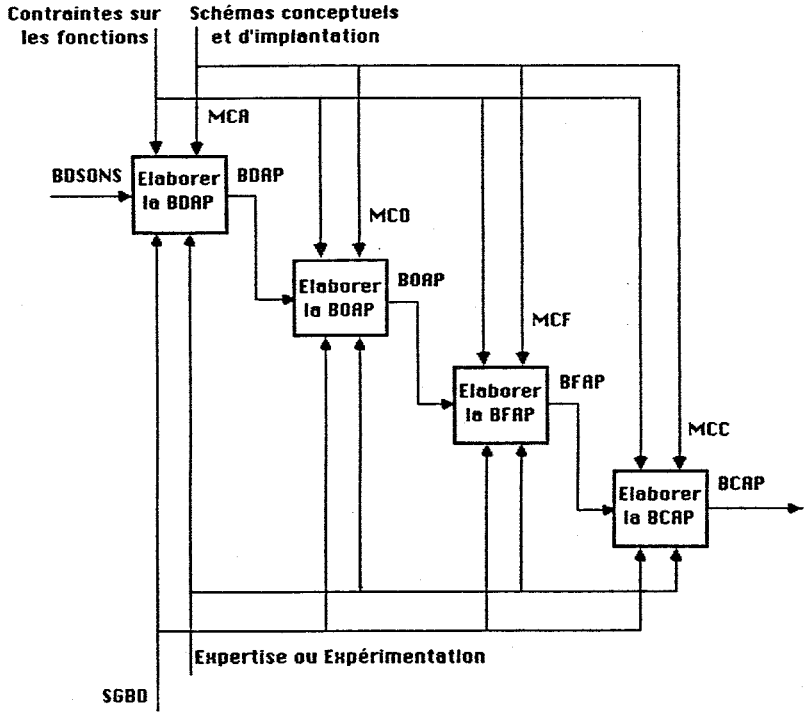


Fig. 2 : Architecture du système ARCANE.

BDSONS	*Base de données des sons,*
BDAP	*Base de données acoustiques et phonétiques,*
BOAP	*Base des observations acoustiques et phonétiques,*
BFAP	*Base des faits acoustiques et phonétiques,*
BCAP	*Base de connaissances acoustiques et phonétiques,*
MCA	*Modèle conceptuel et d'implantation au niveau des échantillons,*
MCO	*Modèle conceptuel et d'implantation au niveau des observations,*
MCF	*Modèle conceptuel et d'implantation au niveau des faits,*
MCC	*Modèle conceptuel et d'implantation des connaissances.*

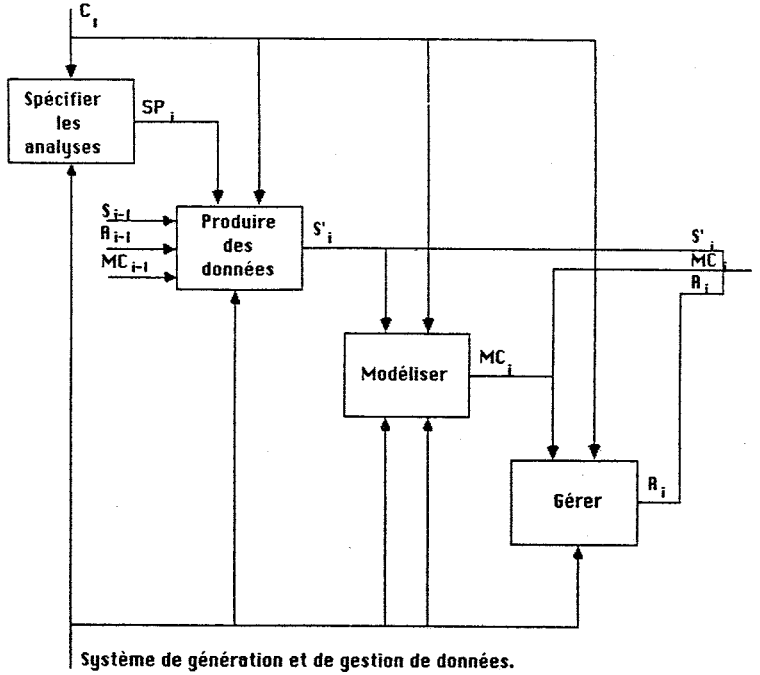


Fig. 3: Architecture d'un système de génération et de gestion de données: "Elaborer une base" au niveau i.

C_i	*Contraintes d'intégrité,*
S_{i-1}	*Données produites par le système de niveau i-1,*
MC_{i-1}	*Schémas conceptuels et d'implantation du niveau i-1,*
B_{i-1}	*Répertoires élaborés par le niveau i-1.*
SP_{i-1}	*Spécifications pour le module "Produire des données.*

DEFINITION ET REALISATION D'UNE BASE DE DONNEES DES SONS DU FRANCAIS

O.Cervantes (1,2), J.F.Serignat (1), R.Descout (3), R.Carré (1)

(1) ICP Grenoble, (2) LGI IMAG, (3) CNET Lannion

ABSTRACT

"BDSON" is a "Speech Communication GRECO" project. The methodology used for the speakers and corpus choices, recording, labeling and segmentation is described. The recordings require 4 Giga-bytes. BDSON can be distributed thanks to a video-betamax recorder associated with a PCM Sony AD converter and a computer interface. BDSON is implemented by means of a Relational Data Base Management System of which the conceptual scheme, the retrieval and updating possibilities are also presented.

I. INTRODUCTION

L'action "Base de Données des Sons du Français" (BDSON) est menée sous l'égide du GRECO COMMUNICATION PARLEE qui regroupe une quinzaine d'équipes de recherche depuis 1983. Cette action répond à la demande des chercheurs de disposer d'une large base de sons, utile tant pour l'étude de la langue française que pour les recherches en traitement automatique de la parole. Ces dernières présentent aujourd'hui deux pôles d'activité principaux : l'évaluation des algorithmes de reconnaissance et l'étude de nouveaux modèles en reconnaissance et en synthèse de la parole.

S'il existe actuellement des Bases de Données Lexicales en documentation automatique, en traduction assistée par ordinateur et en reconnaissance de la parole pour la plupart des langues, les Bases de Données de Sons ne font leur apparition que depuis quelques années seulement à cause de leur complexité de mise en oeuvre : choix des corpus et des locuteurs, volume de stockage important et modalités d'échange et d'accès. Ces bases de données constituent les premiers éléments de ce que l'on commence à appeler les "Industries de la Langue".

II. CONTENU DE LA BDSON

II.1 Choix des corpus et des locuteurs.

Ne pouvant couvrir dès le départ l'étendue des besoins de tous les laboratoires représentés, nous nous sommes limités, dans une première étape, aux deux sous-ensembles suivants :

A) Le corpus "Evaluation", enregistré par 32 locuteurs et composé de 3 parties :

- **Ajustement** : un texte de 5 phrases et 54 groupes CVCV comportant les trois voyelles /a/, /i/, /u/ dans les environnements consonantiques extrêmes /p/, /s/, /f/ en phrase porteuse,
- **Chiffres et nombres** : 400 chiffres isolés, 200 suites de 3 chiffres, 100 suites de 4 chiffres et 100 suites de 5 chiffres "connectés", ainsi que 200 nombres de 0 à 99 et 100 numéros de téléphone,
- **Lettres et noms** : 432 lettres de l'alphabet isolées, 52 noms épelés "en isolé" et 50 noms épelés "en continu".

B) Le corpus "Acoustique", destiné plus particulièrement aux études fondamentales sur les sons du français, enregistré par 12 locuteurs et composé de deux parties :

- **des mots** : 600 groupements CVCV comportant les 17 consonnes du français et les 3 voyelles cardinales /a/, /i/, /u/, 200 groupes consonantiques, et les tests de rimes [1] pour les consonnes et les voyelles par paires et par triplets,
- **des phrases** : 50 phrases phonétiquement équilibrées [2], 44 phrases pour l'étude des nasales et 192 phrases contenant des mots réels avec toutes les consonnes et toutes les voyelles du français.

Le choix des locuteurs, et donc de la norme de prononciation, a été le suivant : 12 locuteurs de base ont enregistré les corpus (A) et (B). Ils ont été sélectionnés par un ensemble de 6 phonéticiens comme étant représentatifs de la prononciation "standard" du français. Les 20 locuteurs complémentaires, qui ont enregistré le corpus (A) sont répartis ainsi : 10 représentent divers accents régionaux et 10 ont été choisis en raison de leur élocution particulière qui les rend difficiles à reconnaître par les systèmes automatiques. Il y a autant de locuteurs féminins que de masculins.

II.2 Enregistrements, fichiers, supports.

Les enregistrements ont été directement numérisés sur 16 bits à une fréquence d'échantillonnage de 16 KHz (bande passante : 7,5 KHz) au CNET de Lannion. La chaîne complète avait un rapport signal/bruit de 85 dB. Les éléments à prononcer étaient présentés sur une console de visualisation et synchronisés avec l'acquisition de façon identique pour tous les locuteurs.

Les fichiers ont une structure qui constitue une norme nationale et en outre ils sont compatibles avec le système ILS. Ils sont composés d'une en-tête de 6 blocs contenant toutes les informations nécessaires à l'accès aux différentes parties de l'enregistrement du signal numérique lui-même, et d'une fin de fichier image du listing détaillé accompagnant celui-ci.

L'ensemble des enregistrements est stocké sur 320 bandes magnétiques numériques. Le corpus (A) occupe un volume de 72 Méga-octets par locuteur, soit un total de 2,3 Giga-octets (200 bandes magnétiques). Le corpus (B) occupe un volume de 126 Méga-octets par locuteur, soit un total de 1,5 Giga-octets (120 bandes magnétiques).

La diffusion de cette base sera faite à l'aide d'un support d'archivage particulièrement intéressant : l'enregistrement numérique sur cassettes vidéo au standard Bétamax, réalisé grâce à un équipement constitué d'un convertisseur PCM Sony associé à un enregistreur vidéo Bétamax et à son interface calculateur OROS-AI (société OROS de Grenoble). Cet équipement permet l'échange de données numériques entre calculateur et cassette avec une fréquence d'échantillonage pouvant varier entre 5 et 44 KHz. De plus, ce dispositif effectue, sur la deuxième piste, un "marquage" d'informations en code ASCII permettant de repérer les fichiers. Ce nouveau type de support est extrêmement économique puisqu'une cassette de 2 heures peut contenir l'équivalent d'environ 18 bandes magnétiques. Une fois dupliqué, le corpus (A) sera contenu dans 12 cassettes, et le corpus (B) dans 7 cassettes seulement.

II.3 Segmentation et étiquetage.

La richesse de la consultation de la base de données BDSON dépendra principalement de la qualité de la segmentation et de l'étiquetage qui seront appliqués sur ces corpus. Le GRECO a choisi de disposer à la fois d'un étiquetage dit "large", utile pour l'interrogation de la base et le repérage des réalisations en contexte phonémique large, et d'un étiquetage dit "fin" destiné plus particulièrement à l'étude fondamentale des sons du français.

Un groupe de travail a défini les principes de la segmentation et de l'étiquetage [3] qui seront appliqués. L'étiquetage "large" consiste à repérer et à étiqueter les centres des réalisations de segments de taille phonémique alors que l'étiquetage "fin" correspond à un repérage d'événements qui traduisent des discontinuités majeures apparaissant sur le signal de parole et/ou sur son spectre. Pour l'étiquetage "fin" deux méthodes complémentaires sont mises en oeuvre :

- la première méthode, entièrement manuelle, se situe dans le domaine temporel et opère directement sur l'onde acoustique. A cet étiquetage fin temporel sera superposée ensuite une structure de macro-classes à plusieurs niveaux.
- la seconde qui est une méthode semi-automatique, procède à partir d'une représentation fréquentielle.

L'étiquetage "large", quant à lui, sera mené de façon entièrement automatique, mais l'évaluation des diverses méthodes possibles et le choix de l'une d'elles reste à faire. On évalue à environ 1 homme-année le temps nécessaire à l'étiquetage "fin" des groupes CVCV du corpus acoustique seulement.

III. STRUCTURE DE LA BDSON.

Etant donné le grand volume de données à gérer dans la BDSON et afin de suivre une démarche modulaire de conception, nous réalisons la BDSON en deux étapes qui correspondent aussi à deux niveaux de gestion des fichiers-sons.

Dans un premier temps, le Système de Gestion de Bases de Données (SGBD) gèrera les descripteurs des fichiers contenant la parole digitalisée. Les fichiers-sons eux-mêmes seront stockés hors de la base sur des supports comme les bandes magnétiques et les cassettes PCM. Puis, dans une deuxième étape, lorsque l'étiquetage "large" et "fin" auront été effectués, nous fournirons à l'utilisateur des outils pour la gestion du contenu des fichiers-sons (répérer des mots, des phonèmes, des phases acoustiques, etc) ainsi que les résultats des traitements associés (spectres, énergie, etc.).

III.1 Gestion du catalogue de la BDSON.

Dans ce premier niveau nous gérons les données décrivant les locuteurs, les corpus et leur contenu ainsi que les réalisations (un LOCUTEUR l prononce un CORPUS c et produit une REALISATION r qui contient e ELEMENTS). Il s'agit donc d'un catalogue qui répertorie tous les corpus enregistrés et disponibles dans la BDSON.

Ces informations permettront à l'utilisateur de connaître le type, la quantité et l'emplacement des réalisations qui sont disponibles dans la base.

Ce niveau de la BDSON est défini à l'aide d'un SGBD Relationnel qui grâce à sa simplicité, sa généralité et ses possibilités d'extension, facilite la définition et la manipulation des données [4]. Il permet ainsi, l'archivage précis et la consultation des fichiers-sons.

Schéma conceptuel

Pour cette première étape, nous présentons les entités de base qui composent le schéma conceptuel de la BDSON, ainsi que leurs attributs associés :

CORPUS (code-corpus, titre, type-corpus, nb-élém, type-élém, fréq-échant)

LOCUTEUR (code-locuteur, nom, sexe, âge, caract-linguist, adresse, tel)

Les détails de l'enregistrement des corpus par les locuteurs figurent dans la relation :

REALISATION (code-corpus, code-locuteur, répétition, support, no-unité, longueur, durée, date, lieu)

Le contenu de chaque corpus sous la forme orthographique et phonétique est décrit élément par élément dans la relation :

CONT-CORPUS (code-corpus, no-élém,
desc-ortho, desc-phoné)

Et chacun des ces éléments prononcés par chaque locuteur est décrit et répéré dans la relation :

ELEMENT (code-corpus, code-locuteur,
no-élém, pos-ds-réalisation)

Interrogation de la BDSON

L'objectif principal à ce premier niveau de gestion de la BDSON est de permettre à l'utilisateur de s'informer sur le contenu de la base ainsi que sur l'emplacement des données auxquelles il s'intéresse. Pour ceci, nous avons conçu un menu qui met à sa disposition les possibilités de consultation dans la BDSON. L'utilisateur pourra ainsi poser ses requêtes en allant du général jusqu'au détail dans la sélection des corpus, des locuteurs et/ou des réalisations.

Les requêtes peuvent porter d'abord sur toutes ou chacune des caractéristiques des entités : corpus, locuteur et réalisation. Dans le cas de CORPUS par exemple, il pourrait obtenir la liste des corpus existants, en les sélectionnant par :

- type de corpus (évaluation, acoustique, préliminaire) et/ou
- type des éléments qu'il contient (mots, phrases, texte) et/ou
- code identificateur (Ex. SYL*) et/ou
- titre du corpus

En ce qui concerne les LOCUTEURS, il pourrait obtenir la liste des locuteurs sélectionnés par :

- code-locuteur et/ou
- nom du locuteur et/ou
- sexe et/ou
- âge et/ou
- caractéristiques linguistiques

Pour les REALISATIONS, l'utilisateur pourrait les sélectionner en choisissant les caractéristiques des CORPUS prononcés et des LOCUTEURS qui les ont prononcés.

Lorsque l'utilisateur aura choisi les réalisations qui l'intéressent, il pourra obtenir une liste avec des informations associées : type-support, no-unité, et l'emplacement où elles se trouvent. Il pourrait aussi repérer certains éléments en particulier, contenus dans une ou plusieurs réalisations. Dans ce cas, l'utilisateur pourra sélectionner les éléments qui l'intéressent après avoir obtenu les corpus qui les contiennent et les locuteurs qui les ont prononcés.

Pour toutes les requêtes, l'utilisateur peut décider d'obtenir la sortie soit sur l'écran, soit sur l'imprimante. Le menu présenté à l'utilisateur est développé avec l'interface programmable du SGBD et chacun de ses choix est traduit par un ensemble de requêtes exprimées dans le langage de manipulation de données.

Mise à jour

La structure de la BDSON est dynamique. Il est toujours possible de l'enrichir et de la modifier selon les besoins de ses utilisateurs.

Il existe donc des procédures qui effectuent la mise à jour : ajout de nouvelles données, modification et/ou effacement de celles déjà existantes. Par exemple : ajouter de nouveaux locuteurs, de nouveaux corpus, et/ou de nouvelles réalisations. Afin de maintenir la cohérence des données gérées par la BDSON, nous exécutons des procédures de vérification des contraintes d'intégrité. Ceci évite par exemple d'ajouter des données correspondant à une nouvelle réalisation pour laquelle la BDSON ne gère pas le corpus prononcé et/ou le locuteur qui l'a prononcé.

On est aussi capable de créer de nouveaux types de données et/ou d'éliminer certains déjà existants. Ceci est utile pour l'ajout de nouvelles caractéristiques aux entités gérées par la BDSON. Par exemple, ajouter d'autres paramètres d'enregistrement associés à chaque réalisation.

Dans tous les cas, toutes les opérations de mise à jour sont réservées à l'administrateur de la base. Chaque laboratoire pourra avoir sa propre version de ce niveau de la BDSON, mais il ne devra pas la modifier. Il ne pourra que la consulter et accéder postérieurement aux réalisations cherchées, qui existent sur le support correspondant.

Cette décision s'appuie sur la nécessité de garantir la cohérence de l'agrandissement et de l'enrichissement de la base. Ainsi, toutes les modifications à réaliser sur la BDSON se feront dans le centre serveur, qui aura la responsabilité de la distribution aux autres laboratoires des versions actualisées.

III.2. Gestion du contenu des fichiers-sons et des traitements associés

Dans ce deuxième niveau, nous accéderons au contenu des fichiers-sons. Nous visons à la manipulation des portions du signal répérées par des étiquettes.

Ces étiquettes suivent les principes définis par le groupe de travail d'étiquetage du GRECO. Ce groupe de travail a défini un code de représentation standard pour l'étiquetage phonétique dit "large". Par contre, pour les deux méthodes complémentaires mises en oeuvre dans l'étiquetage "fin", chacune possède son propre code de représentation. La BDSON devra gérer ces différents types d'étiquetage et permettre la construction d'autres corpus à partir de segments du signal extraits à l'aide des étiquettes qui les définissent.

En plus de la gestion du contenu des fichiers-sons et de l'étiquetage correspondant, nous envisageons la gestion des résultats d'autres traitements appliqués à chaque réalisation. Par exemple, le spectre, les formants, les niveaux d'énergie, etc.

Certaines requêtes des utilisateurs peuvent se traduire par des accès et des opérations complexes sur les sons. Pour simplifier la manipulation de données, un ensemble d'utilitaires développés autour de la BDSON prendra en charge ces opérations. De plus d'autres utilitaires seront employés pour le traitement de sons qui tireront ainsi partie de la présence des informations dans la base. L'ensemble de ces outils constituera donc un Environnement de Gestion de Sons permettant à l'utilisateur d'accéder simplement aux sons contenus dans les réalisations et à toutes leurs informations associées.

L'ensemble des informations, que l'on peut juger dès maintenant nécessaire de gérer dans la BDSON, est donc constitué par :

- les descripteurs du contenu et de l'emplacement sur les supports des fichiers-sons,
- les fichiers-sons eux-mêmes,
- l'étiquetage associé,
- les résultats d'autres traitements, associés aux réalisations.

En considérant la capacité de stockage (de l'ordre de 4 Gigaoctets) nécessaire pour avoir l'accès direct aux fichiers-sons, il est impératif de prévoir l'utilisation d'un support de type Gigadisque au niveau du centre serveur. Toutes les autres données de la BDSON resteront sur support magnétique. La diffusion de la BDSON sera faite par cassettes video-PCM pour les réalisations des corpus et par disquettes magnétiques pour les autres informations associées.

IV. ETAT D'AVANCEMENT

Actuellement, la BDSON gère 87 corpus, avec 32 locuteurs et environ 1800 réalisations. Ce premier niveau de gestion de la BDSON est developpé avec un volume de stockage d'environ 3 Méga-octets. Quelques problèmes de performance se posent lors de l'accès simultané à plusieurs relations. Spécialement avec des recherches qui portent sur les relations CONT-CORPUS (environ 3500 tuples) et ELEMENT (environ 75000 tuples) en combinaison avec CORPUS et/ou LOCUTEUR.

Une première maquette de la BDSON a été définie avec l'aide du SGBD relationnel MICROBE*. Le premier niveau de la BDSONS est implanté aussi sur un micro-ordinateur compatible IBM-PC de l'ENSERG en utilisant le SGBD relationnel KMAN/2**. Cette version facilitera sa diffusion parmi les laboratoires du GRECO.

Nous travaillons sur la définition et la réalisation de critères plus précis d'évaluation qui nous permettront de décider si le SGBD utilisé actuellement satisfait nos besoins pour le développement de la deuxième étape.

* MICROBE : SGBD développé par G.T.Nguyen LGI-IMAG
** KMAN/2 : SGBD commercialisé par ISE-CEGOS.

Il nous reste encore beaucoup de détails à préciser pour la gestion du deuxième niveau de la BDSON. Nous comptons sur les expériences de la première étape pour afiner l'évaluation des performances du SGBD Relationnel et des moyens matériels de stockage et de visualisation à mettre en oeuvre pour répondre efficacement aux besoins des utilisateurs.

CONCLUSION

La BDSON est le résultat d'un vaste consensus des ingénieurs et des phonéticiens français. De nombreuses réunions de concertation ont abouti à un accord sur la méthodologie adoptée pour définir les caractéristiques des enregistrements, le choix des locuteurs et des corpus, l'étiquetage et la segmentation. Un premier niveau de gestion de la BDSON mis un oeuvre à l'aide d'un SGBD relationnel est déjà opérationnel. La deuxième phase du développement de la BDSON commence : étiquetage "large" et "fin", distribution à la demande de corpus particuliers, intégration des résultats d'analyse.

Par ailleurs, il a été proposé d'interconnecter BDSON et BDLEX (également soutenue par le GRECO Parole). Ainsi sera couvert un champ important : "du son jusqu'aux mots".

REFERENCES BIBLIOGRAPHIQUES

[1] PECKELS J.P., ROSSI M., 1971, "Le test de diagnostique par paires minimales", 2èmes Journées d'Etude sur la Parole, Aix-en-Provence.

[2] COMBESCURE P., 1981, "20 listes de 10 phrases phonétiquement équilibrées", Revue d'Acoustique n. 56, pp. 34-38

[3] ABRY C. et al., 1985, "Propositions pour la segmentation et l'étiquetage d'une base de données des sons du français", 14èmes Journées d'Etude sur la Parole, Paris, pp. 156-163.

[4] DELOBEL C., ADIBA M., 1982, "Bases de données et systèmes relationnels", DUNOD Informatique.

RECONNAISSANCE

Présidents

Jean-Sylvain LIENARD

L.I.M.S.I. d'Orsay

Jean-Pierre TUBACH

E.N.S.T. de Paris

RECONNAISSANCE DE GRANDS VOCABULAIRES: UTILISATION ET EVALUATION DE TRAITS GROSSIERS.

G.Adda, M. Eskénazi, P.E. Stern

LIMSI-CNRS, B.P. 30, 91406 Orsay Cedex, France

Abstract

Within the framework of recognition of very large vocabularies, the use of rough spectral features has proven useful for English. Their use allows for a significant reduction of number of words in a vocabulary to be dealt with during the recognition phase.

We present an example, for French, of the use of rough spectral feature on a very large (> 15,000-word) vocabulary. We also evaluate the latter by comparing results for two differents French lexicons - one of 267,000 forms, and the other of 17,000 - with studies carried out on English vocabularies.

The evaluation shows that the reduction we obtain is comparable to the ones for English considering our use of several different possible pronunciation of each word.

I. Introduction

Des chercheurs expérimentés dans la lecture de spectrogrammes et dans l'analyse acoustifque de la parole ont montré l'existence de certains faits acoustiques qui, indépendamment du locuteur, sont détectables de manière très fiable. Nous appellerons ces faits acoustiques **traits grossiers.**

L'utilisation de traits grossiers dans le cadre d'un système de reconnaissance en mots isolés de grands vocabulaires nous semble intéressante. Ce système comprendrait une première étape de reconnaissance de traits grossiers permettant d'accéder à un sous-ensemble du vocabulaire, suivie d'une analyse plus fine afin de déterminer le mot exact à l'intérieur de ce sous-ensemble. Il faut donc d'une part, que la détection de ces traits grossiers conduise de manière sûre à un sous-vocabulaire qui contienne le mot prononcé et d'autre part que celui-là soit d'une taille suffisamment limitée pour que la tâche de la reconnaissance fine soit réalisable.

Diverses études ont montré l'intérêt de cette approche sur la langue anglaise [1][2][3]. Pour la langue française, des études ont déjà été menées sur le langage écrit [4], et sur un vocabulaire limité (1000 mots) [5].

Nous présentons ici une évaluation de l'utilisation de traits grossiers pour la langue française dans le cadre du système de reconnaissance de grands vocabulaires (> 15000 mots) que nous voulons réaliser.

II. Utilisation de traits grossiers

Nous avons effectué un certain nombre de choix préliminaires:

- des traits grossiers, i.e. des faits acoustiques que nous détecterons au cours d'une première étape de l'analyse,
- du système de codage du vocabulaire en fonction de ces traits grossiers, i.e. le choix de la méthode d'accès à un sous-vocabulaire.

A. Définition des traits grossiers

Le premier critère de choix pour ces traits a été la fiabilité de leur détection dans le signal. Ceci nous a conduit, contrairement à d'autres études où ces traits correspondent à des classes phonétiques ou à des notions linguistiques (CV), à choisir des traits liés essentiellement au signal acoustique. Nous avons écarté des traits couramment employés, comme nasal, car il ne nous semble pas possible de déterminer actuellement un tel trait avec un taux d'erreur de moins de 5%.

Dans un premier temps, nous avions retenu 5 traits: **Silence, Silence-voisé, Bruit, Bruit-voisé, Vocalique** mais une évaluation préliminaire de ces traits nous a poussé à effectuer une partition de la classe vocalique en deux classes: **non(A), non(I). non(A)** (resp. **non(I)**) représente les traits vocaliques qui à coup sûr ne seront pas reconnus comme le phonème /a/ (resp. /i/).

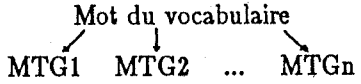
B. Définition des sous-vocabulaires

Dans la phase de reconnaissance, le mot prononcé est décodé en une suite de traits grossiers par un système de détection de traits. Le mot composé de traits grossiers, (noté **MTG**), doit permettre d'accéder à un sous-vocabulaire contenant le mot prononcé.

Dans une phase préliminaire à la reconnaissance, il est nécessaire de réaliser une partition du vocabulaire, chaque partie du vocabulaire étant représenté par un MTG. On appelle **cohortes** les sous-vocabulaires représentés par des MTG.

La taille des vocabulaires utilisés dans ce type d'application étant très grande (> 15000 mots), la détermination des cohortes, à partir d'un vocabulaire donné, doit se faire de manière automatique.

On doit donc, à partir d'un mot du vocabulaire phonétisé par le programme de conversion graphème-phonème GRAPHON, pouvoir prédire les différents MTG susceptibles d'être détectés dans le signal correspondant.



Obtention d'un MTG à partir d'un mot du vocabulaire

Nous avons choisi d'avoir la possibilité de réécrire chaque mot en plusieurs MTG [5], afin de tenir compte de variantes phonologiques, de différentes prononciations et des aléas de la détection des traits dans certains contextes. Aussi, nous utilisons des règles de réécriture, qui fournissent, pour une même prémisse, une ou plusieurs suites de traits grossiers.

L'idée de base pourrait être de choisir comme prémisse chaque phonème. Cependant, nous avons décidé d'utiliser comme prémisses des contextes plus longs, afin à la fois de limiter le nombre de suites possibles de traits grossiers générées, et surtout pour faciliter la réécriture des phonèmes qui sont fortement influencés par leur contexte (en particulier /R/). L'unité contextuelle retenue, dont la taille correspond à celle de la syllabe, commence soit avec le début du mot, soit avec une voyelle et se termine à la prochaine voyelle ou à la fin du mot. Cette définition permet, étant donné les traits grossiers choisis, de rendre mieux compte des variations possibles, qui sont essentiellement dues aux consonnes. Une liste de toutes les prémisses(pseudo-syllabes) possibles, a été dressée à partir du vocabulaire de test. Il faut ajouter que les conclusions des règles de réécriture ont été obtenues à partir de connaissances formalisées dans le système expert en lecture de spectrogrammes SONEX [6].

Afin de limiter le nombre de prémisses, les phonèmes ayant le même comportement, vis-à-vis des règles contextuelles, ont été regroupés en classes phonétiques. Ces classes sont: occlusive sourde, occlusive sonore, occlusive nasale, /f/, /s/, /ʃ/, fricative sonore, /R/, /l/, /a/, /i/ et voyelles autres que /a/ et /i/. Les prémisses des règles sont donc constituées de ces classes phonétiques.

Les différents MTG sont obtenus à partir d'un mot du vocabulaire, par concaténation de toutes les conclusions des règles déclenchées.

Nous avons, de plus, utilisé la notion de **compactage**. En effet, prenons l'exemple du mot *acte* (/akt/). Pour ce mot l'explosion du /k/ peut ne pas apparaître. La réécriture de ce mot en MTG génére un certain nombre de possibilités:

/akt/ → non(I).Silence.Bruit.Silence.Bruit

...

non(I).Silence.Silence.Bruit

Cette dernière réécriture se compacte en: **non(I).Silence.Bruit.**

En effet, il est peu réaliste de vouloir détecter deux silences successifs. D'une manière générale, nous avons compacté toute suite de traits identiques, à l'intérieur d'un MTG, en un seul trait. Cette action de compactage contribue à rendre le traitement indépendant des problèmes de segmentation.

III. Evaluation

Nous avons effectué notre étude préliminaire sur deux vocabulaires.

- Un dictionnaire (DICO1) de 267291 formes orthographiques, correspondant à 136305 formes phonétiques et 69453 mots composés des classes phonétiques utilisées dans les règles de réécriture. Ce dictionnaire est dérivé de celui utilisé dans le logiciel d'intérrogation de bases de données SPIRIT [7] .
- Un dictionnaire (DICO2) de 17040 formes orthographiques, correspondant à 10083 formes phonétiques et 6663 mots en classes phonétiques . Ce dictionnaire a été constitué à partir des formes les plus fréquentes de la langue française écrite, et a été développé au LIMSI [8] .

Les résultats obtenus sur DICO1 et DICO2 sont présentés sur les figures 1 et 2 . La figure 1 représente, pour chaque taille possible de cohorte, le nombre de cohortes qui ont une telle taille. On utilise ici les dictionnaires sous leur forme phonétique. Les échelles sont représentées en logarithme à base 2.

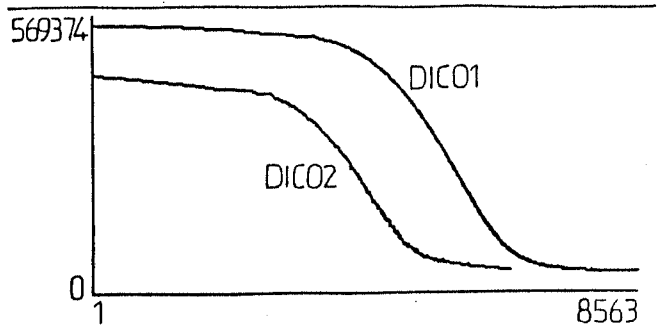


Fig 1: Nombre de cohortes(en log_2) en fonction de leur taille(en log_2).

La taille attendue d'une cohorte, c'est-à-dire la somme de toutes les tailles divisée par le nombre total de cohortes (1.390000 pour DICO1 et 83000 pour DICO2), n'est pas une mesure significative pour un mot du vocabulaire de la taille moyenne d'une cohorte qui le contient. Nous utilisons une notion statistique, **la taille attendue**, dont la définition [2] est:

$$T_{attendue} = \sum_{i \in E} S_i P_i$$

où E est l'ensemble des cohortes,
S_i représente la taille de la cohorte i,
et P_i est la probabilité pour un mot d'appartenir à la cohorte i.
Nous utiliserons ici

$$P_i = \frac{S_i}{\sum_{i \in E} S_i}$$

Cette notion est une mesure plus réelle que la taille moyenne, du facteur de réduction du vocabulaire.

Taille	maximale	moyenne	attendue
DICO1 (orthographique)	17234	13	1056
DICO1 (phonétique)	8563	8	510
DICO2 (orthographique)	2023	6	200
DICO2 (phonétique)	1063	4	108

Fig 2: Tailles attendue, moyenne et maximale pour DICO1 et DICO2.

On obtient, pour DICO1(phonétique), sur un ensemble de cinq traits grossiers (cf II.A), une taille attendue de 2100 et une taille maximale de 12000. La différence entre ces résultats et ceux présentés figure 2 montrent l'intérêt d'utiliser deux traits pour la classe vocalique.

Des études sur les traits grossiers ont déjà été menées pour la langue anglaise sur de grands vocabulaires(20000 mots), et il semble nécessaire de les confronter avec nos résultats obtenus sur le français.

En anglais, pour un vocabulaire de 20000 mots et un ensemble de six traits grossiers, on obtient [1][3] une taille attendue de 20 et une taille maximale de 200 . Le facteur de réduction de la taille du vocabulaire semble a priori en français moins intéressant. Cependant, contrairement aux études antérieures, nous avons choisi d'effectuer plusieurs réécritures en MTG pour un même mot du vocabulaire et de compacter les traits, afin de prévoir tous les cas de reconnaissance, et en particulier le pire. Pour comparer effectivement nos résultats avec ceux de l'anglais, on doit évaluer l'impact de ces choix.

La figure 3 permet de matérialiser la différence entre notre méthode de réécriture multiple et une méthode de réécriture unique.

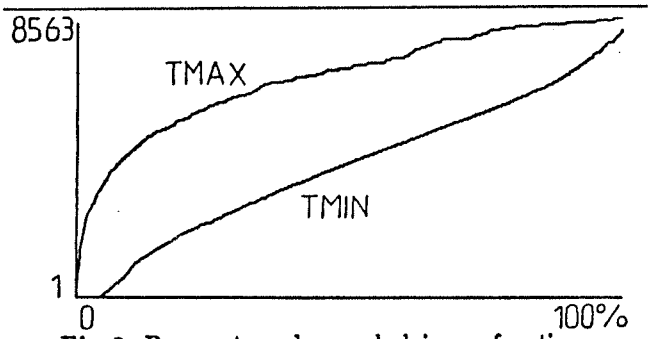


Fig 3: Pourcentage du vocabulaire en fonction de la taille des cohortes(en log_2).

La courbe TMIN (resp. TMAX) représente le cas le plus favorable (resp. le plus défavorable) de reconnaissance, c'est-à-dire que pour chaque mot du vocabulaire, on aboutit à la cohorte qui contient ce mot et qui a la taille la plus petite (resp. la plus grande). On montre ici le pourcentage du vocabulaire DICO1 que l'on peut reconnaître dans le meilleur des cas (TMIN) et dans le pire (TMAX), en n'utilisant que des cohortes d'une taille inférieure à un seuil. Ainsi, on peut voir que, pour 50% du vocabulaire, on obtient sur TMIN une taille maximale de 82 et pour TMAX une taille maximale de 1881, soit un facteur de réduction de 23 en faveur de TMIN. De même, on obtient pour 90% du vocabulaire, pour TMIN une taille de 1452, et pour TMAX une taille de 7549, soit un facteur de réduction de 5,2.

Nous avons constaté de plus une très nette diminution des tailles des cohortes en fonction du nombre de traits composant les MTG. Ceci confirme l'idée intuitive que, pour un mot donné, la taille de la cohorte correspondante sera d'autant plus petite que le nombre de traits grossiers

détectés sera important.

Une réécriture unique, correspond en fait à une transcription en une suite de classes phonétiques. Ceci revient, pour nos traits grossiers, à réécrire un mot à l'aide du plus grand nombre de traits possible, i.e. du MTG le plus long.

Dans l'hypothèse de réécriture unique, on se place délibérément sur la courbe TMIN, alors que pour une réécriture multiple, on se trouve entre TMIN et TMAX. Les résultats présentés sont donc comparables aux résultats pour l'anglais, la courbe TMIN conduisant à une taille maximale de 2800 pour un vocabulaire de 136000 mots. Il faut de plus ajouter que nous n'avons pas tenu compte ici du compactage. Celui-ci a pour effet d'accentuer les disparités entre cohortes dans la mesure où il augmente les tailles des cohortes déjà grande(avant compactage) et laisse les autres inchangées.

Il ne suffit pas de connaître la taille de chaque cohorte pour savoir, en phase de reconnaissance, quelle sera la taille de la cohorte qui sera trouvée, car ceci dépend du vocabulaire de test utilisé. En se plaçant dans le cadre d'une dictée vocale, on peut faire l'hypothèse que les mots les plus employés dans le vocabulaire de test seront ceux du français écrit, et ceux-là sont les mots les plus courts. Donc, en donnant, pour chaque longueur possible de mot phonétique les courbes TMIN et TMAX de DICO1, on peut voir la taille des cohortes qui seront les plus utilisées (fig 4).

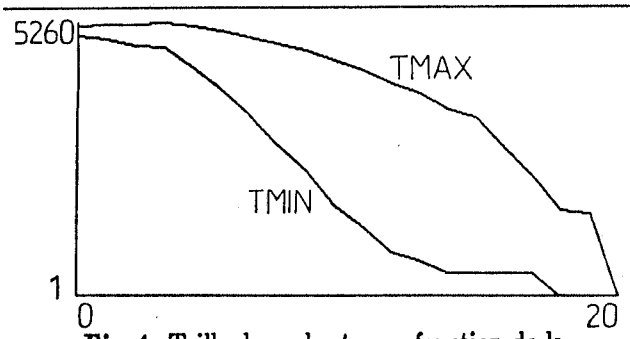


Fig 4: Taille des cohortes en fonction de la longueur du mot phonétique.

On peut noter que pour les mots d'une taille comprise entre deux et quatre phonèmes (et qui ont a priori une fréquence lexicale élevée), la courbe TMAX présente un maximum . La pondération par la fréquence lexicale conduirait à une dégradation des résultats, et en particulier ceux des tailles attendues . Ce phénomène a également été remarqué pour l'anglais [3] .

IV. Conclusion

L'évaluation de l'utilisation de traits grossiers pour la partition de grands vocabulaires n'est qu'une première étape. A l'issue de celle-ci, nous disposons d'un outil qui génère automatiquement la partition d'un vocabulaire donné en cohortes . De plus l'utilisation de règles de réécriture nous permet, grâce à leur forme déclarative, de nous adapter facilement à une évolution des traits grossiers (changement de contexte, ou ajout d'autres traits) .

Les conclusions de l'évaluation ont pu montré que, en comparaison des études faites pour l'anglais, l'utilisation de traits grossiers en français peut amener une forte diminution de la taille du vocabulaire. Cependant, les résultats semblent globalement plus défavorables pour le français que pour l'anglais, en particulier à cause des réécritures multiples et du compactage . Notons que l'on pourrait diminuer l'influence de celui-ci en introduisant une notion de durée sur les traits . Par exemple, du fait du compactage, la cohorte **non(I)** contient les deux mots du vocabulaires *a* et *araméen*, qui sont pourtant facilement discernables par leur durée .

Bibliographie

[1] **D.W. Shipman, V.W Zue,** "Properties of large lexicons. Implications for advanced isolated word recognition systems", *Proc. IEEE ICASSP*, Paris ,1982.

[2] **A. Waibel,** "Towards very large vocabulary word recognition", Tech. report 144, CMU, 1982.

[3] **D.P. Huttenlocher,** "Acoustic-phonetic and lexical constraints in word recognition: lexical access using partial information", Master thesis, MIT, 1984.

[4] **R. Carlson et al.,** "Phonetic and orthographic properties of the basic vocabulary of five european languages", *Symposium franco-suédois*, Grenoble, 1985.

[5] **J.P. Haton,** "Accès lexical et reconnaissance de grands vocabulaires", *Journées d'Etudes sur la Parole*, Bruxelles, 1984.

[6] **P.E. Stern, M. Eskenazi, D. Memmi,** "An expert system for speech spectrogram reading", *Proc. IEEE ICASSP*, Tokyo, 1986.

[7] **A.Andreewsky et al.,** "Les dictionnaires en formes et leur utilisation dans la transformation lexicale et syntaxique correcte de chaînes phonétiques", *Journées d'Etudes sur la Parole*, Grenoble, 1979.

[8] **J.L. Gauvain,** "A syllable-based isolated word recognition experiment", *Proc. IEEE ICASSP*, Tokyo, 1986.

Reconnaissance multilocuteur de mots isolés fondée sur une approche phonétique

A.MESSAOUDI, J.L.GAUVAIN

LIMSI/CNRS BP.30 91406 ORSAY Cedex

Résumé

Dans cet article nous présentons d'une part des expériences portant sur le choix des paramètres retenus après l'étage d'analyse du signal, et d'autre part un système de reconnaissance multilocuteur de mots isolés fondé sur une approche phonétique. Chaque phonème est représenté par un dictionnaire de spectres obtenu par classification automatique sur un corpus multilocuteur. La phase d'apprentissage se résume simplement par l'introduction de la liste du vocabulaire sous forme phonétique ou graphémique. A chaque mot est associé un modèle de production que nous représentons par un automate fini.

Introduction

L'approche globale s'appuyant sur des méthodes statistiques (multi-référence [1] ou modèle Markovien [2]) est une solution très efficace au problème de la reconnaissance multilocuteur de mots isolés. Toutefois, pour chaque nouvelle application, cette approche nécessite l'enregistrement de l'ensemble du vocabulaire par une centaine de locuteurs. L'approche phonétique s'appuyant sur des méthodes statiques ou sur des traits multilocuteur permet d'éviter ce problème [3], [4]. Nous utilisons le phone comme unité de décision; ce choix se justifie par le faible nombre de phonèmes qu'il y a dans chaque langue. A chaque phonème est associé un dictionnaire de spectres correspondant aux différentes réalisations de ce phonème. Ces dictionnaires sont construits à partir d'un corpus, multilocuteur, de 136 mots contenant 34 phonèmes (selon le code LIMSI [9]) dans plusieurs contextes. La segmentation du corpus a été faite manuellement à l'aide d'un éditeur de spectrogrammes numériques. Un algorithme de classification automatique (k-moyennes [10]) est appliqué sur l'ensemble des spectres d'un phonème pour en extraire un sous-ensemble limité, reflétant les différentes réalisations spectrales de ce phonème. La phase d'apprentissage se fait simplement par la description phonétique ou graphémique de chaque mot du vocabulaire. Le système génère pour chaque mot un modèle de production tenant compte des durées moyennes des phonèmes et des variantes phonologiques du mot. La phase de reconnaissance se fait en deux étapes. La première étape permet de calculer les distances phonétiques de chaque échantillon de parole. La deuxième étape consiste à calculer la distance globale du mot inconnu à chaque modèle de référence.

Analyse et paramétrisation du signal

Analyse du signal

La méthode d'analyse du signal employée dans le cadre de notre travail, est une analyse spectrale par banc de 16 filtres répartis selon une échelle de Bark. Ces filtres sont simulés par une FFT calculant un spectre tous les 12,8 ms, d'un segment de signal de 25,6 ms pondéré par une fenêtre de Hamming. Après cet étage d'analyse, un mot est représenté par une succession d'échantillons représentés par des vecteurs d'énergies exprimées en dB dans les 16 canaux.

Paramétrisation

Au lieu de considérer les vecteurs d'énergies, on utilise couramment d'autres paramètres. Nous avons utilisé quatre types de paramètres dans nos expériences.

- "Pentes" : 15 paramètres obtenus par la différence d'énergie entre deux canaux successifs.
- Energies normalisées : 16 paramètres obtenus par soustraction de la valeur moyenne.
- Coefficients cepstraux : 1 à 15 paramètres obtenus par une transformation en cosinus sur le spectre.
- Indices acoustiques [4]:
  5 indices (Grave/Aigu, Fermé/Ouvert, Doux/Strident,

Bémolisé/Diésé, Compact/Ecarté) simulés sur les 16 canaux dont les fréquences centrales et largeurs de bandes sont légèrement différentes de celles des 24 canaux utilisés par Caelen [4].

La distance que nous avons utilisée sur ces paramètres est la distance de Minkovski d'ordre 1 (distance $L1$).

type de paramètres	*pentes*	*énergies normalisées*	*cepstres (7)*	*indices*
taux(%)	77,5	92,5	94	91

Tableau I: Taux de reconnaissance en fonction des quatres types de paramètre.

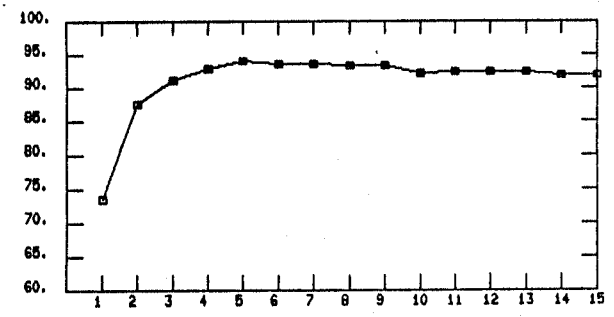


Figure 1: Taux de reconnaissance en fonction du nombre de coefficients cepstraux.

Choix des paramètres

Afin de choisir des paramètres robustes et pertinents, nous avons réalisé des tests de reconnaissance inter-locuteur de mots isolés. Le corpus employé pour ces tests est constitué de deux répétitions des 10 chiffres par 20 locuteurs: 10 locuteurs de référence et 10 autres de test. Une répétition des 10 premiers locuteurs est utilisée comme référence, et les deux répétitions des 10 autres locuteurs servent de corpus de test. Les résultats obtenus sont donnés dans le tableau I. On peut remarquer que les "pentes" donnent de mauvais résultats en reconnaissance inter-locuteur; (plus particulièrement nous avons constaté que le taux de reconnaissance est de l'ordre de 60% quand les locuteurs de test et de référence sont de sexe différents). Par contre avec 7 coefficients cepstraux nous avons obtenu le meilleur taux de reconnaissance. Ceci confirme les résultats obtenus en monolocuteur par C.Gagnoulet et M.Couvrat sur des vocabulaires difficiles [5]. Nous nous sommes alors intéressés à l'influence du nombre de coefficients sur le taux de reconnaissance. Sur la courbe de la figure 1 nous remarquons que 4 à 8 coefficients suffisent pour caractériser le signal. Dans la suite de notre travail, nous avons choisi les coefficients cepstraux comme paramètres de codage des échantillons du signal.

Présentation du système

Construction du dictionnaire

Afin de constituer un dictionnaire de formes spectrales pour chaque phonème, nous avons construit un corpus de mots isolés prononcés par 18 locuteurs (9H et 9F). Le vocabulaire utilisé est constitué de 136 mots, extraits d'un dictionnaire de 10400 mots (formes phonétiques) [6], contenant tous les phonèmes dans des contextes différents. Nous avons segmenté et étiqueté manuellement le corpus à l'aide d'un éditeur de spectrogrammes. Nous envisageons une segmentation semi-automatique afin d'étendre le corpus à une large population. Un segment phonétique comprend l'ensemble des spectres situés entre deux transitions. Nous avons donc placé les marqueurs de début et de fin de segment aux frontières des deux parties transitoires. Pour les occlusives sourdes, nous n'avons tenu compte que de l'explosion. Le nombre de segments étiquetés est de l'ordre de 10800, ce qui donne une moyenne de l'ordre de 350 segments par phonème et un nombre de spectres de l'ordre de 3500 par phonème. Pour réduire cette quantitée d'information, nous avons appliqué l'algorithme des K-moyennes [10] qui permet une classification en un nombre fixé de classes. Une classe est représentée par son centroide, i.e. l'élément dont la distance moyenne à tous les autres éléments de la classe est minimisée. Le nombre de classes par phonème a été fixé, pour la suite de nos expériences, à la valeur 64 (soit au total 2200 spectres de référence qui peuvent éventuellement être traités en temps réel par un processeur de programmation dynamique [7])

Représentation des mots

Un mot de référence est décrit par sa transcription phonétique. L'utilisateur introduit la liste des mots sous forme graphémique qui est traduite par le module de traduction graphèmes-phonèmes [9]. Afin d'aider le module de comparaison dynamique à trouver le chemin optimal, un mot est représenté par un automate fini. La figure 2 montre l'exemple du mot /ALFA/. A l'état i, correspondant à l'émission d'un phonème, on associe un paramètre

(a_i) qui est la durée moyenne de ce phonème. En pratique nous avons affecté à ce paramètre la moitié de la durée moyenne du phonème associé, calculée sur tous les segments d'apprentissage. La transition de l'état i à l'état suivant, ne peut être réalisée que si le système est resté au moins (a_i) fois dans l'état i. Les occlusives sourdes sont représentées par la succession de deux états; le premier est associé au silence, le deuxième est associé à l'explosion. Pour tenir compte du E muet prononcé par certains locuteurs, nous avons ajouté un état après les consonnes finales et la semi-voyelle /j/. Enfin, nous envisageons de représenter un mot par un modèle "complet" tenant compte de ses principales variantes phonologiques [11].

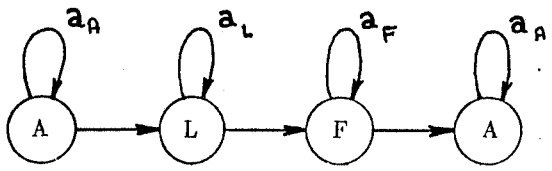


Figure 2: Modèle associé au mot /ALFA/.

Phase de reconnaissance

Pour éviter les problèmes posés par la segmentation phonétique, nous utilisons un modèle centiseconde pour représenter les énoncés à reconnaître. Durant l'étape de reconnaissance, une distance globale est évaluée entre le mot inconnu et chacun des modèles de production. La solution est le mot correspondant au modèle qui minimise cette distance. Le calcul de la distance globale entre un modèle de production d'un mot et la suite des éléments centisecondes de l'énoncé à identifier se décompose en deux phases. Durant la première phase, le système évalue les distances locales entre les éléments centisecondes et les états du modèle de production. La distance locale entre éléments centisecondes et états du modèle est égale à la plus petite distance spectrale entre cet élément centiseconde et les spectres du dictionnaire associé à l'état. Durant la seconde phase, la distance globale est évaluée par un algorithme de programmation dynamique [8] qui recherche la succession d'états du modèle qui minimise la somme des distances cumulées de l'état initial à l'état final du modèle.

Expériences et résultats

Les premières expériences ont été réalisées en mode monolocuteur sur un corpus de 660 syllabes contenant tous les phonèmes dans des contextes différents. Nous avons extrait de ce corpus 1880 segments phonétiques. L'algorithme des K-moyennes a été appliqué sur l'ensemble des spectres de chaque phonème. Le nombre de spectres retenu, par phonème, est de 10. Le représentant d'une classe est donné par son centroide. Les résultats obtenus en autocohérence sur les segments phonétiques d'apprentissage sont de 78% sur le premier candidat et de l'ordre de 93% sur les trois premiers. Le taux de reconnaissance sur des vocabulaires de 20 syllabes prononcés par le même locuteur est de l'ordre de 98%.

Pour les expériences en mode multilocuteur, nous avons effectué des séries de tests sur deux vocabulaires de tailles différentes.

Le premier vocabulaire est celui qui a servi à la constitution du dictionnaire de 136 mots. Nous avons enregistré 11 locuteurs pour ce vocabulaire. Parmi ces locuteurs six (3H, 3F) ont participés à la phase d'apprentissage. Le taux de reconnaissance est de 89% pour les locuteurs connus et de 81% pour les locuteurs inconnus. Parmi les erreurs commises nous avons pu distinguer quatre sous-ensembles. Le premier regroupe les erreurs dues à la prononciation du E muet par 4 locuteurs. Le second concerne les mots dont la production spectrale n'est pas équivalente à la chaîne phonétique (par exemple le /B́/ du mot /OBTU/ est pronnoncé /P/). le troisième est constitué des mots comportant plusieurs variantes phonologiques (par exemple /KATR/ devient /KAT/). Enfin le dernier type d'erreur concerne les mots acoustiquement voisins (par exemple *deux* et *boeufs* , *pointe* et *plainte*). Nous avons corrigé les erreurs des deux premiers types en modifiant la transcription phonétiques des mots et en ajoutant au modèle de mot un état optionnel pour les mots se terminant par une consonne ou la semi-voyelle /j/. Après ces modifications nous avons obtenu 94% pour les locuteurs connus et 88% pour les locuteurs inconnus. Le taux de reconnaissance pour les locuteurs inconnus peut être amélioré si on étend le corpus d'apprentissage à une large population. Le tableau 2 donne les résultats obtenus en fonction des locuteurs. Nous avons constaté que le taux de reconnaisance sur les deux premières positions est de 97% pour les locuteurs de référence est de 94% pour les locuteurs de test.

Le deuxième vocabulaire est constitué de 51 mots: 32 nombres de 0 à 31, les jours de la semaine, et les 12 mois de l'année. Deux répétitions de ce vocabulaire par 6 locuteurs (3H, 3F) inconnus ont été enregistrées. Soient 612 tests de reconnaissances réalisés. Le tableau 3 donne

les résultats sur ce vocabulaire pour chaque locuteur. Le taux de reconnaissance obtenu est de 94,5%.

Conclusion

Nous avons présenté un système de reconnaissance multilocuteur de mots isolés fondé sur une approche phonétique. Le système ne nécessite pas de phase d'apprentissage oral. Le taux de reconnaissance obtenu est de 94,5% pour un vocabulaire de 51 mots et de 88% pour un vocabulaire de 136 mots. Néanmoins ces résultats ne sont qu'une première estimation. L'évaluation des limites du système doit être faite sur des vocabulaires de tailles différentes et pour une large population. Dans ce cas, une prise en compte des principales variantes phonologiques est nécessaire. Nous envisageons une segmentation automatique pour l'extention du corpus à une centaine de locuteurs.

JLG	*JJG*	*PL*	*FNE*	*MA*	*CBI*	*taux moyen*
97	88	96	89	98,5	97,5	94

a)

GA	*HB*	*PH*	*JT*	*FB*	*taux moyen*
85	92,5	86	86	89,5	88

b)

Tableau II: Taux de reconnaissance par locuteur sur le vocabulaire de 136 mots.
a) 6 locuteurs ayant participés à l'apprentissage
b) 5 locuteurs inconnus

CP	*CD*	*FNY*	*CBE*	*HB*	*SK*	*taux moyen*
92	99	97	95	96	89	94,5

Tableau III: Taux de reconnaissance par locuteur pour le vocabulaire de 51 mots.

Références

[1] L.R.Rabiner, S.E.Levinson, A.E.Rosenberg, et J.G.Wilpon, "Speaker independent recognition of isolated words using clustering techniques", *IEEE Trans. Acoust., Speech, and Signal Processing*, vol ASSP-27 (4), pp.336-348, 1979.

[2] S.E.Levinson, L.R.Rabiner, et M.M.Sondhi, "Speaker independent isolated digit recognition using hidden Markov models", *Proc. ICASSP-83*, Avril 1983.

[3] P.Fonsale, "Simulation informatique d'un système multilocuteur de reconnaissance de parole (mots isolés) sans apprentissage oral", *Thèse de Docteur Ingénieur*, Grenoble, 1984.

[4] J.Caelen et G.Caelen, "Indices et propriétés dans le projet ARIAL II", *Actes du Séminaire Encodage et Décodage Phonétique*, pp.129-143, Toulouse, Septembre 1981.

[5] C.Gagnoulet et M.Couvrat, "Seraphine : a connected word speech recognition system", *Proc. ICASSP-82*, Paris, 1982.

[6] J.L.Gauvain, "A syllabe-based isolated word recognition experement", *Proc. ICASSP-86*, Avril 1986.

[7] G.Quenot, J.L.Gauvain, J.J.Gangolf, et J.Mariani, "A dynamic time warp VLSI processor for continuous speech recognition", *Proc. ICASSP-86*, Avril 1986.

[8] H.Sakoe et S.Chiba, "Dynamic programming optimisation for spoken word recognition", *IEEE Trans. Acoust., Speech, and Signal Processing*, vol ASSP-26, (1), pp.43-49, 1978.

[9] B.Prouts, "Traduction phonétique de textes écrits en Francais", *10èmes Journées d'Etudes sur la Parole*, Grenoble, 1979.

[10] J.MacQueen, "Some methods for classification and analysis of multivariate data", *Proc. 5th Berkeley Symp. Probability and Statistics*, Berkely, CA, 1979.

[11] F.Néel, M.Eskenazi, J.Mariani, "Cadrage automatique pour la constitution de dictionnaires d'entités phonétiques", *Speech Communication*, vol.2, pp.193-195, 1983.

APPRENTISSAGE INDUCTIF DE REGLES POUR LE DECODAGE ACOUSTICO-PHONETIQUE

Jacques Guizol

G.I.A., Faculté de Luminy
70 Route Léon Lachamp 13288 Marseille Cedex 09

ABSTRACT

The object of this work is automatic inference of a set of rules which will be used during speech analysis to determine a phoneme or an acoustic category of phonemes from a sequence of acoustic events.

Examples are context-dependant realizations of phonemes constitued by a lattice of patterns representing characteristics of some of the used parameters on parts of signals. These training instances can be grouped in phoneme or acoustic feature classes and treated, by the learning process, on the induced criterion.

Rules of strategy, inference and specialy, generalization are defined in PROLOG II likewise object descriptions specific of the study domain.

The acoustico-phonetic characterization rules so obtained are context-dependant and valued according to the determination accuracy of events they use and to the complexity of the relevant relations.

I - INTRODUCTION

Le système que nous présentons constitue une phase d'acquisition automatique de connaissances symboliques en vue du décodage acoustico-phonétique de la parole.
A partir d'un ensemble d'exemples constitués par un codage de portions de signal issues des réalisations de phrases types, l'apprentissage a pour tâche de fournir une caractérisation du concept induit par le choix de ces exemples.

Les règles de stratégie, d'inférence et en particulier, celles de généralisation sont définies en PROLOG, de même que celles permettant de décrire les objets propres à l'application ou les contraintes de généralisation.

Les règles produites par la méthode comportent une information contextuelle et sont valuées en fonction du nombre d'exemples qu'elles vérifient et de la precision de détermination des objets qu'elles utilisent.

II - CODAGE DU SIGNAL

Notre système opérant un apprentissage par acquisition de concept (événement acoustico-phonétique, phonème, trait acoustique, etc.), les exemples sont constitués d'une représentation symbolique d'une portion de signal, donc de nature tout à fait différente par rapport à celle utilisée auparavant [1]. Celle-ci est obtenue grâce à un ensemble de prédicats évaluables permettant, pour chacun des paramètres du signal, de déterminer maxima, minima, moyennes, pentes afin de modéliser les évolutions temporelles sous la forme de collines, vallées et portions monotones reliées entre elles par des relations situationnelles (coïncidence, succession, intersection, etc.) [2].

A l'issue de ce traitement, nous disposons donc d'une part, de formes élémentaires issues de l'analyse du signal, représentant l'évolution dans le temps des divers paramètres, et d'autre part, de la chaîne phonémique associée. A partir de ces données, le module d'apprentissage détermine les diverses formes associées aux paramètres reconnus comme étant caractéristiques du concept étudié (classe de phonèmes ou trait acoustique), leur configuration ainsi que le contexte ayant éventuellement une incidence sur la caractérisation.

III - PRESENTATION DE LA METHODE

III . 1 - Justification de la Méthode

Dans le contexte de l'application que nous désirions en faire, la panoplie des techniques d'apprentissage nous offrait plusieurs choix. La définition du scénario d'utilisation du système, d'une part, le mode d'acquisition et le type des données, d'autre part, nous ont conduit à choisir les techniques d'apprentissage à partir d'exemples.

Grâce à celles-ci, nous pouvions décider d'obtenir une description de concept selon un critère taxonomique, discriminant ou de caractérisation. Vu le type d'utilisation des règles produites, c'est ce dernier que nous avons choisi. Toutefois, ne disposant pas de contre-exemples (tout au moins dans la description de classes de phonème), il peut

advenir que les ensembles définis par les fonctions de reconnaissance de deux phonèmes acoustiquement proches soient non disjoints.

Dans ce cas, afin de rétablir la consistence du système obtenu, deux attitudes étaient possibles :

a) effectuer une caractérisation en choisissant pour exemples les instances d'un des deux phonèmes et pour contre-exemples, celles du second.

b) opérer un apprentissage caractérisant chacun des deux phonèmes, puis un traitement délivrant des règles permettant la discrimination.

La constatation de recouvrement des domaines de validité ne pouvant intervenir qu'après l'obtention des fonctions de reconnaissance et afin de ne pas doubler le traitement de caractérisation (plus long que celui de discrimination), c'est la deuxième solution que nous avons choisie.

En ce qui concerne le processus de caractérisation, n'utilisant comme nous l'avons déjà dit que des exemples "positifs", nous devions perdre le moins d'informations possible.

La méthode donnée par Hayes-Roth [3] présente l'avantage d'être un processus incrémental permettant un apprentissage dynamique (ce qui peut s'avérer intéressant dans notre application), mais satisfait mal la contrainte enoncée précédemment.

Celle proposée par Michalski [4][5] y répond mieux, en particulier grâce à sa moins grande sensibilité à l'ordre et à la nature des exemples et à la facilité qu'elle offre de pouvoir obtenir des formes généralisées disjonctives. Mais n'ayant volontairement aucun à priori sur la forme et à fortiori le contenu des résultats, nous voulions éviter la phase préliminaire de production d'hypothèses.

La technique à stratégie mixte developpée par Nicolas [6] est séduisante, mais en ce qui concerne notre application nous pouvons faire à son sujet la même remarque que pour la méthode précédente

En définitive, nous avons construit notre système en nous inspirant des procédés ci-dessus dans ce que chacun d'eux offrait de plus intéressant pour notre application (processus incrémental, disjonction interne, caractérisation sous forme disjonctive, outils de généralisation performants, module fournissant les connaissances et stratégies applicables au domaine, etc.).

III . 2 - Structures des Exemples

Conçu dans l'optique d'être indépendant de l'utilisation qui en sera faite, le système admet des exemples dont la syntaxe est peu contraignante. Toutefois, dans un but d'efficacité, aucune interface n'a été prevue et ils doivent donc avoir une structure de termes PROLOG. Plus précisément, chacun d'eux est constitué d'une liste de n-uplets décrivant une conjonction de propriétés et de relations s'appliquant sur des objets.

Les propriétés sont des termes formés d'une part d'un doublet *<type , spécification>*, d'autre part d'un objet identifié par un terme. Les relations sont des triplets dont le premier élément est le symbole relationnel, le second un objet et le troisième une liste (de longueur quelconque) d'objets en relation avec le précédent.

Exemple :

*<phoneme(*II*),61>.<contexte-gauche(*TT*),61>.*
*<contexte-droit(*PP*),61>.<colline-er0(niveau1),62>.*
<precede,62,61>...

III . 3 - Organisation du Système

L'ensemble des règles constituant le système se sépare en deux parties totalement distinctes :

- une partie constituant le moteur d'induction proprement dit et dont les règles de production ou de stratégie présupposent uniquement les exemples mis sous la forme décrite précédemment ;
- une partie contenant la base de connaissance propre au domaine étudié (BCD) constituant un module facilement modifiable ou même interchangeable et qui permet de fournir des informations qui seront utilisées par le système : description arborescente des propriétés (nécessaire à la généralisation par hiérarchie), modèle imposé à la forme généralisée, propriétés des relations, méta-règles de stratégies liées à l'application, etc.

III . 4 - Structure des Résultats

La structure des résultats dans sa forme la plus simple est identique à celle des exemples, mis à part le fait que propriétés et relations voient leur(s) argument(s) représenté(s) par une v ariable PROLOG.

La structure des termes représentant une propriété peut toutefois être plus complexe :

- Lorsque deux propriétés de même *type* mais de *spécification* différente ont été couplées par l'intermédiaire de leur argument apparaissant dans des relations mises en correspondance structurelle, la forme généralisée représentera cette propriété par le *type* commun et la disjonction des *spécifications*.
- L'association de deux objets vérifiant des propriétés de *types* différents mais déterminés dans la BCD comme équivalents donnera naissance à une propriété dont le *type* lui-même sera formé par la disjonction des *types*.

Si l'appariement structurel, offre plusieurs possibilités, la forme generale engendrée sera constituée d'une

conjonction de formes. Sur le plan logique, cette conjonction est justifiée par le fait que tous ses éléments vérifient tous les exemples, mais elle sera interpretée comme une disjonction au meta-niveau. En effet, lors de l'utilisation au cours de l'analyse, des différentes règles ainsi produites et sous l'hypothèse de consistance du système obtenu, la caractérisation pourra être déduite de la première règle vérifiée.

Il peut advenir que les exemples décrivant un concept soient totalement disjoints. Dans ce cas de figure, la méthode de Hayes-Roth échoue. Dans notre système, la forme généralisée devient disjonctive. Dès lors que cette situation apparaît, chacun des exemples devra être vérifié par un seul des éléments de la disjonction.

Exemple:

*k= {ou(<phoneme(*II*),x>.*
<contexte-droit(non.liquide),x>.
<colline-esp(ou.niveau1.niveau2),y>.
<intersecte(1),x,y>.<pr ecede,y,x>.
<ou.colline-ebf.colline-ap1(niveau1),z>.
<intersecte(0),x,z>.<pr ecede,z,x>.
<intersecte(1),y,z>.<pr ecede,y,z>.nil,
*<phoneme(*II*),x>.<c ontexte-droit(liquide),x>.*
<vallee-esp(niveau1),y>.<intersecte(1),x,y>.
<precede,y,x>.nil)}

IV - PRINCIPE DE FONCTIONNEMENT

La caractérisation d'un concept à partir des exemples s'opère pas-à-pas. A chaque étape, disposant d'une forme généralisée FG (issue de l'étape précédente) constituée dans le cas général d'une disjonction de conjonctions de règles, un nouvel exemple est proposé.

Notons $FG = \bigvee_i FG^i_C$ où $FG^i_C = \bigwedge_k FG^i_k$

Le système va tenter d'introduire ce dernier dans l'élément de la disjonction $FG^{i_0}_C$ donnant de lui la meilleure description (intersection maximale). Si chacun des membres de $FG^{i_0}_C$ 'subsume' l'exemple, celui-ci est ignoré puisqu'il n'apporte aucune information. Ceci se vérifie très simplement sous PROLOG en transformant chaque terme de l'exemple en une clause unaire et en démontrant successivement chaque élément $FG^{i_0}_C$ sur l'ensemble ainsi obtenu.

Dans le cas contraire, l'exemple devra pouvoir être introduit dans chacune des règles composant $FG^{i_0}_C$. En cas d'échec sur l'une d'entre elles, celle-ci est supprimée de la conjonction (ambiguité de caractérisation diminuée, voire résolue). En cas d'échec total, l'exemple devient un nouvel élément de la disjonction.

Nous décrivons ci-dessous les étapes successives de l'introduction de l'exemple dans FG^i_C

1) On procède tout d'abord à une factorisation de l'exemple en regroupant les arguments d'une propriété par conjonction interne :

$$P[\,A\,/\,\wedge\,P'[\,B\,/\,\rightarrow\,P''[\,A\,\wedge\,B\,/$$

où $\begin{cases} \text{soit } P = P' = P'' \\ \text{soit EQUIVALENT}(P\,,\,P') \;\&\; P'' = \text{ou.P .P'} \end{cases}$

2) On considère ensuite la disjonction de l'exemple factorisé avec chaque règle FG^i_k de FG^i_C transformée de la même manière. En utilisant la distributivité de $\wedge$ par rapport à $\vee$ et la disjonction interne sur les arguments, on dégage les propriétés communes.

En supposant, par exemple que la règle considérée FG^i_k est de la forme : $P[\,A\,/\,\wedge\,R$ et l'exemple de la forme: $P'[\,B\,/\,\wedge\,E$, on aura la transformation suivante :

$$\{\,P[\,A\,/\,\wedge\,R\,\}\,\vee\,\{\,P'[\,B\,/\,\wedge\,E\,\}\,\rightarrow\,P[\,A\,\vee\,B\,/\,\wedge\,\{\,R\,\vee\,E\,\}$$

où $\begin{cases} \text{soit } P = P' \\ \text{soit } \begin{matrix} P = \text{ou.P''.P'} \\ P = \text{ou.P'.P''} \end{matrix} \end{cases}$

A noter que A et/ou B peuvent être des conjonctions introduites par la factorisation préliminaire.

3) Des disjonctions d'arguments ainsi obtenues, on déduit des couples (formés d'un objet de FG^i_k et d'un objet de l'exemple considéré) qui sont ensuite valués en fonction du nombre de propriétés puis de relations vérifiées. Cette valuation correspond à un degré de pertinence du couple considéré, nécessaire lorsque l'on ne dispose que d'exemples positifs.

4) Le système réduit ensuite les ensembles de couples issus d'une ambiguïté d'appariement en suivant toujours la stratégie visant à perdre le moins d'information possible. Pour cela, il utilise les connaissances de la BCD indiquant les propriétés des relations (symétrie, transitivité, inclusion, etc.).

Exemple :

L'appariement structurel ayant donné deux couples <A,B> et <A,C> où :
B vérifie P[B / ∧ R(E,B) ∧ S(B,D)
C vérifie P[C / ∧ R(C,E) ∧ S'(C,D)
Si S ⇒ S'
Si R est symétrique

... alors, seul <A,B> sera retenu
puisque S' est plus générale que S.

5) A l'issue de cette série de filtres sur les couples, une forme généralisée est produite pour chaque regroupement cohérent des couples liés par relation. Elle est déterminée par la conjonction des propriétés et relations vérifiées par ceux-ci.

A noter que la BCD contient des règles permettant d'influer sur le résultat des phases précédentes. En

particulier, nous pouvons dans la dernière imposer que tout couple soit défini par au moins une propriété et n'apparaisse pas seulement dans une relation.

6) Il peut être nécessaire selon l'application réalisée d'imposer des hypothèses sur le contenu de la forme généralisée.

Exemple :

Dans notre application, nous imposons que la partie significative d'un phonème chevauche de façon significative avec une forme obtenue de l'analyse. Ceci se traduit par la règle :

modele → <*phoneme(x),y*>.<*interse cte(1),y,z*>

Ce modèle minimum sera décrit dans la BCD. Dans le cas où ce dernier est non vide, seules seront conservées les généralisations vérifiant ce modèle.

A ce stade, chaque couple étant remplacé par une variable, si les propriétés de même *type* ont une même *spécification* pour chaque élément du couple, le terme correspondant de la forme généralisée aura une structure identique. Dans le cas où les *spécifications* diffèrent, c'est leur disjonction qui spécifiera la propriété apparaissant dans la forme généralisée.

7) La généralisation par hiérarchie s'effectue en fin de traitement. On dispose pour cela dans la BCD d'autant de descriptions arborescentes que de *types* de propriétés, la spécificité des nœuds augmentant avec la profondeur. Les feuilles constituent en fait l'ensemble des valeurs possibles de la *spécification* d'une propriété dans les exemples de départ.

Par exemple, dans notre application, l'arbre décrivant les propriétés "*phoneme*", "*contexte-gauche*" ou "*contexte-droit*", peut être structuré selon une description acoustique ou articulatoire (le système totalement disjoint de la BCD rendant le remplacement ou la modification de cette description très facile).

Cette généralisation va intervenir sur les propriétés dont la *spécification* est une disjonction. Après recherche du nœud de plus bas niveau, dont dépendent tous les éléments de la disjonction, elle s'opère de la façon suivante :

- Si ce nœud est la racine :
  - si tous les identificateurs de feuilles sont présents dans la disjonction, la propriété, devenue non significative, est alors supprimé (tout le domaine de valeurs étant couvert).
  - dans le cas contraire et si le passage au complementaire n'aboutit pas, la propriété demeure inchangée.
- Si ce nœud se situe en dessous de la racine, la disjonction est remplacée par l'identificateur affecté à ce nœud.

IV - CONCLUSION

Le système d'apprentissage que nous avons présenté constitue un outil très utile pour caractériser des concepts de façon automatique. En particulier, dans l'application que nous en faisons, il nous permet d'obtenir des règles décrivant des réalisations d'unités acoustiques ou phonétiques propres à un locuteur, nous dispensant ainsi de la laborieuse mise au point de règles ad-hoc.

Les temps de calcul sont assez conséquents (45 s par exemple traité sur VAX 730), mais ce traitement devant être effectué une seule fois par locuteur, nous jugeons que cela ne constitue pas un réel problème et compense de toute manière le temps passé à déterminer les règles "à la main". D'autre part, le caractère systématique de la production des règles constitue un net progrès.

BIBLIOGRAPHIE

[1] Guizol J., Meloni H., Gispert J.
Inférence de règles d'adaptation au locuteur dans un système de R.A.P.C ; 14èmes J.E.P., pp 315-318, Paris 10-13 juin 1985

[2] Meloni H., Bulot R.
Décodage Acoustico-Phonétique en PROLOG ; 15èmes JEP, Aix-en Provence, 27-29 mai 1986

[3] Hayes-Roth F., Mc Dermott J.
An Interference Matching Technique for Inducing Abstractions ; Communications of the A.C.M, Vol. 21, N° 5, pp 401-410, 1978

[4] Michalski R. S.
Inductive Learning as Rule-Guided Generalization and Conceptual Simplification of Symbolic Descriptions ; Workshop on Current Developments in Machine Learning, CMU, Pittsburgh, July 16-18, 1980

[5] Dietterich T. G., Michalski R. S.
Inductive Learning of Structural Descriptions : Evaluation Criteria and Comparative Review of Selected Methods ; Artificial Intelligence 16 pp. 257-294

[6] Nicolas J.
Les Stratégies de Contrôle dans l'Apprentissage à partir d'Exemples ; J.F.A, Orsay, Fevrier 1986

DECODAGE ACOUSTICO-PHONETIQUE EN PROLOG

H. Meloni, R. Bulot

G.I.A., Faculté de Luminy
70 route Léon Lachamp 13288 Marseille Cedex 09

ABSTRACT

We present a system for continuous speech acoustic and phonetic decoding implemented in PROLOG II. A restricted set of general predicates allows the high cost numerical processings. The different sources of knowledge, represented by clauses, can be regrouped by their abstraction level. Some patterns show parts of the signal marked by a specific property of one or several parameters. Arrangements of that patterns (regrouping, coincidence, succession, overlaping, etc.) define acoustic events. The next level is constitued by pseudo-phonetic events corresponding to a first context-dependant interpretation of acoustic elements (phonemic nucleis, burst, transition, etc.). In addition to their formal labeling, these units are marked by pseudo-phonetic features which the definition is context-dependant. All results usable by the upper levels of knowledge are inserted in a network encoded by clauses of PROLOG II.

I - INTRODUCTION

Les travaux que nous avons accomplis dans le cadre de la reconnaissance de la parole [1], [2], [3] séparaient assez nettement les traitements numériques, exécutés dans un langage algorithmique classique, des traitements de données symboliques effectués en PROLOG. Cette dichotomie artificielle interdisait l'interaction et l'optimisation contextuelles des deux processus. Afin de permettre une coopération simple entre toutes les sources de connaissances, nous proposons, sous la forme d'un ensemble de prédicats du langage PROLOG II [4], un environnement souple et efficace pour l'acquisition, la manipulation, l'évaluation, la représentation et le traitement d'informations acoustiques, phonétiques et linguistiques. Les outils développés permettent, de manière interactive, de produire diverses paramétrisations du signal, de décrire et reconnaître des formes simples, de définir et identifier des événements, des propriétés, des indices et des traits acoustico-phonétiques, de coder ces informations et les stratégies qui les utilisent et de structurer l'ensemble des résultats produit sous la forme d'un treillis d'unités valuées. Nous illustrons les possibilités nouvelles de cet environnement en présentant les premiers constituants d'un système de décodage acoustico-phonétique réalisé entièrement sous PROLOG II.

II - PARAMETRISATION DU SIGNAL

Le but visé à ce stade du traitement est de caractériser de manière précise et peu coûteuse une portion de signal au moyen d'une suite de vecteurs de paramètres. Les limites de la zone codée, la nature des attributs retenus ainsi que leurs conditions d'évaluation sont déterminées en examinant des connaissances de niveau acoustique, phonétique ou phonologique.

II - 1 - Prédicats évaluables pour la paramétrisation

Nous disposons de 2 prédicats évaluables qui effectuent le calcul des paramètres et leur chargement dans une mémoire accessible à d'autres fonctions réalisant des opérations numériques complexes. Chaque vecteur, produit à intervalles réguliers de 10 ms, est constitué d'une vingtaine d'attributs temporels et spectraux (répartition spectrale de l'énergie, densité des passages par zéro, position, amplitude, émergence et largeur des pics, etc.). Les spectres lissés sont obtenus à partir des coefficients cepstraux ou de LPC dont un ensemble de variables définit les conditions de calcul (portion de signal traitée, méthode utilisée, nombre de coefficients, préemphase, rayon, etc.).

II - 2 - Utilisations des prédicats de paramétrisation

Les prédicats de paramétrisation ont été employés, dans la phase d'acquisition des connaissances acoustiques, afin d'évaluer les conditions optimales du codage du signal pour diverses portions caractéristiques des phonèmes. Nous disposons pour cela d'une base de donnée des sons du français constituée d'une centaine de phrases prononcées par 4 locuteurs masculins. Chaque unité phonétique apparaît dans un ensemble de contextes phonémiques des plus neutres aux plus contraignants. Une segmentation semi-automatique des énoncés fournit une séquence de marqueurs temporels indiquant des limites de phonèmes ou de portions de phonèmes. Ces connaissances, associées à la description symbolique des phrases, permettent d'effectuer interactivement des accès aux sons de référence à partir d'attributs quelconques de leur caractérisation formelle (phonèmes, traits, contextes, etc.).

La stratégie du système de décodage conduit à une première représentation d'un énoncé au moyen de vecteurs de paramètres évalués toutes les 10 ms dans des conditions standard (14 coefficients de LPC, sans préemphase, rayon unité, etc.).

Ces données peuvent être localement modifiées afin d'obtenir une meilleure spécification de certains sons. Généralement il s'agit de proposer une description plus satisfaisante des spectres qui est trop riche pour les phonèmes sourds, trop pauvre pour les voyelles nasales ou peu précise dans le cas de pôles rapprochés. Les clauses qui activent ce prédicat correspondent à différents niveaux de traitement (acoustique, phonétique, phonologique) et prennent en compte des informations contextuelles variées.

III - RECONNAISSANCE DES FORMES

Les outils proposés dans ce cadre ont pour objectif la modélisation et la symbolisation des évolutions temporelles de certains groupes de paramètres. Ils constituent les éléments de base minimaux pour la description et la reconnaissance de phénomènes acoustico-phonétiques pertinents. Des techniques de ce type, fondées sur la mise en évidence d'événements acoustique simples, ont donné par ailleurs des résultats satisfaisants [5].

III - 1 - Prédicats évaluables de reconnaissance de formes

Sur un intervalle de temps, ils définissent des fonctions simples d'un paramètre telles que la mesure de ses extrema, le calcul de sa moyenne, la caractérisation de son instabilité, etc. Ils déterminent également des fonctions complexes d'un ensemble d'attributs comme le suivi temporel du trajet de pics spectraux, la valuation de la continuité ou de la monotonie d'un phénomène, etc. Enfin, ils désignent et identifient des schémas de formes parmi lesquels les modèles de collines ou de vallées sont les plus utilisés. Des variables permettent de préciser les contours d'une forme en décrivant des sous-familles d'un schéma donné. Ainsi, la définition d'un type de colline, pour un paramètre quelconque, sera donnée par sa largeur minimale, ses émergences limites à gauche et à droite, le seuil maximum de déviation acceptable ainsi que le seuil de bruit au-dessous duquel le paramètre n'est pas significatif.

III - 2 - Utilisations des prédicats de R.F.

Les prédicats évaluables sont parfois directement utilisés dans les règles mais, le plus souvent, ils apparaissent indirectement sous forme fonctionnelle ou bien contribuent à la définition de phénomènes complexes plus proches des représentations symboliques manipulées par les experts (formants, transitions, barre de voisement, buzz, etc.). Ainsi, un certain type de voisement sur une portion z du signal est caractérisé par la règle suivante :

voise1(z) →
inferieur(50,ebf(z))
inferieur(dpz(z),70)
non(coincidence-sur(z,colline1-dpz))
inferieur(plus(esp(z),10),ap1(z))
inferieur(limite-sup(Fp1,z),900) ;

dans laquelle les identificateurs *ebf*, *esp*, *dpz*, *fp1*, *ap1* désignent les énergies basse fréquence et moyenne, la densité des passages par zéro, la fréquence et l'amplitude du premier pic spectral.

A partir de schémas de formes simples, appliqués pour l'ensemble des paramètres sur les phrases de référence, nous avons sélectionné quelques dizaines de formes représentatives de portions de sons déterminées. Ces éléments (collines, vallées, zones monotones ou stables, etc.) concernent, souvent de plusieurs manières, la plupart des attributs temporellement continus. Par exemple, les règles de segmentation utilisent, pour la description de sons constrictifs (consonnes fricatives, explosions, etc.), plusieurs types de collines du paramètre mesurant la densité des passages par zéro du signal définis de la manière suivante :

colline1-dpz(<i,j>,<k,l>,m) →
Colline(4,i,j,25,5,25,25,1,k,l,m) ;

colline2-dpz(<i,j>,<k,l>,m) →
Colline(4,i,j,25,5,15,15,1,k,l,m) ;

colline3-dpz(<i,j>,<k,l>,m) →
Colline(4,i,j,10,10,10,10,1,k,l,m) ;

les termes $<i,j>$, $<k,l>$ et m désignant respectivement la zone de la recherche indéterministe de la forme, les bornes d'un des éléments satisfaisant les contraintes et la position du maximum. Dans les règles, seul le nom d'une sous famille d'un schéma est mentionné, les zones sont manipulées par des prédicats qui masquent à ce niveau les effets de la stratégie.

Dans le système de décodage, les formes retenues définissent des événements acoustiques et phonétiques et constituent des repères pour guider le processus de reconnaissance vers les phénomènes les plus saillants.

Des techniques d'apprentissage automatique nous ont permis d'obtenir un ensemble de règles qui associent à un phonème en contexte les formes de base qui le caractérisent ainsi que les relations qui lient ces événements [6]. Les exemples sont constitués d'une part, du phonème et de son environnement (phonèmes adjacents) et, d'autre part, de toutes les sous familles de formes qui ont une intersection avec l'unité traitée et des relations temporelles qui unissent ces formes.

IV - NIVEAUX SYMBOLIQUES DU SYSTEME

Les connaissances acoustico-phonétiques formelles d'un même type sont regroupées en niveaux pour leur présentation, mais chaque règle peut être sollicitée indépendamment de sa classe.

IV - 1 - Prédicats prédéfinis de l'environnement

Ces outils contribuent à rendre plus naturelle l'expression de la connaissance et définissent pour l'essentiel les fonctions suivantes :

- relations temporelles entre des unités du treillis de résultats (coïncidence, intersection, succession, union, adjacence, inclusion, etc.),
- démonstrations particulières d'une liste de prédicats pour la gestion du contrôle et la visualisation des parcours (effacement déterministe ou complet, vérification de l'existence d'une ou de plusieurs solutions, saturation des effacements, impression de traces, etc.),

- opérations logiques sur des listes de prédicats (conjonction, disjonction, négation, implication, etc.),
- opérations arithmétiques diverses acceptant des fonctions en paramètres,
- manipulations complexes sur les arbres.

IV - 2 - Evénements acoustiques et phonétiques

Les événements acoustiques sont définis par regroupement de formes au moyen des prédicats qui décrivent des relations temporelles entre les éléments de base. Les unités engendrées ne reçoivent pas d'interprétation phonétique ; elles mettent en évidence la conjonction de propriétés acoustiques du signal et caractérisent généralement des segments infra-phonémiques.

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.

Les clauses qui définissent ces connaissances opèrent dans des contextes souvent très différents suivant qu'il s'agisse d'événements "évidents" ou de segments tributaires de l'identification préalable de l'environnement. Ces règles sont indépendantes du locuteur, elles opèrent une partition peu ambiguë d'un énoncé en macro-classes pseudo-phonétiques. Les exemples ci-dessous décrivent respectivement un type particulier d'événement vocalique évident et des conditions de recherche d'un événement vocalique secondaire :

```
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) ) ;

evenement-vocalique secondaire  →
      domaine-de-recherche(<i,j>)
      appartient(longueur(<i,j>),<4,20>)
      debute-en(s,j)
      evenement-consonantique(s)
      termine-en(t,i)
      evenement-consonantique(t)
      trace-regle(evenement3-voc(N))
      ajout-treil(N) ;
```

IV - 3 - Traits pseudo-phonétiques

Chaque événement pseudo-phonétique est caractérisé par un faisceau de traits hiérarchisés dont chacun est défini par un ensemble de clauses. Les règles qui représentent ces connaissances utilisent de nombreuses informations contextuelles sur la nature, les paramètres, les propriétés, les indices ou les traits des sons adjacents. Chacune correspond à une conjonction d'indices qui déterminent un trait dans une situation particulière. Toutefois, les conditions d'application de plusieurs règles désignant un même trait ne sont pas toujours disjointes et un résultat identique peut être produit de plusieurs manières. Cette étape de la reconnaissance fonctionne comme un filtre phonétique limitant le nombre des phonèmes candidats. Le choix des solutions les plus vraisemblables résulte de l'évaluation d'un score à partir de paramètres sélectionnés et ajustés en fonction de caractéristiques des segments contigus.

L'acquisition et l'évaluation de ces règles sont effectuées de manière interactive sur les phrases de référence. Une étude statistique des paramètres ou de certaines fonctions de plusieurs d'entre eux permet de désigner les attributs les plus discriminants pour la détermination d'un trait d'un phonème dans un environnement précis. Les règles qui en résultent sont immédiatement testées sur l'ensemble des situations où elles sont susceptibles de s'appliquer. L'analyse des résultats erronés permet de préciser nos connaissances des phénomènes de coarticulation et d'intégrer les informations contextuelles indispensables à la production d'éléments fiables. Pour la description et l'évaluation des traits *grave* et *aigu* des occlusives sourdes nous disposons de plusieurs règles. Les clauses suivantes représentent deux ensembles de conditions permettant d'affecter le trait d'acuité dans certains types d'environnements :

```
acuite-occ-sourde(z,grave(2))  →
      inferieur(cgh(z),3200)
      inferieur(afmedian(z),ebf(z))
      inferieur(afhaut(z),ebf(z))
      si-alors(inferieur(moins(fois(2,ebf(z)),10)
                         ,plus(afmedian(z),afhaut(z))),
               inferieur(afhaut(z),afmedian(z)))
      si-alors(inferieur(300,fbas(z)),
               inferieur(afbas(z),plus(6,ebf(z)))) ;

acuite-occ-sourde(z,aigu(2))  →
      inferieur(2800,fem(z))
      etendre(z,0,5,z')
      non(tres-aigu(z')) ;
```

les identificateurs *fem, cgh, fbas, afbas, afmedian, afhaut* représentent la position fréquentielle du maximum d'énergie spectrale (dans une fenêtre de 800 Hertz), le centre de gravité de la partie haute du spectre, les positions et amplitudes des pics dominants dans les zones 0-1200 Hz, 1200-3000 Hz et 3000-5000 Hz.

Le traitement des informations contextuelles induit une circularité du processus de décodage car la détermination d'un trait peut exiger la reconnaissance préalable de certaines caractéristiques des sons environnants qui sont eux-mêmes soumis aux altérations du phonème traité. Nous avons écarté la solution consistant à effectuer des insertions conditionnelles dans le treillis afin d'exploiter au maximum l'information présente dans le signal. Chaque règle est donc indépendante, elle conduit parfois à réaliser des calculs qui seront par ailleurs nécessaires à l'identification d'une autre unité.

V - TREILLIS DES RESULTATS

Certains résultats, considérés comme définitivement acquis au cours du décodage d'un énoncé, sont conservés dans un treillis constitué de clauses PROLOG. Cette structure est rendue souple et efficace au moyen d'un ensemble de prédicats qui permet de réaliser des opérations telles que l'ajout et la suppression d'unités en un point quelconque, des parcours multiples, le repérage des zones libres, des accès diversifiés à une unité (par la position de ses bornes, par son type ou ses caractéristiques, etc.), la récupération des arguments et des limites d'un élément, etc.

Cette technique de mémorisation temporaire ou permanente de résultats partiels du processus de décodage s'apparente, par certains aspects, à celle employée dans le système HEARSAY II [7]. D'autre part, elle permet de mélanger les productions de divers niveaux d'interprétation du message (acoustique, phonétique, lexical, syntaxique, sémantique) et de positionner dans le temps les unités identifiées. Ainsi, les stratégies utilisées pour contrôler les opérations de reconnaissance d'une phrase peuvent être indépendantes à la fois de la hiérarchie des connaissances et de l'ordre temporel des unités.

Dans la phase de décodage acoustico-phonétique, la stratégie employée conduit à exécuter non séquentiellement les étapes suivantes :

- calcul des paramètres sur l'ensemble de l'énoncé,
- reconnaissances des formes constituant les noyaux nécessaires à la définition d'événements sûrs,
- identification et mémorisation dans le treillis des événements acoustiques évidents,
- recherche dans les zones libres des événements secondaires et transitoires,
- regroupement des segments étiquetés pour produire et ajouter des événements consonantiques,
- affinement de la segmentation des noyaux vocaliques pour déterminer, quelquefois de manière ambiguë, des événements vocaliques qui vont enrichir le treillis,
- identification des traits pseudo-phonétiques puis ajout des phonèmes les plus vraisemblables après filtrage et calcul du score,
- interprétation des zones non reconnues au moyen de l'ensemble des unités du treillis.

VI - CONCLUSION

La brièveté de l'exposé ne donne qu'une image imparfaite de la puissance potentielle de l'environnement proposé. Son utilisation pour les tâches d'apprentissage et d'acquisition des connaissances acoustico-phonétiques nous a permis de constituer très rapidement un important ensemble de règles dont certaines demeurent encore perfectibles. Le traitement de ces connaissances fournit des résultats plus fiables que ceux que nous obtenions au moyen de techniques classiques. Les améliorations obtenues sont liées d'une part à l'enrichissement des informations disponibles et, d'autre part, à la prise en compte de phénomènes subtils de coarticulation. La durée du processus de décodage est tout à fait raisonnable sur un minicalculateur, et nous pouvons envisager de réaliser sous PROLOG un système complet de reconnaissance automatique de la parole.

Les outils dont nous disposons suffisent pour le traitement de connaissances acoustico-phonétiques décrites au moyen d'un certain type de paramétrisation du signal (spectres lissés). Les prédicats et fonctions de plus bas niveaux diffèreraient sensiblement pour d'autres techniques de codage. Nous envisageons d'étendre notre environnement pour intégrer et évaluer des règles proposées par différents experts. Par ailleurs, il nous semble également indispensable, pour améliorer l'efficacité du système, de disposer de moyens d'interaction graphique sous PROLOG pour l'acquisition et la modification de données visualisées (segmentation semi-automatique, étiquetage, suivi de formants, présentation de nouveaux paramètres, etc.).

BIBLIOGRAPHIE

[1] Méloni H.
Etude et réalisation d'un système de reconnaissance automatique de la parole continue ; Thèse de Doctorat d'Etat, Université d'Aix-Marseille II, Faculté de Luminy, février 1982

[2] Méloni H.
Traitement des Contraintes Linguistiques en Reconnaissance de la Parole ; TSI, vol. 2, n° 5, septembre-octobre 1983, pp 349-363

[3] Méloni H., Gispert J., Guizol J.
Traitement de connaissances pour l'identification analytique de Mots dans le discours continu ; Congrès AFCET Informatique 5ème Génération, Paris 5-7 mars 1985, pp 339-350

[4] Colmérauer A., Kanoui H., Van Caneghem M.
PROLOG : Bases théoriques et Développements actuels; TSI, vol. 2, n° 4, juin-juillet 1983, pp 271-311

[5] De Mori R., Laface P., Mong Y.
Parallel Algorithms for Syllable Recognition in Continuous Speech ; IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. Pami-7, n° 1, janvier 1985

[6] Guizol J.
Apprentissage inductif de règles pour le décodage acoustico-phonétique ; 15ème JEP, Aix-en-Provence, 27-29 mai 1986

[7] Erman L.D., Lesser V.R.
The HEARSAY II Speech Understanding System : a tutorial ; in *Trends in Speech Recognition* ; W.A. Lea éditeur, Prentice-Hall, 1980

RECONNAISSANCE DE MOTS ISOLÉS PAR LA MÉTHODE DESCRIPTIVE DE TRAITS PHONÉTIQUES

R. GUBRYNOWICZ, K. MARASEK, W.W. WIĘZLAK

Institut des Recherches Fondamentales de Technologie
Academie Polonaise des Sciences - Varsovie - Pologne

ABSTRACT

A speaker independent isolated word recognition system based on approximate articulatory description of speech signal is briefly described. In presented system a few global spectral parameters extracted by special purpose hardware unit is used to evaluate the manner of articulation and for labeling of speech segments. Each word is represented by a string of labeled segments and given to the input of a net of fuzzy automata for further, more complex analysis. Each word is divided into three parts, initial one before the first vowel in the word, final after the last vowel and the intermediate between these two vowels. These parts are analysed independently by corresponding sequence of automata. At the output a fuzzy description of the word using 8 articulatory classes is given.

INTRODUCTION

Le problème de décodage au niveau acoustico-phonétique est toujours un des points majeurs dans chaque système de reconnaissance de parole. Le choix d'ensemble de paramètres et leur traitement sont déterminant pour obtenir une bonne reconnaissance.

En général, au premier niveau d'analyse du signal de parole on mesure les paramètres les plus faciles à déterminer et à interpréter. Parmi les paramètres les plus fréquemment utilisés, on peut mentioner tels que les variations du niveau d'énergie totale du signal, du niveau d'énergie dans les basses fréquences, dans les autres bandes des fréquences etc. Les valeurs absolues ou relatives des paramètres mesurés servent à déterminer les indices acoustiques et sont à la base de leur transformation en traits phonétiques. Ce dernier processus, en réalité, consiste à émettre des hypothèses sur la présence dans le signal des traits choisis.

Dans notre système, comme point de départ on prend une caracteristique des valeurs mesurées en employant des termes tels que "petit", "moyen", "grand", "très grand" etc., pour représenter d'une manière descriptive les relations entre les indices acoustiques et les traits phonétiques [1]. Cette forme de présentation est d'ailleurs souvent utilisée dans les traveaux en phonétique.

Il semble que la théorie des sous-ensembles flous soit particulièrement bien adaptée à exprimer ces relations d'une façon formelle, et en allant plus loin, à établir à quel point un segment de parole possède une propriété donnée [2].

Nous allons présenter une méthode de description articulatoire développée pour un système de reconnaissance de mots isolés. La description est réalisée par des reseaux d'automates flous. Nous pensons cependant, que la méthode reste valable non seulement pour les systèmes de reconnaissance de mots isolés mais aussi pour les systèmes prévus pour la parole continue.

L'ENSEMBLE DE PARAMÈTRES ANALYSÉS

Le signal d'entrée est préaccentué et analysé à l'aide d'un analyseur paramétrique qui permet de mesurer 7 paramètres suivants: le niveau d'amplitude dans les basses fréquences (LP) et dans les hautes fréquences (HP), le niveau d'amplitude de l'enveloppe du signal (AO), la densité des passages à zéro, la fréquence fondamentale et les fréquences de deux premiers formants. Toutes les 10 ms les valeurs mesurées des paramètres sont envoyées au mini-ordinateur. Dans la version actuelle du système on utilise les trois premiers paramètres seulement. En plus, à partir de ces parametres on calcule 4 indices qui caractérisent leurs variations relatives par rapport aux valeurs maximales pour un mot donné et par rapport entre eux.

L'ensemble de paramètres employés contient donc encore:

$DA = AO_{max} - AO$; $DL = LP_{max} - LP$;
$HL = HP - LP$; $HA = HP - AO$.

EVALUATION DES VARIATIONS DU MODE D'ARTICULATION DANS LE MOT ANALYSÉ

Nous avons admis au départ que pour un vocabulaire restreint la reconnaissance de mots isolés peut s'appuyer sur la description articulatoire laquelle caractérise l'évolution des traits suivants: le mode et le lieu d'articulation, sans avoir recours à l'identification des phonèmes (au moins au stade préliminaire).

Le premier niveau d'analyse est destiné à déterminer les variations du mode d'articulation dans le mot prononcé. L'analyse du lieu d'articulation, qui n'est pas encore entièrement introduite dans notre système, s'effectue au niveau suivant et uniquement dans le cas où il y a des mots différents ayant le même "squelette" articulatoire.

Le mode d'articulation n'est pas aussi explicitement lié aux indices acoustiques que le lieu d'articulation (voir par exemple, la relation entre le lieu et les fréquences de formants). Pour cette raison, nous utilisons les relations de dénomination [3] qui caractérisent d'une façon formelle le rapport entre la description des paramètres par des valeurs linguistiques floues telles que, "grands", "moyen", "petit" etc..., et l'interprétation phonétique donnée. En effet, le raisonnement sur la présence dans le signal d'un trait phonétique ne peut être que rarement fondé sur l'analyse d'un seul indice acoustique. Pour obtenir une évaluation assez fiable, plusieurs indices doivent être combinés, chacun caractérisé par sa fonction d'appartenance, dotée de la restriction correspondante. A la sortie, chaque échantillon de 10 ms est décrit par une séquence de variables linguistiques, associée aux restrictions floues définies dans l'univers des paramètres. Ces restrictions déterminent le sens de variables d'une manière descriptive. Les variables sont combinés pour former des compositions complexes dans lesquelles des conjonctions comme "et" ou des disjonctions comme "et/ou" sont utilisées.

Dans notre système 4 classes articulatoire sont distinguables après l'analyse indépendante du contexte: résonante définie par les sons prononcés avec une configuration du conduit vocal assurant le passage libre du soufle d'air sur toute la longueur du conduit; obstruente caractérisée par les sons articulés en formant dans le conduit une obstruction complète ou partielle. Il en résulte que le passage de l'air est, soit bloqué, soit fortement freiné et devenant turbulent. Dans ce groupe on peut distinguer les stridents et le reste que nous avons appelé "obstruent candidates" formant deux classes relativement homogènes. Les stridents sont engendrés à partir d'une forte friction de l'air à l'intérieur du conduit vocal. La dernière classe articulatoire c'est la tenue sourde qui englobe toutes les parties silencieuses présentes dans le signal, correspondant soit au blocage de l'air devant les occlusives non-voisées, soit a l'aspiration ou au rétard dans le rétablissement du voisement dans une séquence sourde-voisée dont le lieux d'articulation sont très différenciés, ce qui est dans le cas des suites /su/, comme dans le mot "ressource" ou /sw/ dans le mot "soit", par exemple.

Chaque mot, sur la base de ces quatres classes articulatoires est représente au premier niveau d'analyse par une séquence de segments classifiés indépendament du contexte (Fig. 1) et obtenus à l'issue de la description floue de 7 fonctions paramétriques énumérées plus haut.

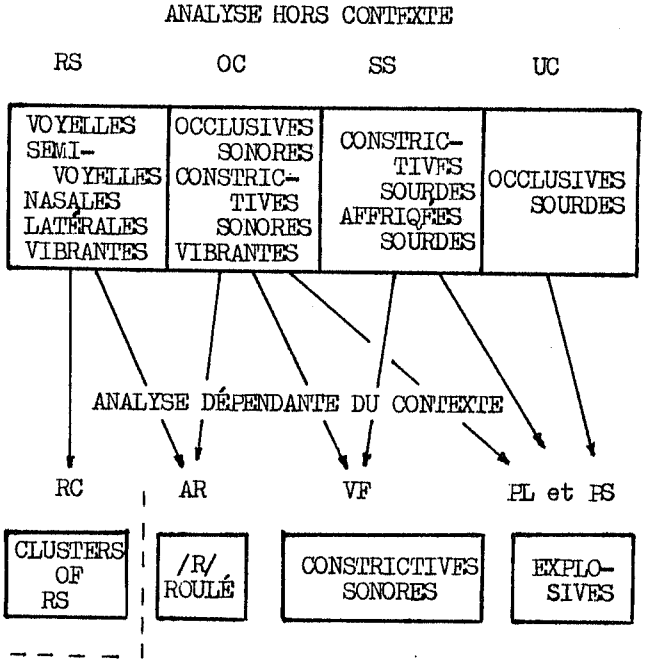


Fig. 1

Dans la transformation des indices acoustiques en traits phonétiques, les valeurs de ces fonctions sont calculées en tenant compte de l'ambiguïte naturelle de la parole. Au lieu de décrire précisement dans le domaine paramétrique les notions articulatoires plus ou moins vaguement définies, il est plus simple et éfficace d'employer une description linguistique de variations des paramètres. Dans notre cas, cette description s'obtient à l'aide du questionnaire suivant:

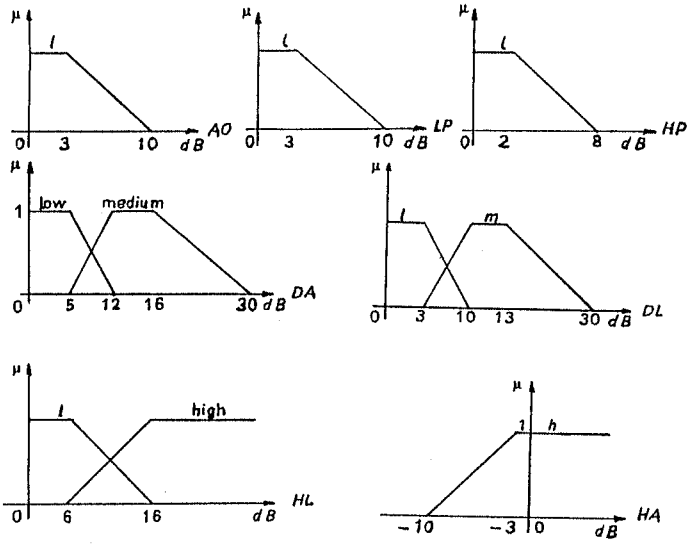
Q_1 = l'échantillon est-il resonant?
Q_2 = l'échantillon est-il strident?
Q_3 = l'échantillon est-il "tenue sourde"?
Q_4 = l'échantillon est-il potentiellement obstruent?

Chacune de ces questions complexes se compose de questions simples concernant les variations des niveaux d'amplitude dans les zones de fréquences mentionées. Les définitions formelles de quatres classes articulatoires indépendantes du contexte sont suivantes:

résonante RS = $l(DA) \circ [l(DL) \vee l(HL)]$
stridente SS = $h(HL) \wedge h(HA)$
tenue sourde UC = $l(AO) \wedge l(LP) \wedge l(HP)$
obstruent candidates
OC = $\overline{UC} \wedge \overline{SS} \wedge [m(DA) \circ m(DL)]$

avec:
o conjonction forte $AoB = Max(a+b-1, 0)$
$\wedge$ conjonction rigide $A \wedge B = Min(a,b)$
$\vee$ disjonction $A \vee B = Max(a,b)$

Et, par exemple, dans notre système les résonants sont caractérisés par un "grand" niveau d'amplitude de l'enveloppe donc par petite différence DA et une petite valeur DL ou HL. Les stridents par contre sont caractérisés par grande différence de niveaux HL et grande différence HA. Sur la Fig.2 on a presenté pour chaque terme employé les fonctions d'appartenance qui déterminent son signification dans le domaine du paramètre analysé.



$$\mu(RS) = \max\{\mu_{l(DA)} + \max[\mu_{l(DL)}, \mu_{l(HL)}] - 1, \emptyset\}$$

$$\mu(SS) = \min\{\mu_{h(HL)}, \mu_{h(HA)}\}$$

$$\mu(UC) = \min\{\mu_{l(AO)}, \mu_{l(LP)}, \mu_{l(HP)}$$

$$\mu(OC) = \min\{1-\mu(SS), 1-\mu(UC), \max[\mu_{m(DA)} + \mu_{m(DL)} - 1, \emptyset]\}$$

Fig. 2

Au bas de la figure, nous avons les formules qui permettent d'évaluer le degré d'appartenance, donc la possibilité qu'une propriété correspondant a la classe articulatoire donnée soit présente dans l'échantillon étudié. Chaque segment de 10 ms est ainsi décrit par un ensemble de 4 degrés d'appartenance et on donne au segment l'étiquette de cette classe qui a obtenu la plus grande valeur. Les segments sont ainsi définis et, en utilisant quelques règles syntaxiques simples, on les regroupe en segments homogènes, plus longs. En outre, on peut en même temps faire quelques corrections mineures sur la description primaire il s'agit surtout des segments qualifiés de "tenue sourde" . Il faut mentionner que dans certains cas, il y a aussi des segments non classifiés, c'est-à-dire, tous les 4 degrés egal 0. Néaumoins, on les a utilisé dans l'analyse ultérieure, surtout pour la détection de faibles plosives prononcées dans la partie finale du mot.

Voici quelques exemples des descriptions articulatoires obtenues pour le mot "zéro" à ce stade d'analyse.

ZERO /zero/
le numéro de l'échantillon initial
OC2/11;SS13/1;OC14/4;RS19/40
nombre d'échantillons

OC2/14;RS16/15;OC31/3;RS34/23

OC1/8;SS9/2;RS11/24;OC35/4;RS39/16

L'ANALYSE DU MODE D'ARTICULATION DEPENDANTE DU CONTEXTE

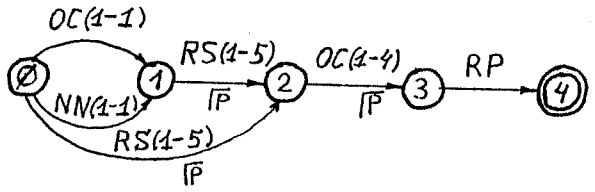
La description obtenue à l'étape précédente d'analyse sert de base à la discrimination de trois unités articulatoires dépendante du contexte (Fig.1)- les fricatives voisées, le /r/ roulé et les explosions d'un type different. En plus, on a identifié le regroupement de résonants nommé résonant cluster RC. Cette étape d'analyse du mot est realisé à l'aide d'un ensemble de réseaux d'automates flous.

Chaque mot est divisé en trois parties: initiale, finale et centrale. Ce partage est determine par la première et la dernière résonante sûre RP, c'est-à-dire, la résonante pour laquelle le degré d'appartenance égal 1. Donc, en pratique, la partie initiale se termine pour la plupart de mots sur la première voyelle et la partie finale commence après la dernière.

Pour chaque partie du mot on a élaboré un réseau d'automates à part, mais certains d'entre eux étaient utilisé plusieurs fois dans le deux ou trois réseaux, en différentes places. Au cours d'analyse de la séquence d'étiquettes de classes articulatoires l'automate choisi est celui, qui arrive à son état final pour la séquence d'entrée donnée. A ce moment, il peut géné-

rer une ou plusieurs séquences de sortie et l'analyse passe au réseau suivant, soit, si la séquence n'est que partielle, on passe à un autre automate.

Sur la Fig. 3 on a donné un exemple d'un automate simple qui réalise la détection de /r/ roulé prononcé dans la partie initiale du mot. L'une des séquences d'entrée acceptée par l'automate peut être suivante:
OC1/1;RS2/3;OC5/4;RP9/20



Matrice de transitions

Entrée / Etat	OC 1-1	OC 2-4	NN 1-1	RS 1-5	RP
∅	1	/	1	2 ⌈p	/
1	1	/	/	2 ⌈p	/
2	3 ⌈p	3 ⌈p	/	/	/
3	/	/	/	/	4
4					

séquences de sortie

AR
OC — RP
#

Les conditions de la description finale:

$LEN(RS_d) + LEN(OC_d) > 3 \longrightarrow \mu(AR) = \mu_A(\delta_A); \mu(OC) = \mu_l(\delta_A);$ $\mu(\#) = \emptyset$.

$LEN(RS_d) + LEN(OC_d) \leqslant 3 \longrightarrow \mu(AR) = \emptyset \;;\; \mu(\#) = 1$.

Fig. 3

où ⌈p - à ce état on met sur pile de l'étiquette correspondante l'adresse et la longuer du segment; / - case vide; ⟶ - génération des séquences de sortie; $LEN(RS_d)$ $LEN(OC_d)$ - les longeurs de derniers segments RS et OC; $\mu(\delta_A)$ - fonction d'appartenance décrivant la "profondeur" de l'abaissement temporaire de l'enveloppe du signal.

RESULTATS

La méthode présentée a été verifiée pour un ensemble de 60 mots prononcé par 11 personnes. Apres la première étape d'analyse la description des mots donne le nombre de squelettes de référence ne dépassants en général 3-4 par mot. La classification des mots decrits à l'aide de 8 classes articulatoires donne une liste de 7 mots au plus, candidats pour la réconnaissance ultérieure. Il faut souligner, malgré sa extrême simplicité, la grande fiabilité de la procédure, surtout dans le cas des unités articulatoires indépendantes du contexte. Par exemple, pour l'ensemble de mots analysés, on a eu 1155 occurences de résonantes et il n'y avait que trois omissions et deux fausses insertions. Pour les stridents on n'a pas noté d'erreurs. En outre, 29% des mots prononcés étaient déjà reconnus à ce niveau d'analyse il n'y avait pas d'autres mots "potentiels" . Pour un ensemble de 30 mots permettant de construire un simple langage de communication homme-machine on a obtenu (pour le mêmes locuteurs) un taux de reconnaissance proche de 97%.

L'avantage de la méthode présentée est surtout la simplicite de génération de la description articulatoire du mot à reconnaitre. Le fait que les règles de cette description sont locales, facilite leur définition et leur modification. D'autre part, l'application de la théorie des sous-ensembles flous dans l'analyse de la parole fait que l'interprétation articulatoire est entièrement paramétrable malgré son caractère descriptive, et par sa nature même, imprecise et approximative.

REFERENCES

[1] R. GUBRYNOWICZ, Mapping acoustic cues into phonetic features, Proc. IV F.A.S.E. Symp. on Acoustics and Speech, vol. 2, 71-85, Venezia, 1981.

[2] R. GUBRYNOWICZ, Application de la théorie des sous-ensembles flous à l'analyse et à la reconnaissance automatique de la parole Note Technique, NT/LAA/TSS/157, CNET, Lannion, 1983.

[3] L.A. ZADEH, A fuzzy algorithmic approach to the definition of complex or imprecise concepts, Int. J. Man-Machine Studies, 8, 249-291, 1976

RECONNAISSANCE DE PARTIES STABLES DE PAROLE CONTINUE POUR LE DECODAGE ACOUSTICO-PHONETIQUE.

Laurent MICLET* Dominique VICARD**

*ENST Paris - **ENST/TELIC-ALCATEL Strasbourg

ABSTRACT

We present in this paper an automatic segmentation and labelling module for the steady parts of continuous speech. The segmentation stage splits the signal into five classes, and the labelling stage uses vector quantization techniques as well as an original decision rule mixing the Bayes and K Nearest Neighbor criteria. The resulting (single speaker) recognition scores are very encouraging. This project is supported by grants from ANRT(+) and TELIC-ALCATEL.

INTRODUCTION

Le décodage acoustico-phonétique est une étape clef dans le processus de reconnaissance de la parole continue, et de sa qualité dépendent les futures performances des systèmes de reconnaissance [1]. Notre approche est purement analytique et notre but est de fournir à des systèmes plus évolués de traitement, utilisant des techniques d'intelligence artificielle, un treillis phonétique le plus "propre" possible. Le système complêt comportera au besoin des circuits intégrés spécifiques dédiés aux parties les plus coûteuses en calcul.

Le module présenté dans cet article s'occupe de la reconnaissance des parties stables d'un signal de parole continue. Les phonèmes reconnus sont définis comme ceux de la langue française. Il s'agit des voyelles, des fricatives dévoisées, ainsi que de certaines plosives voisées (b,d) , de nasales (m,n) et de (z). Ce module sera complété par un module de reconnaissance de parties transitoires, utilisant la programmation dynamique, et qui intégrera les informations de contexte délivrées par le premier module afin d'optimiser la reconnaissance.

(+) A.N.R.T.: Association Nationale de la Recherche Technique

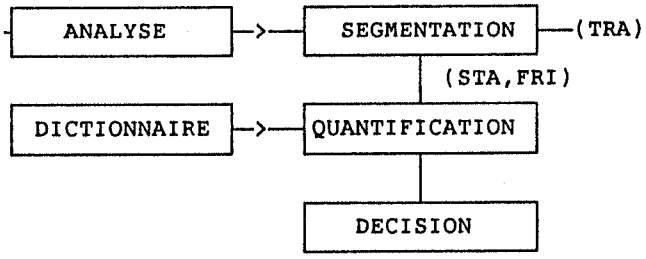


Schéma d'ensemble du decodeur.

ANALYSE

Le signal ,numerisé à 8 Khz et quantifié sur 12 bits, est analysé en LPC d'ordre 10 sur des fenètres de 128 échantillons recouvrantes à 50%. De cette analyse sont dérivés des PARCORs, dont les 3 premiers seront utilisés par l'algorithme de segmentation, ainsi que des coefficients cepstraux (LPCC). Ces conditions donnent des résultats satisfaisants, mais pourraient être adaptés en fonction des outils d'analyse disponibles. Pour tout le traitement ultérieur, nous utiliserons donc des vecteurs de 10 coefficients cepstraux normalisés (C0=0.0). Des essais ont cependant été menés avec d'autres représentations paramétriques.

SEGMENTATION.

Le rôle de cette partie segmentation est de séparer le signal en 5 classes représentant les parties stables (STA), les fricatives stables (FRI), les parties transitoires franches (TRA), les parties transitoires douces (STR) ainsi que le silence (SIL). Pour cela nous utilisons trois indices:

- La position des maxima locaux de l'énergie.
- Le taux de passages par zéro.
- Un indice de stabilité spectrale calculé en fonction des variations des 3 premiers PARCORS.

Ce dernier indice est calculé comme suit:

$$S_t = \sum_{i=1}^{3} D_i \quad \text{avec} \quad D_i = \sum_{j=-2}^{+2} (K_{i}^{t+j} - K_{i}^{t})$$

où K_i^t représente le i-ème PARCOR de la trame t

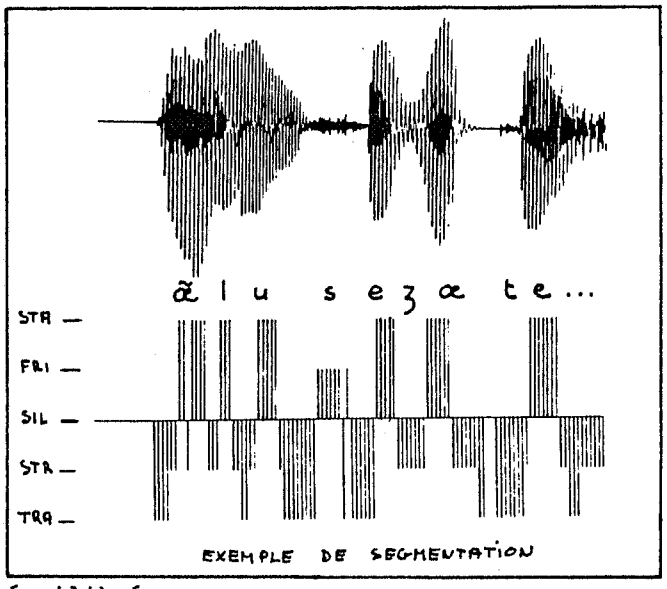
La détermination des zones spectralement stables se fait avec un simple seuillage avec hystérésis sur l'indice de stabilité. Cette décision est complétée par la détection des maxima d'énergie significatifs, ainsi que par une modification dynamique du seuil en fonction du taux de passage par zéro, autorisant ainsi une meilleure détection des fricatives. Les parties non stables sont alors elles-même coupées en deux classes, grâce également à un seuillage sur l'indice de stabilité. Enfin, la détection du silence se fait en fonction du niveau d'énergie sur 5 fenètres.

L'utilisation des maxima d'amplitude pour compléter la fonction de segmentation s'est avérée nécessaire afin de détecter des noyaux vocaliques brefs. La qualité de la reconnaissance est cependant moins bonne à ces endroits.

Cet algorithme de décision par seuillage (assez rustique) est suffisant pour la détection des zones stables lors de la phase d'apprentissage automatique. Il est cependant insuffisant pour assurer une grande fiabilité lors de la reconnaissance sur un signal inconnu, et les informations délivrées sont de fait utilisées avec réserve.

APPRENTISSAGE

Le dictionnaire de forme de référence est composé de 210 vecteurs (plusieurs tailles ont été testées). Ces vecteurs, représentant les parties stables, sont extraits d'un corpus d'apprentissage par une méthode de classification automatique [2] appelée "Boules optimisées" [3]. Ce corpus est lui-même extrait automatiquement par la fonction de segmentation d'un flot de parole continue issu de 30 phrases phonétiquement équilibrées.



EXEMPLE DE SEGMENTATION

Les Boules optimisées

Cet algorithme repose sur une agrégation en boules de rayon fixe dont le placement des centres est optimisé.

- Etape 1:
  - * Pour tous les vecteurs du corpus d'apprentissage:
  - * Comparer un nouveau vecteur V aux elements du dictionnaire: soit D la distance au plus proche vecteur VR.
  - * Si D < rayon alors affecter V à la classe de VR
  - * Sinon créer une nouvelle classe avec V pour représentant.
- Etape 2:
  - * Pour toutes les classes:
  - * Calculer le centre d'inertie de chaque classe; le barycentre devient le nouveau représentant.
- Etape 3:
  - * Recommencer 1 et 2 tant que des vecteurs changent de classe. Eliminer les classes vides à chaque étape.

Cet algorithme présente les avantages suivants:

- Souplesse dans le choix du nombre de représentant par ajustement du rayon.

- Respect des singularités du nuage.
- Quantification ultérieure avec une distance bornée par le rayon.

Par comparaison, des dictionnaires furent crées avec un simple algorithme à seuil [4]: outre la disparition de l'effêt de chaine, l'algorithme des boules optimisées réduit considérablement la distance moyenne des vecteurs du corpus d'apprentissage aux éléments du dictionnaire (test de distorsion).

Les classes ainsi obtenues se recouvrent dans l'espace des coefficients, mais il faut remarquer qu'il en est de même pour des classes dont le critère serait purement phonétique.

Corpus d'apprentissage

Le corpus d'apprentissage est phonétiquement équilibré [5], c'est-à-dire que les fréquences d'apparition des divers phonèmes sont celles de la langue française. Cette propriété permet une optimisation statistique implicite du taux de reconnaissance, favorisant par un plus grand nombre de représentants les phonèmes les plus courants (qui sont malheureusement souvent proches les uns des autres dans l'espace de représentation choisi).

Apprentissage phonétique

Les vecteurs délivrés par l'algorithme de classification sont anonymes et le plus souvent "non réels", car issus d'un calcul de moyenne. Pour assurer la correspondance entre ces vecteurs et les phonèmes à reconnaitre, nous procédons à une phase d'apprentissage phonétique.

Le corpus d'apprentissage des parties stables étiqueté manuellement est quantifié par le dictionnaire, et les répartitions des phonèmes sur ces vecteurs sont notées par simple comptage. Pour chaque vecteur V du dictionnaire, on dérive de ce comptage un vecteur PV de probabilités conditionnelles traduisant la distribution des phonèmes. C'est ce vecteur qui sera utilisé en phase de reconnaissance.

Le corpus d'apprentissage automatique peut-être différent de celui de l'apprentissage phonétique, ce qui permet d'envisager avec moins de craintes une adaptation au locuteur.

COMPARATEUR

Cette partie effectue un classique codage vectoriel des vecteurs de LPCC et utilise la distances de MINKOWSKY. Plusieurs distances ont été testées (Euclidienne, Minkowsky,Tchebycheff) pour cette parties, mais les différences ne sont pas significatives. L'avantage de cette métrique est la faible quantité de calculs nécessaires ainsi que sa simplicité (surtout pour une intégration sur VLSI).

Ce module délivre donc en sortie les 5 étiquettes du dictionnaire dont les représentants sont les plus proches du vecteur à reconnaitre, ainsi que les 5 distances associées.

DECISION

La phase de décision utilise à la fois une maximisation des probabilité, se rapprochant ainsi de la décision Bayesienne, et le critère des plus proches voisins. Cette méthode est assez proche de celle des K plus proches voisins pondérés [6].

Un vecteur inconnu V à reconnaitre est quantifié par le dictionnaire et les distances Dj aux 5 plus proches voisins Vj sont notées. On pondère ensuite les vecteurs de probabilité PVj par ces distances pour obtenir le vecteur de probabilités estimées PV:

$$PV = \frac{1}{\sum_{1}^{5}(1/Dj)} \sum_{1}^{5}(PVj/Dj)$$

La décision est ensuite prise en choisissant la probabilité marginale la plus forte (ou les 3 plus fortes).

RESULTATS

Cette méthode a été testée sur deux corpus, l'un (T2) contitué de 20 phrases phonétiquement équilibrées (représentant 2100 trames de parties stables) , l'autre (T1) d'une histoire célèbre ("La bise..."). Ces deux corpus n'ont bien entendu pas participé à l'apprentissage (constitué de 2400 vecteurs), et les parties stables à reconnaitre ne sont pas dans le même contexte phonétique. Le tableau ci-dessous donne les scores réalisés en confrontant l'étiquetage automatique à un étiquetage manuel.

Pour valider notre méthode, nous l'avons comparée à une décision par les K plus proches voisins choisis dans l'ensemble d'apprentissage. Les résultats obtenus donnent une idée de la difficulté des corpus choisis.

CHOIX	1	2	3
T1	48	63	70
T2	56	70	79

Résultats de reconnaissance trame par trame en premier, second et troisième choix (pourcentage).

K	1	5	20
T1	42	48	51
T2	55	58	60

Résultats sur les mêmes données de la décision par les K-PPV avec K=1,5, et 20 (premier choix)

Nous étudions actuellement d'autres critères de décision finale, basés sur la stabilité de l'étiquetage, qui permettront de consolider a posteriori les résultats de la segmentation.

Les informations issues de ce module seront principalement utilisées pour aiguiller efficacement une recherche sélective dans un dictionnaire de transitions, lesquelles seront traitées par un algorithme de comparaison dynamique continue (étude en cours).

REFERENCES

[1] M. DESI F. POIRIER
Le système SHERPA: étiquetage et classification automatique par apprentissage pour le décodage de la parole continue.
Thèse Orsay 1985.

[2] E. DIDAY J.C. SIMON
Clustering analysis.
Digital Pattern Recognition.
Springler Verlag 1976.

[3] G. FLAMENBAUM and Al.
Agrégation en boules de rayon fixe et centres optimisés.
Les cahiers de l'analyse de données Vol. IV 1979

[4] M. DABOUZ
Transmission de la parole a faible débit par vocodeur a classification.
These de l'ENST Janvier 1984.

[5] P. COMBESCURE
Phrases phonétiquement équilibrée.
Recherches Acoustique Vol VI
CNET Lannion.

[6] A. M. DILL
Extensions de la méthode des plus proches voisins.
Thèse de l'Université CLAUDE-BERNARD (Lyon) 1978

RECONNAISSANCE DE LA PAROLE MULTI-LOCUTEUR PAR PROGRAMMATION DYNAMIQUE

Anne BOYER, Joseph di MARTINO, Jean-Paul HATON

Equipe Reconnaissance des Formes et Intelligence Artificielle
Centre de Recherche en Informatique de Nancy
Université de Nancy 1
Campus Scientifique BP 239
54506 Vandoeuvre-les-Nancy

ABSTRACT

This paper describes a method for speaker-independent isolated word recognition using vector quantization, dynamic time warping, and which is related with methods using hidden Markov models.

1. INTRODUCTION

Les techniques de programmation dynamique sont largement utilisées dans des systèmes de reconnaissance à vocabulaire limité, à la fois pour des applications de reconnaissance de mots isolés et de mots enchaînés. Ces méthodes de reconnaissance donnent d'excellents résultats, avec un taux d'erreur voisin de 0 en monolocuteur lorsque l'on se place dans des conditions expérimentales bien précises:
1-hauteur de la voix stable
2- bruit ambiant pas très élevé
3-vocabulaire de difficulté moyenne (e.g.: les chiffres)

Lorsqu'une de ces conditions n'est pas respectée, on s'aperçoit que les taux de reconnaissance s'écroulent. Par exemple, lorsque l'on passe du vocabulaire des chiffres au vocabulaire des lettres de l'alphabet, le taux d'erreur est multiplié par un facteur 10 environ.

A l'heure actuelle, on s'accorde à dire que les techniques de reconnaissance de mots isolés sont pratiquement au point. Quant-à nous, nous pensons au contraire qu'il reste encore beaucoup de travail à réaliser afin d'éliminer les faiblesses des systèmes actuels.

Les problèmes que l'on rencontre en reconnaissance de mots isolés sont encore plus accentués lorsque l'on se place dans le cas où plusieurs locuteurs peuvent inter-agir avec le système de reconnaissance -on parle alors de reconnaissance multi ou pluri-locuteur-. Si ce dernier a été conçu pour fonctionner avec un seul locuteur, les taux de reconnaissance que l'on obtient lorsque l'on teste le système de reconnaissance avec 2 locuteurs -l'apprentissage ayant été effectué avec un seul des deux locuteurs- sont tout simplement déplorables: de l'ordre de 70% dans le meilleur des cas. On améliore sensiblement les choses si au cours de la phase d'apprentissage des formes vocales relatives aux deux locuteurs sont utilisées pour créer le vocabulaire de référence. Mais cette méthode n'est guère utilisable en pratique car le nombre de formes de référence devient rapidement trop important -pour ce type de systèmes- lorsque le nombre de locuteurs croît. D' autres méthodes plus sophistiquées, faisant appel à la classification automatique des données [1], ont été utilisées pour pouvoir compenser les variabilités inter-locuteur. Ces techniques fonctionnent relativement bien dans le cas de vocabulaires pas trop difficiles. Leur inconvénient majeur est qu'elles nécessitent un accroissement de l'espace mémoire des données par un facteur 12 environ en prenant comme référence l'espace mémoire des données d'un système mono-locuteur de reconnaissance de mots isolés et aussi, par voie de conséquence, un accroissement du temps de calcul par le même facteur.

Le but de ce papier est d'introduire une nouvelle méthode de reconnaissance de mots isolés multi-locuteur qui s'apparente aux méthodes de reconnaissance par modèles de Markov cachés proposées récemment par Levinson [2]. Mais avant d'entrer dans le vif du sujet, il est nécessaire de faire quelques rappels sur les techniques de programmation dynamique avec et sans relâchement des contraintes aux frontières.

2. QUELQUES RAPPELS SUR LA PROGRAMMATION DYNAMIQUE

Soient T(i) et R(j) deux formes vocales où T(i) désigne le ieme vecteur de T et R(j) le jeme vecteur de R. Le but de la programmation dynamique est de mettre en correspondance les échelles des temps de R et T afin de compenser les distorsions dues à la variabilité de l'élocution. Pour cela il

faut rechercher dans le plan de comparaison, visualisé par la figure 1, le chemin optimal de recalage.

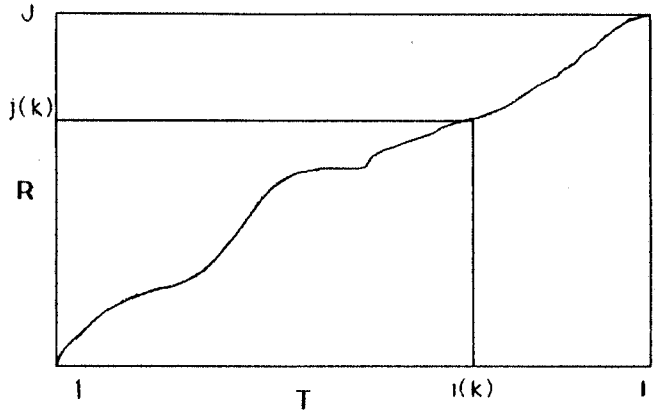


Figure 1- Un chemin de recalage. Le point (i(k), j(k)) correspond à la mise en coincidence entre le i(k) ieme prélèvement de la forme T et le j(k) ieme prélèvement de la forme R.

Celui-ci peut être obtenu en évaluant de manière itérative les relations suivantes:

$$D(i,j) = MIN \{ D(i-1,j)+d(i,j), \quad (1\text{-}1)$$
$$D(i-1,j-1)+2*d(i,j), \quad (1\text{-}2) \quad (1)$$
$$D(i,j-1)+d(i,j)\} \quad (1\text{-}3)$$

où D(i,j) est la distance cumulée associée au chemin de recalage partiel optimal aboutissant au point (i,j) et d(i,j) est la distance locale entre le ieme prélèvement de la forme T et le j eme prélèvement de la forme R, et en gardant pour chaque point (i,j) le point (i*,j*) = P(i,j) qui a permis de minimiser (1) e.g.: D(i,j) = D(i*,j*) + d(i,j) si (1-1) est l'expression qui a permis de minimiser (1).

Le processus de comparaison se termine lorsque D(i,j) a été évaluée pour tout i de 1 à I et pour tout j de 1 à J, où I est le nombre total de prélèvements de la forme T et J celui de la forme R. Le chemin de recalage optimal est alors obtenu par la suite de points suivante: (I,J), P(I,J), P(P(I,J)),.....,(1,1) (2)

3. LES RELATIONS DE PROGRAMMATION DYNAMIQUE AVEC RELACHEMENT AUX CONTRAINTES.

Afin de pouvoir compenser les erreurs de détection parole-non parole, il est intéresant de pouvoir relâcher les contraintes aux frontières,i.e. ne plus obliger les chemins de recalage à débuter au point (1,1) et à se terminer au point (I,J). Voir figure 2

Les chemins de recalage doivent être autorisés à débuter dans un voisinage du point (1,1) et à se terminer dans un voisinage du point (I,J). Or si l'on effectue ce relâchement aux frontières, il est

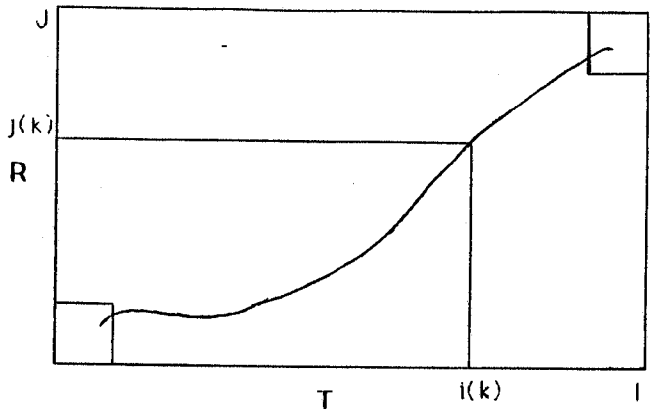


Figure 2. Un exemple de chemin de recalage avec relâchement aux frontières.

aisé de montrer que les relations de programmation dynamique (1) ne sont plus applicables. Ceci est dû au fait que la propriété de constance de la longueur des chemins de recalage en un point quelconque du plan de comparaison n'est plus respectée. Afin de pallier cette difficulté, nous avons introduit de nouvelles relations de programmation dynamique [3] qui opèrent dans une première phase une normalisation des distances cumulées par la longueur des chemins de recalage associés. Celles-ci peuvent être exprimées a l'aide des relations suivantes:

$$l\text{-}d = MIN\{(D(i-1,j)+a*d(i,j))/(L(i-1,j)+a) = d1 \quad (3)$$
$$(D(i-1,j-1)+b*d(i,j))/(L(i-1,j-1)+b) = d2$$
$$(D(i,j-1)+a*d(i,j))/(L(i,j-1)+c) = d3\}$$

2-Si d=d1 alors D(i,j) = D(i-1,j) + a*d(i,j)
L(i,j) = L(i-1,j) + a
fsi
Si d=d2 alors D(i,j) = D(i-1,j-1) + b*d(i,j)
L(i,j) = L(i-1,j-1) + b
fsi
Si d=d3 alors D(i,j) = D(i,j-1) + c*d(i,j)
L(i,j) = L(i,j-1) + c
fsi

où L(i,j) est la longueur du chemin de recalage optimal aboutissant au point (i,j).

L' intérêt des relations (3) est qu'elles permettent de rendre l' algorithme de programmation dynamique paramétrable. Ainsi les paramètres (a,b,c) peuvent être optimisés pour chacun des mots du vocabulaire en rendant, par exemple, minimal le taux de distorsion relatif à un ensemble de formes de références, prononcées par un ou plusieurs locuteurs.

Le deuxième intérêt de ces relations est que les paramètres (a,b,c) peuvent être considérés comme étant des paramètres similaires aux probabilités de transition des modèles de Markov. La figure (3) visualise les probabilités de transition associées aux arcs d'un modèle de Markov.

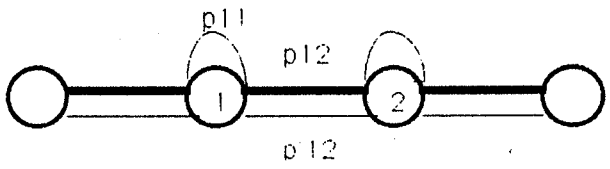


Figure 3- Un automate partiel associé à un modèle de Markov.
-p11 est la probabilité de transition de l'état 1 vers lui même avec émission d'un prélèvement.
-p12 est la probabilité de transition de l'état 1 vers l'état 2 avec émission d'un prélèvement.
-p'12 est la probabilité de transition de l'état 1 vers l'état 2 sans émission de prélèvement.

Les paramètres (a,b,c) peuvent être considérés comme étant l'équivalent des probabilités p11,p12 et p'12 respectivement, des modèles de Markov.

Il est intéressant de constater que la transition de l'état 1 à l'état 2 de probabilité p'12 sans émission de prélèvement n'a guère été utilisée jusqu'à présent dans les modèles de Markov cachés d'une part du fait que le lien avec les algorithmes de programmation dynamique classiques n'a jamais été clairement mis en évidence et d'autre part du fait que l'algorithme de Viterbi -qui est l'algorithme de programmation dynamique communément utilisé dans le cas des modèles de Markov- ne le permet pas.

Ainsi, comme nous venons de le voir, les paramètres (a,b,c) peuvent être associés à un état d'un modèle de Markov -il y a dans ce cas autant d'états que de triplets - qui lui même peut être associé à une liste de vecteurs prototypes -c' est l'approche communément adoptée- ou à un couple de vecteurs prototypes. Nous avons retenu cette dernière solution afin de modéliser plus convenablement la structure temporelle des formes vocales, structure temporelle qui est perdue lorsque l'on associe une liste de prototypes à un état d'un modèle de Markov.

4. L' ALGORITHME DE RECONNAISSANCE DE MOTS ISOLES MULTI-LOCUTEUR.

4.1. LA PHASE D' APPRENTISSAGE

Pour chaque mot du vocabulaire qui a été prononcé plusieurs fois par plusieurs locuteurs, nous déterminons un ensemble de vecteurs prototypes par quantification vectorielle à l'aide de l'algorithme de Linde et Buzo [4]. Les formes de référence utilisées dans cette étape sont extraites d'un corpus d'apprentissage.

Chaque forme de référence associée à un mot du vocabulaire est comparée par programmation dynamique à toutes les autres formes de référence correspondant à ce même mot. Pour chaque couple de prototypes -concrétisé par un point d'un chemin de recalage - nous évaluons le nombre des déplacements locaux horizontaux, verticaux et en diagonale qui ont provoqué le passage du chemin optimal par le couple en question. L' inverse des fréquences de ces déplacements locaux nous permet de déterminer les coefficients (a,b,c) associés à chaque couple de prototype.

4.2. LA PHASE DE RECONNAISSANCE

Le mot inconnu est codé à l'aide des différents codeurs vectoriels, associés à chaque mot du vocabulaire, qui ont été déterminés au cours de la phase d'apprentissage. Les différentes formes codées obtenues sont comparées par programmation dynamique à un ensemble de formes de références, codées par le même quantifieur vectoriel que celui qui a été utilisé pour coder la forme inconnue, qui représentent le mieux la classe des formes de référence associée au mot du vocabulaire dont les références ont permis de construire le quantifieur vectoriel en question. Le mot reconnu est le mot associé au quantifieur vectoriel qui a permis d' obtenir le score de programmation dynamique minimal. Les résultats obtenus par cette méthode seront commentés lors de l'exposé.

5. REFERENCES

[1] L.R. RABINER, S.E. LEVINSON, A.E. ROSENBERG, J.G. WILPON: "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques", IEEE Trans. Acoust. Speech Signal Processing, ASSP-27, pp.336-349, 1979.

[2] S.E. LEVINSON, L.R. RABINER, M.M. SONDHI: "Speaker Independent Isolated Digit Recognition Using Hidden Markov Models", ICASSP 83, pp.1049-1052,1983.

[3] J. di MARTINO: "Dynamic Time Warping Algorithms for Isolated and Connected Word Recognition", in Nato Asi Series, Vol F16, New Systems and Architectures for Automatic Speech Recognition and Synthesis, R. DE MORI, C.Y. SUEN, Springer-Verlag Heidelberg, 1985.

[4] Y. LINDE, A.BUZO, R.M. GRAY: "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications COM-28, pp. 84-95, January 1980.

UTILISATION DE LA QUANTIFICATION VECTORIELLE EN RECONNAISSANCE DE LA PAROLE CONTINUE

SU Huan-Yu

IRISA, Campus de Beaulieu, 35042 RENNES CEDEX

ABSTRACT

The representation of speech signals and the distance measure between two frames are essential for the application of the vector quantization on speech recognition by phoneme. Several such problems and a continuous speech phonetic decoder are discussed in this paper.

The signals are represented by a smoothed version of their spectrum, which is considered as a vector in $\mathbb{R}^N$. After grouping a training sequence into sub-sets corresponding to each phoneme, the method, "Splitting" is used to obtain a sub-dictionary ; the reunion of these sub-dictionaries gives our global dictionary for the phonetic recognition. Each incoming vector is superseded by its nearest neighbours in the dictionary, in order to be recognized.

This method is easy to apply. It gives very good results for the classification and recognition of phonemes. It can thus be used as a primary step for a more advanced recognition process.

I Introduction

Depuis quelques années, nombre d'études [1]-[3] ont été menées sur l'utilisation de la quantification vectorielle en traitement de la parole. La méthode de "Splitting"[1] que nous allons étudier, bien que peu utilisée, fournit un excellent procédé d'obtention d'un dictionnaire (ici, un ensemble fini de $\mathbb{R}^N$, N=256), pour la reconnaissance de la parole.

La méthode de reconnaissance d'un signal de parole inconnu se décompose alors en deux temps : une représentation par un vecteur de $\mathbb{R}^N$ au moyen de certains prétraitements, puis une quantification qui consiste à remplacer ce vecteur par son plus proche voisin dans le dictionnaire.

La qualité d'un système de reconnaissance dépend donc essentiellement du dictionnaire utilisé et de la distance choisie sur $\mathbb{R}^N$.

Nous proposons une construction de ce dictionnaire par réunion de sous-dictionnaires, chacun d'eux correspondant à un phonème. Deux distances différentes sont utilisées : euclidienne pour la construction du dictionnaire, alors que celle pour la reconnaissance ne l'est pas, pour des raisons perceptives.

II Espace de représentation

Il existe plusieurs représentations pour le signal de parole ; par exemple : le spectre, le spectre lissé (dans le domaine fréquentiel) et le cepstre (dans le domaine temporel). Nous avons choisi le spectre lissé, qui semble être la meilleure approche pour garder la forme perceptive des sons.

Pour un son voisé, la contribution essentielle à la perception se situe au voisinage des formants, et pour des fréquences inférieures à 4 KHz ; par contre, pour un son non-voisé (sans formants), toute la bande spectrale (0-6,4 KHz, pour un signal échantillonné à 12,8 KHz) reste utile.

Une fenêtre temporelle (Hamming) de 40 ms,est placée sur le signal (512 échantillons). Le spectre lissé est obtenu par une FFT sur les 18 premiers coefficients cepstraux.

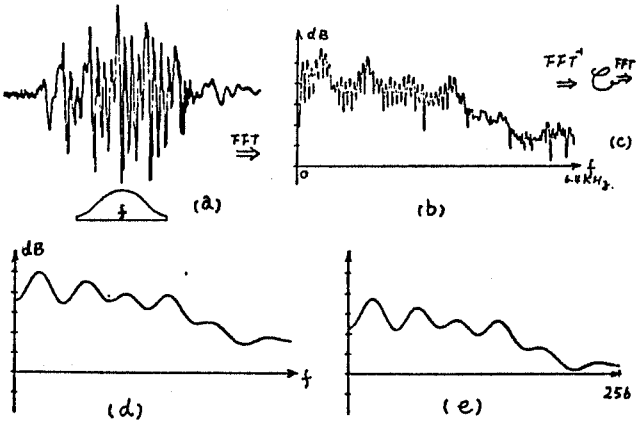


Figure 1

(a) le signal de la parole
(b) son spectre
(c) ses coefficients cepstraux
(d) son spectre lissé
(d) son spectre lissé pondéré

Chaque son (fig.1 a) est donc représenté, dans le domaine fréquentiel, par une courbe (256 points, fig. 1 d) que nous recalons ensuite sur son minimum (fig. 1 e). De plus, pour les sons voisés, le filtre fréquentiel suivant :

$$fi = \begin{cases} 1 & \text{pour } i = 1,170 \\ e^{-\frac{i-170}{20}} & \text{pour } i = 171,256 \end{cases}$$

est employé pour diminuer l'influence néfaste de

la partie hautes-fréquences.

III Dictionnaire

1. Méthode de "Splitting" [1]

A partir d'un ensemble d'apprentissage ($\subset R^N$) la méthode de "Splitting", associée à l'algorithme de Lloyd, a pour but de construire récursivement des dictionnaires de taille de plus en plus importante.

Elle se décompose de la manière suivante :

i) calcul du centre de gravité X_o de l'ensemble d'apprentissage et définition d'un dictionnaire initial, $D^\circ = \{X_o-\varepsilon, X_o+\varepsilon\}$, n=2, à partir d'une perturbation.

ii) application de l'Algorithme de Lloyd pour obtenir un dictionnaire Dn optimal à partir de D°. Une distorsion moyenne est obtenue pour chaque nouvelle classe ;

iii) définition d'un nouveau dictionnaire initial D°_{n+1}, à partir de D_n, en perturbant la classe de distorsion maximale ;

iv) itération de ii) et iii) jusqu'à obtention de la taille finale du dictionnaire recherché (fig. 2).

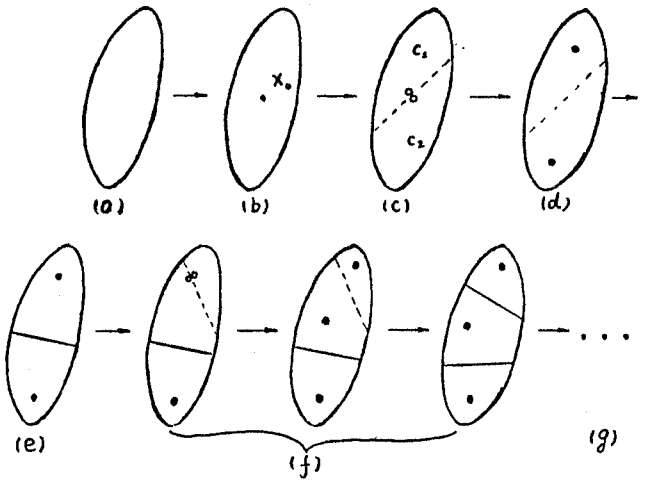


Figure 2

a) ensemble d'apprentissage ;
b) X_o : centre de gravité ;
c) perturbation autour de X_o qui partage l'espace en deux classes C_1 et C_2 ;
d) calcul du nouveau centre de gravité de chaque classe, D_n ;
e) partage de l'ensemble en deux sous-ensembles ;
f) refaire (c) (d) (e) pour le sous-ensemble dont la distorsion est la plus élevée ;
g) répéter (f), jusqu'à obtention du dictionnaire final.

2. Méthode d'obtention du dictionnaire

i) Données d'apprentissage

Les sons sont représentés par leur spectre lissé pondéré; nous disons qu'un spectre est représentatif s'il reste inchangé lors du décalage de la fenêtre de quelques échantillons. Nous ne traitons donc que les parties stables des phonèmes.

Les fenêtres sont placées manuellement afin de conserver le maximum d'information contextuelle. Sur la figure 3, le phonème /a/ est décomposé en trois segments :

- le 1er est influencé par /s/
- le 2ème correspond à la partie stationnaire du phonème
- le 3ème est influencé par /v/

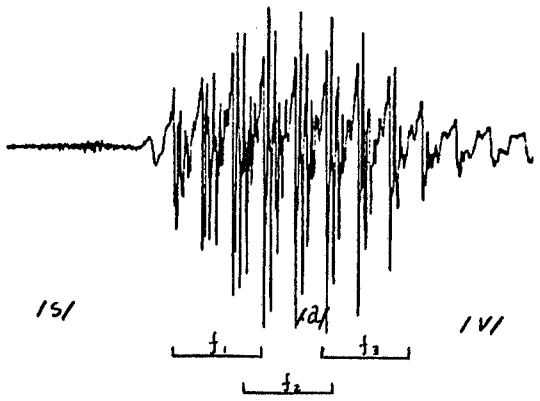


Figure 3 le phonème /a/ entre /s/ et /v/; trois segments sont pris pour l'apprentissage.

Nous avons utilisé vingt phrases, phonétiquement équilibrées, prononcées par un seul locuteur. Nous obtenons 750 segments d'apprentissage pour 21 phonèmes /a,ɑ̃,w,y,i,j,ə,e,ɛ,m,n,bdg (base de voisement de b.d.g.), u,o,ɔ,v,l,r (bruité), s,ʃ,f/.

ii) Dictionnaire

Les 750 vecteurs sont donc regroupés en 21 sous-ensembles correspondant à chacun des phonèmes. La méthode de "Splitting" permet d'extraire un sous-dictionnaire pour chacun d'eux. La taille de chaque sous-dictionnaire est de 4 à 10 classes, selon l'influence du contexte (quantifiée par la valeur de distorsion moyenne), et du cardinal du sous-ensemble d'apprentissage, correspondant.

Pour diminuer l'influence des erreurs d'apprentissage, dues à une mauvaise segmentation manuelle ou à une "mauvaise" prononciation, nous imposons à chaque classe de contenir un minimum de 2 vecteurs, refusant ainsi les classes singulières (fig. 4).

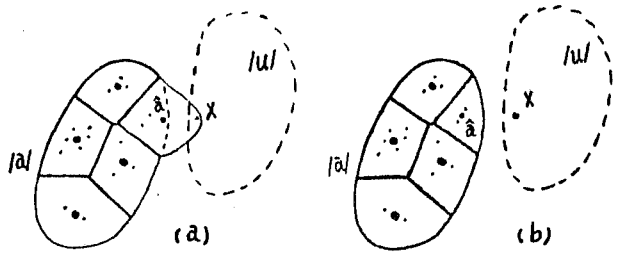


Figure 4

(a) un vecteur x, appartenant à l'ensemble /u/ et attribué par erreur à l'ensemble /a/, aura pour simple effet de déplacer le centre â ;
(b) par contre, si nous gardons les classes singulières, x sera un représentant erroné de /a/ dans le dictionnaire.

Le dictionnaire global est obtenu par simple réunion de ces sous-dictionnaires ; sa taille est

de 147 classes.

IV Reconnaissance

i) Segmentation

Il s'agit de déterminer quels segments nous désirons reconnaître, et de quelle façon nous allons positionner la fenêtre, à l'intérieur de ce segment. Dans un premier temps, nous employons, comme pour l'apprentissage, la localisation manuelle des fenêtres sur les parties stables des phonèmes. Pour une deuxième série d'expérimentations, nous utilisons une segmentation automatique de parole continue [4]. Nous éliminons les segments trop courts (<40 ms) et les segments correspondants au silence, par un seuil sur l'énergie. Nous plaçons la fenêtre au milieu de chaque segment.

ii) Distance

Lors de la construction du dictionnaire, la distance employée est la distance euclidienne : la méthode de "Splitting" est appliquée pour chaque phonème et il est normal de considérer qu'à l'intérieur d'une même famille phonèmique, les coordonnées (éventuellement pondérées) de tout vecteur sont d'égale importance.

Par contre, pour la reconnaissance, nous devons comparer un vecteur à des représentants des phonèmes différents. Pour cela, une distance non euclidienne est utilisée :

$$D(x,\hat{x}) = \sum_{i=1}^{N} (x_i - \hat{x}_i)^2 \, P_i(x,\hat{x})$$

$$\text{avec } P_i(x,\hat{x}) = \begin{cases} \dfrac{x_i^2 \hat{x}_i^2}{S\hat{S}} & \text{pour les sons voisés} \\ 1 & \text{pour les sons non-voisés} \end{cases}$$

$$\text{où } S = \sum_{i=1}^{N} x_i^2 \quad \text{et} \quad \hat{S} = \sum_{i=1}^{N} \hat{x}_i^2$$

Pour les sons voisés, le poids $P_i(x,\hat{x})$ a pour effet d'augmenter l'importance des formants, et de tenir compte de leur placement ; pour les sons non-voisés, sans structure formantique, les coordonnées restent équi-pondérées.

iii) Le plus proche voisin

Pour chaque vecteur inconnu, nous testons d'abord son voisement (par un seuil sur la troisième composante), pour obtenir la pondération P. Ensuite, nous cherchons ses deux plus proches voisins dans le dictionnaire. Le vecteur est alors représenté soit par son plus proche voisin (critère du plus proche voisin), soit par ses deux plus proches voisins, suivant l'application future.

V Résultats

Nous avons utilisé une liste de 10 phrases phonétiquement équilibrées, prononcées par le même locuteur, mais n'ayant pas été utilisées pour l'apprentissage de la construction du dictionnaire. Nous avons ainsi 395 segments de signal stable (segmentés à la main) et 21 phonèmes représentés. Le taux de reconnaissance est de 79,5 % d'après le critère du plus proche voisin, et 88,1 %, d'après celui des deux plus proches voisins. Lorsque les différences entre /w/ et /a/, /e/ et /ɛ/, /i/ et /j/, sont ignorées, les taux sont respectivement de 82 % et 92 %.

Nous avons 25 % et 38 % d'erreurs provoquées par :

- des phonèmes voisins :
  - /m/ et /n/
  - /a/ et la partie orale de /ã/
  - /u/ et /w/
- des phonèmes contextuels :
  - /a/ en contextes /ka/, /pa/ est reconnu comme /ɛ/ et inversement

et 15 % et 23 % d'erreurs sur la reconnaissance des /v/, /l/ et /r/, les phonèmes courts.

VI Conclusion

Nous avons proposé une procédure d'obtention du dictionnaire global par réunion des sous-dictionnaires correspondant à chacun des phonèmes. L'avantage de la méthode de "Splitting" et de la distance pondérée nous permet d'obtenir un dictionnaire optimal, et une meilleure comparaison pour trouver les plus proches voisins de chaque vecteur à coder.

Cette méthode pourra aussi être exploitée comme première approche dans d'autres domaines de classification et de reconnaissance des états physiques.

VII Références

[1] Robert M. GRAY
Vector Quantization, IEEE ASSP Magazine 1984.

[2] Y. LINDE, A. BUZO and R.M. GRAY
"An algorithm for Vector Quantizer Design",
IEEE Trans. on comm. signal processing,
January 1980.

[3] R.M. GRAY, A. BUZO, A.H. GRAY, J.R. and Y. MATSYYAMA
"Distorsion Measures for Speech Processing",
IEEE Trans. on Acoustic, Speech, and signal processing, Vol. ASSP-28 N° 4 August 1980.

[4] R. ANDRE
"Segmentation automatique du signal de parole, sans reconnaissance". Thèse de 3ème cycle - Rennes, 1985.

RECONNAISSANCE MULTILOCUTEUR DES CHIFFRES FRANCAIS PAR ASSOCIATION D'UN PRETRAITEMENT FONDE SUR LA QV AVEC DES MODELES DE MARKOV CACHES.

Alain TASSY

MATRA sa., 3 av. du Centre
78182 Saint-Quentin Yvelines Cedex
FRANCE

Laurent MICLET

E N S T, Dpt. SYC, UA CNRS 820
46 rue Barrault,75634 Paris cedex 13
FRANCE

ABSTRACT

Vector Quantization has recently been used in the realization of a VQ speaker-independent digit recognizer, based uniquely on the spectral content of the speech signals. On the other hand, the Hidden Markov Models proved their ability in modelling temporal distortions between different utterances of a word pronounced by several speakers. In term of recognition rate, HMMs are as efficient as the conventional DTW matching, but they need less computation and memory. This paper presents a speaker-independent digit recognition system that combines word-based VQ with HMM, the cost of which is low enough to be implemented on a single signal processor available today. It is the first result of a cooperation project between ENST and the MATRA company, financially supported by the ANRT

INTRODUCTION

Beaucoup d'applications pratiques demandent l'utilisation de la reconnaissance de parole multilocuteur. Pour la plupart d'entre elles, le vocabulaire à reconnaître est assez restreint et se limite souvent aux dix chiffres, plus quelques mots de commande. Depuis le début des années 80, la technique de la programmation dynamique avec classification des références (Levinson [1]) a donné de bons taux de reconnaissance en multilocuteur. Le temps de calcul et la mémoire nécessaires à cette méthode la rende difficile à mettre en oeuvre, en temps réel, sur un procésseur bas de gamme.

Le principe de compression de données appelé Quantification Vectorielle (QV)[2] a souvent été utilisé afin de réduire le coût mémoire des systèmes de reconnaissance de parole [3]. Cependant, la Quantification Vectorielle est plus qu'une simple technique de codage, comme le prouvent Shore et Burton [4]. D'une séquence d'apprentissage de plusieurs répétitions de chaque mot prononcé par un grand nombre de locuteurs, ils extraient un dictionnaire de vecteurs de coefficients LPC propre au mot du vocabulaire considéré. Dans la phase de reconnaissance, les mots isolés sont rangés selon la valeur de la distorsion moyenne qui résulte de leur codage avec les différents dictionnaires de Quantification Vectorielle. Ce système fondé uniquement sur l'analyse spectrale du mot à reconnaître présente une lacune: les informations temporelles du signal de parole sont perdues. Contre cet inconvénient, Pan & Al. [5] proposent d'utiliser cette technique pour créer un préprocesseur qui élimine les candidats peu probables et limite ainsi le nombre de références à traiter pour réaliser l'alignement temporel.

D'autre part, les Modèles de Markov Cachés (MMC) semblent s'adapter à une vaste gamme de systèmes de reconnaissance [6], [7]. Rabiner & Al. [3] proposent un algorithme de reconnaissance multilocuteur, moins coûteux en temps calcul et mémoire que le DTW, au prix d'une légère baisse de performance. Le premier étage du système exécute une Quantification Vectorielle traditionnelle des mots isolés et donne une représentation symbolique discrète du codage à prédiction linéaire (LPC). Le deuxième étage est constitué d'une série de modèles de Markov, un pour chaque mot du vocabulaire, très efficaces pour modéliser la distorsion temporelle du signal de parole. Pour accroître les performances de ces modèles, il est intéressant de présumer que les paramètres des Modèles de Markov Cachés ont des distributions gaussiennes et d'utiliser des fonctions probabilistes à densité continue. Cependant, il semble que cette technique augmente beaucoup le coût du système [8].

Dans cet article, nous présentons les premiers résultats d'un projet d'implantation d'un système de reconnaissance multilocuteur de chiffres sur un seul processeur de signal (du type TMS320-10). Dans un premier temps, pour s'assurer des possibilités de mise en oeuvre, nous avons conçu un système sous-optimal, très bon marché. Par la suite, nous devrons accroître les performances, au moyen de solutions plus complexes proposées dans la littérature.

La première partie présente un processeur à Quantification Vectorielle qui utilise un dictionnaire de vecteurs spécifiques à chaque mot du vocabulaire. Son étude fut guidée par des concepts similaires à ceux des Bell Labs, notre système étant toutefois complétement différent et plus simple. En effet, la Quantification Vectorielle travaille directement sur des coefficients d'autocorrélation; nous utilisons une distance de Minkowsky et nous créons un nombre variable de classes pour chaque mot. Au lieu d'essayer d'améliorer le taux de reconnaissance, nous avons plutôt analysé le procédé de Shore et Burton sur les chiffres français. Différents critères d'élimination des candidats donnant trop de distorsion ont été expérimentés.

Le préprocesseur produit des segments, résultats de la Quantification Vectorielle d'une occurrence d'un mot avec les dictionnaires retenus. La deuxième partie débat des possibilités d'utiliser les Modèles de Markov Cachés sur ces segments pour profiter des informations temporelles contenues dans le signal, sans avoir à recalculer de nouvelles distances.

La troisième partie présente nos premières expériences effectuées sur de petites bases de données. Jusqu'à présent, la méthode fut testée avec un nombre limité de locuteurs, c'est pourquoi nous n'avons pas pu réellement optimiser ni valider la totalité du système de reconnaissance.

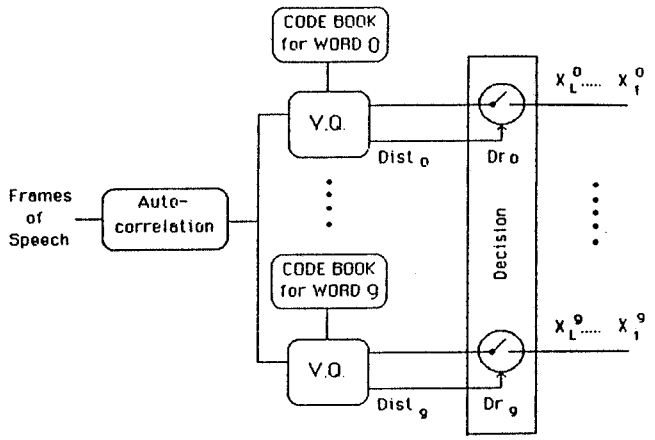


Fig. 1 : Préprocesseur fondé sur la QV

LE PREPROCESSEUR

Beaucoup de systèmes de reconnaissance, et en particulier ceux construits autour d'un processeur de signal, calculent les coefficients d'autocorrelation du signal de parole pour extraire ensuite une représentation spectrale . Ainsi, pour tester les limites de la méthode de Burton et dans l'optique de proposer un préprocesseur plutôt qu'un système complet, nous avons choisi d'effectuer la QV sur des vecteurs d'autocorrélation.

Apprentissage

Pour trouver des dictionnaires, il faut d'abord définir une métrique sur l'espace des vecteurs. Pour des raisons de simplicité, nous utilisons la distance de Minkovsky définie par:

$$d = \sum_{i=1}^{i=10} | R_i - R'_i |$$

Chaque fenêtre temporelle de 20 ms de chaque répétition d'un même mot du vocabulaire est représentée par le vecteur X des coefficients d'autocorrélation du signal échantillonné à 8 kHz. Au fil du calcul de ces X, on exécute l'algorithme à seuil déjà exposé et étudié dans [9]. On crée ainsi un dictionnaire par mot du vocabulaire, chaque dictionnaire contenant un nombre de vecteurs variable [10].

Reconnaissance

En phase de reconnaissance, le préprocesseur réalise le codage du signal de parole sur les différents vecteurs et ne transmet que des chaînes de symboles. Un premier classement des candidats à la reconnaissance est alors possible en utilisant la distorsion moyenne définie par:

$$Dist_j = \frac{1}{N} \sum_{n=1}^{n=N} d(R_n, X_n^j)$$

où R_n représente l'autocorrélation de la n^{ieme} fenêtre et X_n^j est le représentant de la classe de R_n dans le dictionnaire j. Différentes stratégies sont possibles pour éliminer les candidats les moins probables, elles ont en partie été exposées dans [10]. Pour bâtir un premier système, nous avons utilisé une décision des K meilleurs candidats. Quel que soit le résultat du codage, le préprocesseur transmet un nombre fixe de chaînes de symboles. Une autre décision a été proposée par les Bell Labs pour un préprocesseur du même type mais beaucoup plus complexe[5].

Un des grands avantages d'un préprocesseur fondé sur la QV est qu'il transforme le signal de parole en chaînes de symboles sur lesquelles il est possible d'appliquer un grand nombre d'algorithmes traditionnels de la reconnaissance des formes. En fournissant des données d'un haut niveau d'abstraction, puisque les symboles constituant les chaînes n'ont aucun sens intrinsèque mais représentent un rang dans un dictionnaire, ce type de préprocesseur peut constituer un pont entre le traitement du signal et la reconnaissance structurelle de formes ou l'intelligence artificielle. Il faut, dans ce sens, remarquer que le postprocesseur constituant notre système a été programmé en Lisp (Le_Lisp de l'INRIA).

MODELES DE MARKOV CACHES

Etant passé du signal de parole à des symboles abstraits grâce au préprocesseur, nous avons à notre disposition un grand nombre de méthodes de reconnaissance des formes pour prendre la décision finale. Une approche structurelle traditionnelle suppose que les chaînes de symboles à traiter sont produites par une grammaire régulière. De plus, il est possible d'introduire des connaissances statistiques sur les productions et de représenter ce type de grammaire sous forme d'un automate d'états finis. On obtient alors um modèle stochastique de production. Un type particulier de ces modèles a été récemment utilisé pour faire de la reconnaissance de la parole sous l'appellation de Modèles de Markov Cachés (MMC).

Appliquée à la reconnaissance de mots isolés indépendante du locuteur, cette technique repose sur la représentation des différentes façons de prononcer chaque mot du vocabulaire à reconnaître par un modèle optimal. Les chaînes de symboles produites correspondent à des séquences d'états du modèle. Les paramètres de chaque modèle sont déterminés pour représenter au mieux les différentes caractéristiques du mot du vocabulaire auquel ils correspondent. Ces paramètres sont:

- Nombres d'états
- A matrice de transition
- B matrice de production des symboles

Après avoir effectué l'apprentissage, sur lequel nous reviendrons, on peut calculer la probabilité qu'une chaîne de symboles inconnue soit produite par les différents modèles. La décision finale de reconnaissance consistera à exhiber le mot du vocabulaire correspondant au modèle donnant la probabilité maximale. C'est-à-dire, trouver N tel que:

$$P(W/HMM_N) = \max_{j=0,9} P(W/HMM_j)$$

Pour simplifier les calculs, on ne détermine pas réellement cette probabilité. On se contente souvent de trouver le chemin optimum T à l'intérieur de chaque modèle j permettant de produire les observations considérées, et l'on effectue la décision sur les probabilités de ces chemins. La décision consiste alors à trouver N tel que:

$$P(T_N(W)) = \max_{j=0,9} P(T_j(W))$$

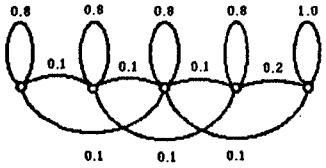


Fig. 2 : Modèle de Markov Caché initial

Apprentissage des modèles

Les différents paramètres d'un modèle sont estimés en utilisant un ensemble de chaînes de symboles représentant différentes élocutions du chiffre correspondant au modèle, prononcé par un grand nombre de locuteurs. En pratique, le nombre d'états de chaque modèle a été fixé arbitrairement entre 0 et 10; mais il n'y a aucune raison théorique de choisir un même nombre d'états pour représenter les différents mots d'un vocabulaire. Des études réalisées sur ce sujet [10] n'ont pas permis de mettre en évidence de loi reliant la structure phonétique du mot et le nombre optimal d'états.

La phase d'apprentissage a pour but de trouver une modélisation optimale de chaque chiffre en partant de modèles initiaux judicieusement choisis. Pour un chiffre du vocabulaire, étant données les matrices initiales A_0 et B_0, on détermine les chemins optimaux permettant de produire les chaînes de symboles de l'ensemble d'apprentissage. On trouve alors les nouveaux paramètres A et B de façon statistique. Une probabilité d'une transition est estimée par sa fréquence d'utilisation dans ces chemins optimaux. De même, une probabilité d'émission $B_{i,j}$ est estimée par le nombre de fois que le symbole j a été produit sur l'état i, divisé par le nombre total de productions de cet état. On constate que les nouveaux paramètres définissent un MMC qui a une probabilité moyenne d'émettre les différentes chaînes d'apprentissage supérieure ou égale à celle du modèle initial. On itère ainsi le processus pour arriver à un optimum.

N'ayant pas fixé a priori les lois de probabilité considérées, les modèles que l'on atteint risquent de n'être que localement optimaux [11][12]. Ceci pose le problème du choix du modèle initial. Dans cette étude, les paramètres de la matrice A_0 ont été arbitrairement initialisés aux valeurs données à la figure 2. Nous proposons un algorithme permettant de trouver une matrice B_0 à partir des données d'apprentissage.

Par la stucture même du modèle fonctionnant de gauche à droite, la phase d'apprentissage peut s'interpréter comme une procédure de segmentation. Pour avoir plus de chances de converger vers un modèle intéressant, nous proposons de déterminer la matrice B_0 par une segmentation grossière des occurrences d'apprentissage en appliquant l'algorithme:

- Chaque chaîne de symboles est divisée en N parties égales (N = nombre d'états)
- Pour chaque symbole j du dictionnaire de QV correspondant au chiffre considéré, on compte le nombre d'apparitions de j dans le $i^{ième}$ segment: $K_{i,j}$. La valeur initiale de $B^0_{i,j}$ est définie par:

$$B^0_{i,j} = \frac{K_{i,j}}{\sum_{\text{tout } l} K_{i,l}}$$

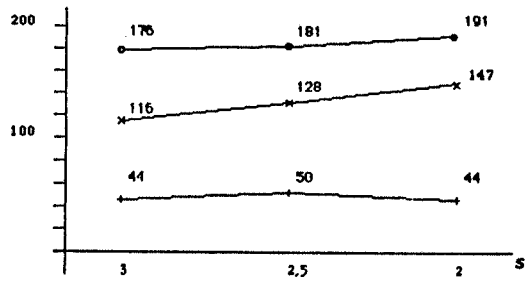


Fig. 3 : Nombre de bonnes reconnaissances en fonction du seuil, o dans les trois premiers, x sur le premier, sur le premier avec critère de confiance

Décision de reconnaissance

Dans les MMC discrets traditionnels, chaque modèle utilise la même représentation du signal de parole entrant. Pour chaque modèle correspondant à un chiffre, on utilise l'algorithme de Viterbi pour trouver la probabilité de la séquence optimale des états pouvant produire la chaîne de symboles observée. La décision de reconnaissance consiste à exhiber le modèle ayant fourni la plus grande de ces probabilités.

Dans notre approche, on ne calcule pas cette probabilité pour tous les modèles mais uniquement pour ceux correspondant aux chiffres retenus par le prétraitement. En contrepartie, si la longueur des chaînes entrant dans chaque automate est identique, la signification des symboles est différente d'un chiffre à l'autre. Il apparait ainsi un problème de normalisation entre modèles car l'on ne sait pas évaluer la qualité de normalisation des différents HMM.

De plus, comme le nombre de représentants dans les dictionnaires de vecteurs est très limité, deux chiffres très différents de celui ayant servi à constituer le dictionnaire, peuvent fournir la même séquence de symboles. A cause des propriétés de l'espace R^{10}, il semblerait que les chaînes de symboles issues d'un codage avec une très forte distorsion, ne contiennent que peu d'informations permettant de les différencier. Ceci nous a montré l'importance du critère de décision au niveau du prétraitement qui ne doit laisser passer que les séquences les plus significatives.

EXPERIENCES DE RECONNAISSANCE

Contrairement aux Etas-Unis, il n'exite pas en France de base de données accessible à tous les Laboratoires et contenant un grand nombre de locuteurs (plusieurs centaines). Un effort important du Greco Communication Parlée a porté sur la constitution de bases de données du Français qui seront très utiles pour l'étude du langage et pour la reconnaissance de parole continue. Les corpus retenus sont énormes et ils n'ont pu être prononcés que par 32 locuteurs, ce qui est beaucoup trop peu pour valider objectivement des systèmes indépendants du locuteur.

On présente ici des résultats d'expériences réalisées avec 20 locuteurs ayant prononcé 4 fois les 10 chiffres. L'apprentissage du système s'effectue avec 15 locuteurs et les tests de reconnaissance utilisent les 4 répétitions des chiffres

par les 5 locuteurs restants. En faisant tourner les locuteurs participant à l'apprentissage, on peut appréhender la dépendance aux locuteurs des résultats obtenus. Cette méthode nous permet de travailler sur les algorithmes, elle ne constitue en aucun cas une validation du système et n'autorise pas la publication de pourcentages de reconnaissance. Ainsi, toutes les valeurs fournies correspondent à des tests sur 200 mots.

La figure numéro 3 présente les résultats typiques du préprocesseur quand on choisit trois types de décision:

- Transmission des 3 meilleurs candidats
- Reconnaissance avec le préprocesseur seul (Burton & Shore)
- Transmission des candidats dont la distance au meilleur est inférieure à un seuil arbitraire.

La figure 4 montre les erreurs réalisées par le système complet (QV+MMC), quand on fait varier le nombre de candidats transmis par le préprocesseur. Nous pouvons observer ici deux choses. D'une part, les résultats obtenus en combinant astucieusement le préprocesseur et les MMC sont meilleurs que ceux fournis par la QV ou par les MMC séparément. Dans les conditions de la figure 4, il est clair que si l'on conserve les 3 premiers candidats du préprocesseur, les MMC prennent une meilleure décision que la méthode de Burton et Shore. D'autre part, on constate que seuls les premières chaînes fournies par le premier étage du système possèdent des informations pertinentes pour discriminer les mots.

Enfin, nous insisterons sur la simplicité de l'algorithme que nous proposons. Il se résume à dix codages vectoriels sur des dictionnaires particulièrement petits (moins de 25 classes) et à l'algorithme de Viterbi exécuté sur trois automates. Son implantation sur TMS320-10 doit facilement tenir en temps réel et l'on peut envisager d'améliorer ses performances en effectuant une meilleure analyse spectrale du système ou en modifiant l'algorithme de classification. Une base de données importante est de toute façon nécessaire pour réaliser un bon apprentissage des dictionnaires de QV et des paramètres des MMC.

CONCLUSION

Cet article propose un système de reconnaissance des chiffres qui associe de manière originale la Quantification Vectorielle et les Modèles de Markov Cachés. Le premier étage utilise les spectres à court terme du signal de parole. Le deuxième étage modélise les transitions possibles entre ces spectres. Ces deux techniques demandent beaucoup de moyens calculs à l'apprentissage mais elles sont particulièrement économiques dans la phase de reconnaissance. Elles sont donc bien adaptées pour faire un système de reconnaissance de mots indépendant du locuteur.

Beaucoup de problèmes théoriques restent à résoudre pour comprendre tous les mécanismes sous-jacents et pour réellement valider le système qui semble particulièrement économique à implanter.

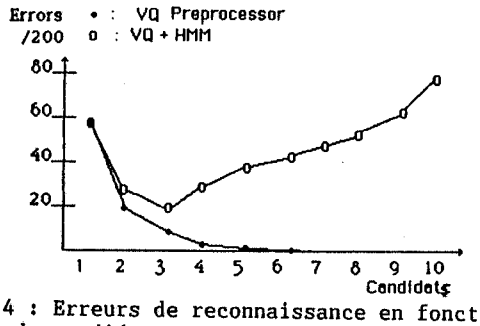


Fig. 4 : Erreurs de reconnaissance en fonction du nombre de candidats transmis

REFERENCES

[1] S.E. Levinson, L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon, "Interactive clustering technics for selecting speaker independent reference templates for isolated word recognition," IEEE Trans. ASSP No. 27, Vol. 2, pp. 134-141, April. 1979.

[2] A. Buzo, H.A. Gray, R.M. Gray, J.D. Markel, "Speech coding based upon vector quantization," IEEE Trans. ASSP Vol. ASSP-28, No. 5, pp. 562-574, Oct. 1980.

[3] L.R. Rabiner, S.E. Levinson, M.M. Sondhi, "On the application of vector quantization and hidden markov models to speaker-independent,isolated word recognition," BSTJ, Vol. 62, No. 4, Apr. 1983.

[4] J.E. Shore, D.K. Burton, "Discrete utterance speech recognition without time alignement," IEEE Trans. IT, Vol. IT-29, No. 4, pp. 473-491, Jul. 1983.

[5] K.C. Pan, F.K. Soong, L.R. Rabiner, A.F. Bergh, "An efficient Vector-Quantization preprocessor for speaker independent isolated Word recognition," Proc. ICASSP-85, pp. 874-877, Mar. 1985.

[6] F. Jelinek, "Continuous speech recognition by statistical methods," Proceedings of the IEEE, Vol. 64, No. 4 , pp. 532-556, Apr. 1976.

[7] H. Boulard, Y. Kamp, C.J. Wellekens, "Speaker dependent connected speech recognition via phonemic Markov models" Proc. ICASSP-85, pp. 1213-1216, Mar. 1985.

[8] L.R. Rabiner, B.H. Juang, S.E. Levinson, M.M Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities," B.S.T.J., Vol. 64, No. 6, pp. 1211-1233, Jul. 1985.

[9] L. Miclet, M.H. Dabouz, "Low bit rate transmission of speech by Vector Quantization of the spectrum," à paraître dans Speech Communication

[10] A. Tassy, L. Miclet, "Quantification Vectorielle et reconnaissance de mots multilocueur" Congres AFCET- RFIA, Grenoble, Oct. 1985

[11] A. Nadas, "Hidden markov chains, the forward-backward algorithm and initial statistics," IEEE Trans. ASSP Vol. ASSP-31, No. 2, pp. 504-506, Apr. 1983.

[12] L.A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov Sources," Trans. on Information Theory, Vol. IT-28, No. 5, pp. 729-734, Sept. 1982.

Reconnaissance de la Parole par des modèles Markoviens: application aux grands vocabulaires

Hélène CERF, Anne-Marie DEROUAULT, Marc EL BEZE,
Bernard MERIALDO, Serge SOUDOPLATOFF

Centre Scientifique IBM France
36 avenue Raymond Poincaré, 75116 Paris FRANCE
Tel: (1) 45 05 14 00

Résumé

Cet article a plusieurs buts:

1. rappeler la formulation du problème de la reconnaissance de la parole dans le cadre de la Théorie de l'Information, et expliquer quels sont, à notre avis, les avantages et inconvénients de cette approche.
2. rappeler quelques notions élémentaires sur l'utilisation des sources de Markov, en particulier l'apprentissage et les divers types de décodage.
3. présenter les résultats des expériences menées au Centre Scientifique en utilisant ces méthodes pour la reconnaissance phonétique et la dictée automatique avec un grand vocabulaire en français.

Théorie de l'information

En se plaçant dans le cadre de la Théorie de l'Information, Jelinek [8,9] a proposé la formulation suivante du problème de la reconnaissance de la parole:

- le locuteur est considéré comme une *source* qui émet des messages faisant partie d'un certain *langage L*.
- chaque message traverse un canal appelé *canal acoustique*, composé de l'appareil phonatoire du locuteur, de l'espace de propagation entre le locuteur et le microphone, et du *processeur acoustique* qui traite le signal du microphone pour produire une *observation acoustique O*.
- à partir de cette observation, un *décodeur linguistique* retrouve un message décodé $\overline{m}$.

Bien entendu, le canal acoustique ne réalise pas une bijection entre m et O, mais il est caractérisé par les probabilités conditionnelles $p(O \mid m)$. Le décodage optimal, c'est-à-dire celui qui minimise le taux d'erreur sur le message décodé, consiste alors à choisir $\overline{m}$ tel que:

$$p(\overline{m} \mid O) = \max_{m'} p(m' \mid O)$$

et en appliquant la formule de Bayes, on aboutit à l'équation fondamentale de la reconnaissance de la parole:

$$\boxed{p(\overline{m} \mid O) = \max_{m'} \frac{p(O \mid m') \times p(m')}{p(O)}}$$

On remarque très simplement que le dénominateur ne dépend pas de m' et qu'il suffit donc de maximiser le numérateur $p(O \mid m') \times p(m')$.

Cette présentation fait ressortir les trois problèmes à résoudre pour construire un système de reconnaissance de la parole:

1. choisir les paramètres acoustiques qui caractérisent l'observation, et réaliser le processeur acoustique qui les extrait du signal.
2. élaborer un modèle acoustique qui indique la façon dont les messages sont transmis en terme de ces observations acoustiques, en permettant de calculer $p(O \mid m)$.
3. élaborer un modèle linguistique qui permette d'estimer la probabilité de chaque message $p(m)$.

Un quatrième problème est de savoir comment trouver pratiquement le meilleur message, ce qui n'est pas trivial lorsque le langage L a un cardinal très grand (ou infini).

La puissance de cette formulation tient dans sa généralité. En effet, on n'a pas encore précisé quel était le langage (ensemble des messages) considéré, ni quelle était l'observation acoustique, ni comment on réalise les différents modèles (en particulier, on n'a pas encore parlé de source de Markov). Mais cette formule indique quels sont les éléments du système, et surtout, comment se réalise **la coopération des niveaux acoustique et linguistique.** Cette coopération se fait à partir des probabilités des modèles acoustique et linguistique, en maximisant leur produit.

Sources de Markov

En Théorie de l'Information, il est courant de modéliser une source quelconque par une source de Markov. Les sources de Markov ont l'avantage d'être simples, leurs propriétés sont bien connues d'un point de vue théorique, il est facile de voir quelle approximation elles font d'une source réelle, et enfin il existe des algorithmes puissants pour les utiliser efficacement.

Définition

Une *source de Markov* est un Automate d'Etats Finis probabiliste, c'est-à-dire un quadruplet (S,A,T,P) où:

- S est un ensemble fini appelé *ensemble des états,*
- A est un ensemble fini appelé *alphabet,* dont les éléments sont appelés *symboles,*
- T est un sous-ensemble de $S \times A \times S$, dont les éléments sont appelés *transitions.* Si $t = (s, a, s')$, on dit que t est une transition qui va de s (état de départ) vers s' (état d'arrivée) en produisant le symbole a.
- P est une application $T \rightarrow [0, 1]$ qui définit une distribution de probabilité sur l'ensemble des transitions partant d'un état donné, c'est-à-dire que:

$$\forall s \quad \sum_{a,s'} p(s, a, s') = 1$$

les valeurs $p(s, a, s')$ sont appelées *paramètres* de la source.

Souvent, on particularise un état appelé *état initial,* et un autre appelé *état final.* Un *chemin* est une suite de transitions t_i, $i = 1, 2..n$ telle que l'état d'arrivée de t_i coïncide avec l'état de départ de t_{i+1}. La probabilité de ce chemin est le produit des probabilités de transitions t_i. Un chemin est dit *complet* si t_1 part de l'état initial, et si t_n arrive à l'état final. Si a_i est le symbole produit par la transition t_i, le chemin t_1^n produit la suite de symboles a_1^n. La probabilité qu'une suite de symboles a_1^n soit produite par la source de Markov est la somme des probabilités de tous les chemins qui produisent cette suite:

$$p(a_1^n) = \sum_{t_1^n} p(t_1^n)$$

Apprentissage

L'intérêt des sources de Markov, en particulier pour la reconnaissance de la parole, réside dans l'existence d'algorithmes qui permettent de calculer les paramètres optimaux d'une source en fonction d'une observation des symboles produits. Il est même possible de faire ces calculs sans connaître la suite exacte des états qui ont produit cette observation *(modèle de Markov caché,* ou *HMM, hidden Markov model).* Cela provient d'une inégalité introduite par Baum [2] qui donne lieu à un algorithme itératif appelé *Forward-Backward* ou *Baum-Welch.*

Si l'on fixe la structure de la source de Markov M, cet algorithme cherche à maximiser la probabilité $p(O \mid M)$ que l'observation O soit produite par la source, en cherchant les valeurs optimales des paramètres. Il fonctionne de façon itérative:

- on fixe des valeurs arbitraires aux paramètres,
- avec ces valeurs, on peut calculer le nombre moyen de fois où chaque transition est utilisée dans la production de l'observation,
- cela fournit de nouvelles estimées des paramètres, dont l'inégalité de Baum assure qu'elles sont meilleures que les précédentes,
- on itère jusqu'à ce qu'un certain critère d'arrêt soit vérifié (souvent on fixe à priori le nombre d'itérations).

Cet algorithme converge vers un maximum **local** de la fonction $p(O \mid M)$. Il est simple à mettre en oeuvre et relativement rapide. Il permet de traiter des sources ayant plusieurs dizaines de milliers de paramètres, et donc de construire automatiquement des modèles complexes, qu'il serait impossible de définir manuellement.

Mentionnons que d'autres algorithmes sont envisageables, par exemple le *'simulated annealing'* [11] qui permet de trouver un maximum global et dont le fonctionnement s'inspire du refroidissement lent d'un matériau pour lui conférer une structure cristalline régulière (minimum d'énergie de cohésion).

On peut aussi envisager de maximiser une autre fonction que $p(O \mid M)$, ce qui conduit à des systèmes plus **discriminants** (voir Bahl [1]).

Décodage

Rappelons les deux grands types d'algorithmes de décodage utilisés en Théorie de l'Information:

1. l'algorithme de *Viterbi* [13], qui a une complexité de calcul linéaire en fonction du nombre d'états, et qui permet de connaître la suite des états correspondant au meilleur chemin,
2. les algorithmes de *décodage séquentiel,* tels ceux de Fano [6] ou de Jelinek [7] qui font une exploration arborescente des chemins possibles.

L'algorithme de Viterbi:

- à un instant donné i, et pour chaque état s du modèle, on garde le meilleur chemin ayant produit le début de l'observation o_1^i et arrivant en s.
- on calcule tous les prolongements de ces chemins qui produisent le symbole o_{i+1}, et on garde le meilleur à chaque état.
- on continue jusqu'à la fin de l'observation.

L'algorithme de Jelinek à pile: (Stack decoding)

- à une étape de l'algorithme, on a une pile de chemins partiels, triés selon une fonction de vraisemblance,
- on prolonge le meilleur chemin, et on le remplace dans la pile par ses prolongements,
- on retrie la pile, et on continue jusqu'à ce que l'on trouve en haut de la pile un chemin complet.

Remarque: le problème du décodage se ramène souvent à la recherche d'un chemin optimal dans un graphe, et il y a donc une forte ressemblance entre les algorithmes ci-dessus et des algorithmes utilisés en recherche opérationnelle ou en Intelligence Artificielle (comme l'algorithme A^* [12]).

Reconnaissance phonétique

Nous quittons maintenant la théorie pour présenter quelques expériences que nous avons réalisées au Centre Scientifique IBM-France en utilisant ce type de modèle.

Observation acoustique

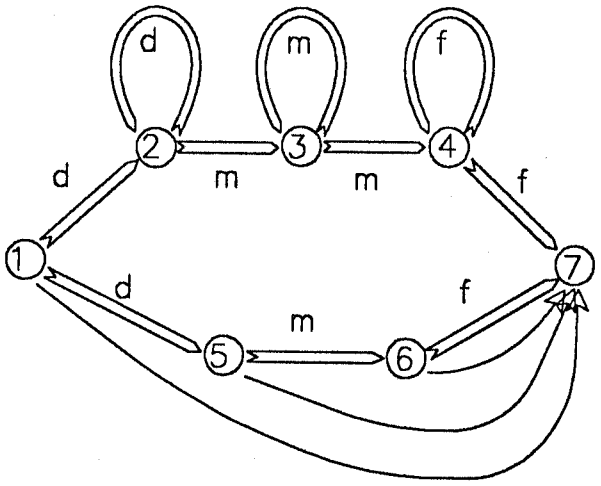
Le signal de parole est échantillonné à 10kHz, 12 bits. Sur chaque fenêtre de 128 échantillons (12.8ms), on applique une fenêtre de Hanning, suivie d'un spectre FFT, de la projection sur une échelle MEL à 20 bandes, et d'une quantification vectorielle VQ sur 200 points (obtenus par l'algorithme de *k-means*).

L'observation acoustique consiste donc en une suite de nombres entre 1 et 200 (appelés *labels*), émis environ chaque centième de seconde.

Les phrases sont prononcées en mode *'syllabes isolées'* en marquant une légère pause entre chaque syllabe. Pour l'instant, les expériences sont réalisées avec la voix d'un seul locuteur masculin.

Modèle phonétique

Chaque phonème est modélisé par une source de Markov ayant la structure suivante:



Les flèches pleines représentent en fait 200 transitions émettant chacun des 200 labels, les flèches fines représentent des transitions *vides* qui n'émettent pas de labels. Les transitions (1,2), (2,2), (1,5) sont *liées* [9] par une même distribution de labels d, c'est-à-dire que leurs probabilités sont de la forme

$$p(s, a, s') = q(s, s') \times q_d(a)$$

de même pour les transitions (2,3), (3,3), (3,4), (5,6) liées par la distribution m, et pour (4,4), (4,7), (6,7) liées par la distribution f. L'algorithme Forward-Backward s'adapte facilement au cas de transitions liées.

Apprentissage

Le corpus d'apprentissage est constitué par un enregistrement de 410 phrases (dont les phrases équilibrées de Combescure [3]). On connaît la transcription phonétique de chaque phrase lue, on peut donc construire la source de Markov de chaque phrase en

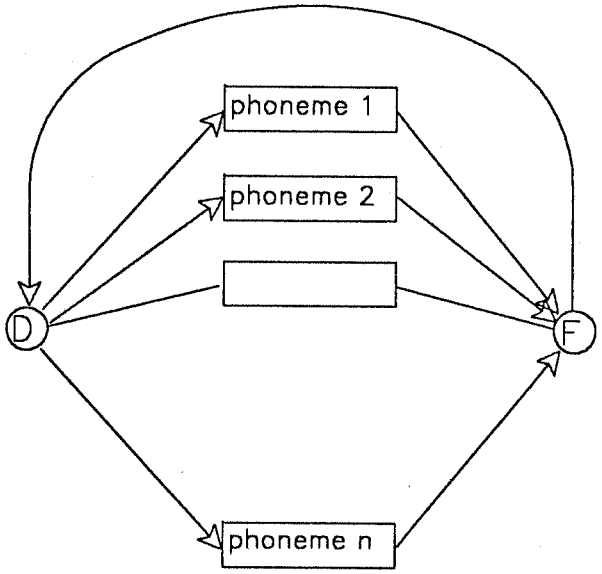
concaténant les machines phonétiques. Ensuite, on applique l'algorithme FB pour avoir les valeurs optimales des paramètres pour chaque machine. On part de distributions équiprobables. Les 90 premières phrases ont été segmentées manuellement. Pendant les 5 premières itérations, on applique le FB sur ces 90 phrases, en le contraignant à respecter cette segmentation. Cette procédure permet de *'boot-straper'* des valeurs initiales des paramètres qui soient déjà proches des valeurs optimales. Ensuite, on applique le FB sur le corpus entier sans restriction de segmentation pendant 5 nouvelles itérations.

Les valeurs obtenues sont *lissées* par un algorithme du type *'deleted estimator'* qui permet d'éviter une trop grande particularisation des paramètres pour le corpus d'apprentissage.

Reconnaissance

Il y a de nombreuses façons d'utiliser ces modèles pour faire de la reconnaissance phonétique, selon la fonction que l'on cherche à maximiser et l'algorithme de décodage. Nous présentons ici quelques variantes utilisant l'algorithme de Viterbi.

On construit une nouvelle machine en mettant en parallèle les différentes machines phonétiques ϕ_l, plus deux nouveaux états D et F reliés comme l'indique le schéma:



On voit qu'un chemin dans cette machine consiste à faire un certain nombre de boucles en utilisant une machine phonétique à chaque boucle et en revenant à l'état D. Pour une observation donnée, correspondant à l'enregistrement d'une phrase, on peut rechercher le chemin le plus probable dans cette machine par l'algorithme de Viterbi. Cela donnera une suite de boucles, donc une suite de phonèmes ϕ_l qui constituera la chaîne phonétique reconnue.

L'interprétation de cette suite dépend du choix des valeurs attribuées aux transitions vides (D, ϕ_l). Nous proposons 4 cas:

1. elles ont chacune une valeur 1 (on ne peut plus alors parler de probabilité, mais on peut toujours appliquer l'algorithme de Viterbi),
2. elles ont chacune une probabilité $\frac{1}{N}$ (phonèmes équiprobables, N = nombre de phonèmes),
3. elles ont chacune la probabilité $p(\phi_l)$,
4. elles dépendent aussi du dernier phonème ψ du meilleur chemin arrivant en D, par $p(\phi_l \mid \psi)$. (cela revient à dupliquer l'état D autant de fois qu'il y a de machines phonétiques).

(les probabilités $p(\phi_l)$ et $p(\phi_l \mid \psi)$ sont fondées sur les fréquences des phonèmes et des diphonèmes relevées sur un corpus phonétisé).

Avant de voir l'interprétation de ces diverses possibilités, remarquons que si les machines de Markov sont suffisamment *'déterministes'*, il y a peu de différence entre $\sum p(t_1^n)$ et $\max p(t_1^n)$. On peut alors identifier la probabilité du meilleur chemin avec la probabilité que l'observation soit produite par la machine. Nous nous plaçons dans cette hypothèse pour la suite, et prenons quelques libertés avec la rigueur mathématique pour mieux faire ressortir les principes fondamentaux de ces phénomènes.

1. dans le cas 1, la probabilité du meilleur chemin correspondra à peu près à $p(O \mid \phi_1^n)$. On voit alors que le décodage revient à chercher la suite de phonèmes pour laquelle l'observation est la plus probable.
2. dans le cas 2, on pondère cette probabilité selon le nombre de phonèmes utilisés dans la chaîne. Cela revient à maximiser
$$p(\phi_1^n \mid O) = \frac{p(O \mid \phi_1^n) \times p(\phi_1^n)}{p(O)}$$
en considérant que tous les phonèmes sont équiprobables, et que donc
$$p(\phi_1^n) = \prod_i \frac{1}{N}$$
3. le cas 3 relève du même principe, mais cette fois on a une connaissance plus précise sur les phonèmes et on considère que
$$p(\phi_1^n) = \prod_i p(\phi_l)$$
4. le cas 4 est identique, sauf que l'on modélise alors les suites de phonèmes selon une source de Markov du premier ordre en prenant
$$p(\phi_1^n) = \prod_i p(\phi_l \mid \phi_{l-1})$$

Bien évidemment, on pourrait continuer à utiliser des contraintes phonétiques de plus en plus précises *(n-phonèmes)*, à condition d'avoir des statistiques satisfaisantes. Nous nous en tiendrons à ces quelques cas.

Résultats

Nous avons donc mené des expériences de décodage relativement à ces quatre approches. Les tests sont réalisés sur un ensemble de 79 phrases. Les résultats sont présentés dans le tableau suivant, où l'on indique le taux de reconnaissance TR (pourcentage de phonèmes originaux reconnus), et le taux d'insertion (pourcentage de phonèmes rajoutés par le décodeur).

	TR	TI
1	78.1	21.5
$\frac{1}{N}$	76.7	8.5
$p(\phi_l)$	78.1	6.7
$p(\phi_l \mid \phi_{l-1})$	80.9	6.1

Ceci montre la facilité avec laquelle la théorie peut utiliser des modèles de plus en plus précis. Bien entendu, l'utilisation de meilleurs modèles améliore en général les performances.

Alignement

Il nous paraît intéressant de mentionner une autre utilisation possible des modèles phonétiques qui résout très simplement le problème de l'étiquetage d'une phrase. Si on dispose de l'enregistrement d'une phrase, et de sa représentation phonétique, il est très simple de former la machine correspondant à la phrase, et par l'algorithme de Viterbi, de trouver le chemin de probabilité maximale dans cette machine. Le long de ce chemin, on sait à chaque instant de la prononciation de la phrase dans quel état on se trouve, donc dans quelle machine phonétique. Cela permet de

d'indiquer sur le signal quels sont les intervalles que couvre chaque phonème de la transcription.

Les expériences que nous avons faites dans ce sens montrent que l'étiquetage ainsi réalisé est du même ordre de qualité qu'un étiquetage manuel.

Dictée automatique

Les travaux que nous décrivons maintenant ont été réalisés dans le cadre de l'étude d'un système expérimental appelé

PARSYFAL:

PARis SYllable-based French Automatic Listener.

L'objectif (à long terme) de ce système est la dictée automatique avec très grand vocabulaire *('Very Large Size Dictionary', VLSD,* contenant 100.000 mots et plus). La méthodologie employée est celle décrite au début de cet article.

Revenant à cette formulation, nous étudions maintenant son application à la dictée automatique. Le langage considéré est alors le français, plus précisément les messages émis par le locuteur sont des phrases. La reconnaissance revient à chercher la suite de mots qui maximise

$$p(O \mid m_1^n) \times p(m_1^n)$$

Dans le cadre d'un projet de transcription de la sténotypie [4,5], nous avons déjà étudié plusieurs façons de réaliser un modèle de langage permettant d'estimer $p(m_1^n)$. Ces modèles s'appliquent au discours naturel sans contrainte syntaxique ou sémantique, et supportent un très grand vocabulaire. Dans les expériences décrites ci-après, nous utilisons le modèle linguistique basé sur la prédiction d'un mot d'après les deux dernières classes grammaticales. Nous ne détaillons pas plus ici ce modèle.

Nos expériences actuelles portent sur l'étude de la reconnaissance avec grand vocabulaire. Plus précisément, nous utilisons un dictionnaire composé des 10.000 formes déclinées les plus fréquentes, plus les mots des 79 phrases à reconnaître. A chaque forme sont associés:

- sa ou ses classes grammaticales,
- la ou les fréquences correspondantes,
- sa description phonétique.

Certains mots ont plusieurs phonétiques possibles.

Un décodage basé sur l'algorithme de Viterbi recherche la meilleure suite de mots au sens de la probabilité combinée:

$$p(O \mid m_1^n) \times p(m_1^n)$$

Nos résultats actuels indiquent 101 erreurs sur 722 mots, soit un taux d'erreur inférieur à 14% (en pourcentage de mots incorrectement transcrits).

Voici deux exemples de phrases reconnues:

"les décisions sont prises à la majorité des voies (voix) "

"cette situation ne s'offrait (saurait) bien entendu se prolonger"

Conclusion

Nous avons montré ici une petite partie des applications possibles des modèles Markoviens dans la reconnaissance de la parole. Nous avons présenté nos travaux sur l'application de ces modèles pour la reconnaissance de la parole avec un grand vocabulaire. Pour une tâche de dictée automatique, sans contrainte syntaxique ou sémantique sur le discours, et avec un vocabulaire de 10.000 formes déclinées, nos résultats actuels nous donnent un taux d'erreur inférieur à 14%.

Il reste à retenir qu'au delà des sources de Markov, **l'approche Théorie de l'Information** de la reconnaissance de la parole fournit un cadre extrêmement puissant pour le développement d'outils de reconnaissance. Loin de restreindre l'utilisateur à un ensemble figé de méthodes, elle laisse au contraire une grande liberté pour définir de nouveaux procédés correspondant au type d'expertise que l'on veut mettre en valeur.

Bibliographie

[1] L. Bahl, P. Brown, P. de Souza, R. Mercer Maximum mutual information estimation of Hidden Markov Model parameters, ICASSP 86, Tokyo..

[2] L.E. Baum An Inequality and Associated Maximisation Technique in Statistical Estimation of Probabilistic Fonctions of Markov Process, Inequalities, Vol 3, 1972, pp 1-8..

[3] P. Combescure, Phonetically balanced sentences, Recherches acoustique vol VI, 1979-1980 CNET Lannion (France).

[4] A-M. Derouault, B. Merialdo, Language modeling at the syntactic level, 7th International Conference on Pattern Recognition, August 1984, Montreal.

[5] A-M. Derouault, B. Merialdo, Probabilistic Grammar for phonetic to French transcription, ICASSP 1985, Tampa..

[6] R.M. Fano, A heuristic introduction to probabilistic decoding, IEEE Trans. Information Theory, IT-9, no97, 1963.

[7] F. Jelinek, Fast sequential decoding algorithm using a stack, IBM Journal of research and development, vol 13, no 6, November 1969.

[8] F. Jelinek, L.R. Bahl and R.L. Mercer, Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech, IEEE Trans. IT, vol IT-21, no 3, May 1975..

[9] L.R. Bahl, F. Jelinek, R.L. Mercer, A maximum likelihood approach to continuous speech recognition, IEEE Trans. PAMI, vol PAMI-5, no 2, March 1983..

[10] F. Jelinek, The development of an experimental discrete dictation recognizer, Proceedings of the IEEE, vol 75, no 11, November 1985.

[11] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science, 1983, 220, 671-680..

[12] N. J. Nilsson, Problem solving methods in Artificial Intelligence, Mc Graw Hill, 1971.

[13] A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Trans. Information Theory, IT-13, 1967.

RECONNAISSANCE DE LA PAROLE CONTINUE PAR PROGRAMMATION DYNAMIQUE

Joseph di MARTINO

Equipe Reconnaissance des Formes et Intelligence Artificielle
Centre de Recherche en Informatique de Nancy
Universite de Nancy 1
Campus Scientifique BP 239
54506 Vandoeuvre-les-Nancy

1. INTRODUCTION

Les techniques de programmation dynamique ont été proposées il y a plus de vingt ans par des chercheurs soviétiques [1-2] lorsqu'il a été acquis que les techniques de normalisation linéaire donnaient des résultats très médiocres. Ces techniques, dans la dernière décénie, ont été extrêmement étudiées par des chercheurs de tout pays et leur efficacité a clairement été mise en évidence. Toutefois, elles ne constituent pas la panacée en reconnaissance de la parole: en effet, elles ont du mal à prendre en compte les fortes variations d'amplitude du signal de parole, à tenir compte des phénomènes de coarticulation dans le cas où les mots sont prononcés continument,i.e. sans pause artificielle entre les mots, à prendre en compte les variabilités inter-locuteur. Ce dernier point est d'ailleurs fortement gênant car il oblige un ré-apprentissage du vocabulaire dans le cas d'un changement de locuteur. Un autre point qui va à l'encontre de ces méthodes est que la quantité d'informations qu'il faut conserver en mémoire est très importante et limite en conséquence la réalisation de systèmes de reconnaissance à grand vocabulaire.

Après un constat aussi peu éloquent pour la programmation dynamique, il faut se demander pourquoi de telles techniques ont eu autant de succès dans les années précédentes et pourquoi elles ont mobilisé autant de chercheurs de part le monde. A mon sens, la réponse à cette question est double. Premièrement, je pense que cela est dû au fait que très vite les chercheurs ont réussi à réaliser de petits systèmes à fort taux de reconnaissance, avoisinant les 100% dans le cas de vocabulaires pas très difficiles. Deuxièmement, je pense que cela est dû au fait qu'aucune autre méthode n'a réussi franchement a détrôner la programmation dynamique.

Des méthodes de reconnaissance fondées sur des considérations phonétiques ont été proposées. Mais le savoir dans ce domaine, certainement très prometteur, est si flou, si peu formalisé, qu'aucun système n'a pu véritablement franchir la première phase de tout processus de reconnaissance, i.e. la phase de paramétrisation de l'onde acoustique ou de décodage acoustique dans ce cas bien précis. De plus, si on relit les conclusions du projet expérimental ARPA [3], on se rend compte que le seul système qui respecte le cahier des charges est le système le moins sophistiqué et surtout celui qui utilise le moins de connaissances phonétiques. Pourtant, les experts en reconnaissance de la parole ont tout à gagner à comprendre les mécanismes qui interviennent dans la production de la parole. Une meilleure compréhension de ces mécanismes permettra sûrement de développer des algorithmes de paramétrisation de l'onde vocale, plus performants que ceux qui ont été proposés jusqu' à présent, et en tout état de cause, de telles connaissances devront être intégrées dans les systèmes de compréhension de la parole continue à venir, si l'on veut véritablement progresser dans ce domaine

Ceci étant dit, je tiens aussi à préciser que la notion de segmentation qui présuppose que l'onde vocale est constituée d'une suite de phonèmes est fondée sur un modèle extrêmement simpliste du fait des recouvrements temporels des productions phonémiques: en effet, lorqu'on visualise sur un spectrogramme un phonème, il faut savoir que sa production a débuté bien avant le moment précis où il est détecté. En conséquence, la notion de segmentation de l'onde vocale en unité phonétique, telle qu'on la conçoit usuellement, est une notion qui mérite d'être revue. De plus, pour un expert en reconnaissance de la parole, une telle représentation de l'onde vocale, entachée d'erreurs -souvent fatales pour les systèmes de reconnaissance- peut ne pas être satisfaisante s'il a à sa dispostion des moyens algorithmiques qui lui permettent d'adopter la stratégie la plus sage en reconnaissance de la parole, à savoir retarder au maximum les prises de décision - prises de décision

qui ont été effectuées, il faut le rappeler, dès les plus bas niveaux, dans la cas des approches phonétiques traditionnelles !?? -.

Cette mise au point étant faite, je peux introduire le sujet de cet article. Je me propose de montrer que les algorithmes de programmation dynamique pour la reconnaissance de mots isolés et de mots enchaînés peuvent être généralisés dans le but de tenir compte d'une classe de langages beaucoup plus larges que celle des langages réguliers -i.e. artificiels- englobant des langages pseudo-natutels. De tels algorithmes permettent de traiter une part non négigeable de la problématique de la parole continue à l'aide de techniques de programmation dynamique. L'approche que je vais développer permet aussi d'unifier de manière très forte tous les algorithmes de reconnaissance fondés sur la programmation dynamique.

2. QUELQUES RAPPELS SUR LES ALGORITHMES DE RECONNAISSANCE DE MOTS ISOLES ET DE MOTS ENCHAINES.

2.1. RECONNAISSANCE DE MOTS ISOLES

Si l'on désigne par D(n,j,i) la distance cumulée associée au chemin de recalage aboutissant au point (n,j,i) où le ieme prélèvement de la forme inconnue est mis en coincidence avec le jeme prélèvement de la forme de référence n, alors la programmation dynamique permet d'évaluer D(n,j,i) de la manière suivante:

$$D(n,j,i)=\underset{(n,j',i')\in V(n,j,i)}{MIN}\ D(n,j',i') + dp((n,j',i'),(n,j,i)) \quad (1)$$

où
- V(n,j,i) est un voisinage du point (n,j,i) qui dépend de la contrainte locale utilisée[4]. Dans le cas de la contrainte locale sans condition de pente-dans toute la suite de cet article, nous nous placerons dans le cas d'une telle contrainte tout en ayant à l'esprit que celle-ci peut rendre les algorithmes de programmation dynamique biaisés du fait du non respect, dans certaines conditions, de la propriété de constance des longueurs de chemins de recalage aboutissant en un point de l'espace de comparaison[4]-, que l'on désigne souvent par contrainte 0, V(n,j,i) s'écrit:

$$V(n,j,i)= \{(n,j,i-1),(n,j-1,i-1),(n,j,i-1)\} \quad (2)$$

-dp((n,j',i'),(n,j,i)) est une distance locale pondérée entre le point (n,j',i') et le point (n,j,i).

Ces notations étant précisées, un algorithme de reconnaissance de mots isolés consiste à évaluer D(n,j,i) pout tout n, pour tout j et pour tout i. Le mot reconnu n* est celui qui fournit un score normalisé minimal.

$$n^* = \underset{n}{ARGMIN}\ D(n,Jn,I) / Nn$$

/* Jn est le nombre de prélèvements de la forme Rn et I est le nombre de prélèvements de la forme inconnue. Nn est un facteur de normalisation. Celui-ci a été indicé par n car il peut être fonction du nombre de prélèvements de la forme de référence Rn */

2.2. RECONNAISSANCE DE MOTS ENCHAINES SANS CONTRAINTES SYNTAXIQUES

Dans l'évaluation des relations de programmation dynamique appliquées à la reconnaissance de mots enchaînés, il faut tenir compte du fait qu'un mot peut être précédé par n'importe quel autre. Ceci peut se faire simplement grâce à la notion de voisinage en un point de l'espace de comparaison. Dans ce cas, deux types de voisinages doivent être définis:

a) Voisinage inter-mot: (j>1)

$$V(n,j,i) = \{(n,j,i-1),(n,j-1,i-1),(n,j,i-1)\} \quad (4)$$

b) Voisinage inter-mot: (j=1)

$$V(n,j,i) = \{(n,j,i-1)\} \cup \bigcup_{n'} \{(n',Jn',i-1)\} \quad (5)$$

Moyennant ces modifications, la relation de programmation dynamique appliquée à la reconnaissance de mots enchaînés s'écrit :

$$D(n,j,i) = \underset{(n',j',i')\in V(n,j,i)}{MIN}\ D(n',j',i') + dp((n',j',i'),(n,j,i)) \quad (6)$$

La séquence de mots optimale est déterminée à la fin du processus de comparaison -i.e. lorsque D(n,j,i) a étéévaluée pour tout n,j,i- grâce à des fonctions pointeurs qui permettent de retrouver le chemin de recalage optimal.

2.3. RECONNAISSANCE DE MOTS ENCHAINES AVEC CONTRAINTES SYNTAXIQUES

Dans ce cas un mot ne peut être précédé que par un ensemble restreint de mots définis par les contraintes syntaxiques. Cet ensemble peut d'ailleurs être vide.

Afin de pouvoir tenir compte des contraintes syntaxiques, la dimension de l'espace de comparaison relative aux formes de référence doit être remplacer par la dimension des contextes des formes de référence. Un point dans cet espace peut être représenté par un triplet:

$$(t,j,i)$$

où le ieme prélèvement de la forme inconnue est mis en coincidence avec le jeme prélèvement de la forme de référence en contexte t, et où t est un triplet (s1,n,s2) relatif à l'arc du graphe syntaxique, étiqueté par le symbole n - i.e. la forme de référence n- qui part de l'état s1 et qui va à l'état s2.

Soient p1,p2,p3 les trois projections élémentaires définies sur l'ensemble des contextes - p1(t)=s1, p2(t)=s2,p3(t)=s3-

Deux contextes sont syntaxiquement liés, i.e. t<=t', si et seulement si p3(t)=p1(t').

Grâce à ce formalisme, il est possible de définir le voisinage d'un point quelconque de l'espace de comparaison dans le cas de contraintes syntaxiques:

a) VOISINAGE INTRA-MOT

$$V(t,j,i)= \{(t,j,i-1),(t,j-1,i-1),(t,j-1,i)\} \quad (7)$$

b) VOISINAGE INTER-MOT

$$V(t,j,i) = \{(t,j,i-1)\} \cup \{(t',Jp2(t'),i-1)\} \quad (8)$$
$$t<=t'$$

Le voisinage d'un point quelconque de l'espace de comparaison étant défini, la relation générale de programmation dynamique dans le cas de contraintes syntaxiques s'écrit:

$$D(t,j,i) = \min_{(t',j',i')\in V(t,j,i)} D(t',j',i') + dp((t',j',i'),(t,j,i)) \quad (9)$$

De même que dans le cas de la reconnaissance de mots enchaînés sans contraintes syntaxiques, la séquence de mots optimale est déterminée à la fin du processus de comparaison grâce à des fonctions pointeurs.

3. RECONNAISSANCE DE LA PAROLE CONTINUE PAR PROGRAMMATION DYNAMIQUE

Un des défauts majeurs des langages réguliers représentés à l'aide d'automates d'états finis réside dans le fait que d'une part la structure du langage n'apparait pas, et d'autre part, qu'il n'est pas possible d'introduire des contraintes sémantiques du type de celles définies dans [5].

Un bon modèle linguistique qui obvie à ces différents inconvénients, et qui par voie de conséquence est un bon modèle pour la reconnaissance de la parole continue est le modèle des machines multi-niveaux. J'appelle un langage multi-niveaux, un langage qui peut être représenté par un automate hiérarchique, i.e. un automate dont les symboles peuvent être des symboles terminaux ou des automates. A cet automate peuvent être associés différents niveaux de construction -d'où le nom machines multi-niveaux-. Le niveau 0 est le niveau le plus profond de l'automate . A ce niveau les automates ne sont constitués que de symbole terminaux et servent essentiellement à définir la structure acoustique des entités lexicales. Les niveaux suivants servent à définir les contraintes syntaxiques et sémantiques du langage.

L'espace de comparaison d'un automate hiérarchique à n niveaux peut être représenté par un espace à n+2 dimensions:

$$(tn,tn-1,\ldots\ldots,t2,t1,t0,i)$$

En ce point, la ieme entité acoustique est mise en coincidence avec le symbole en contexte t0 dans la machine en contexte t1 appartenant à la

machine en contexte t2, elle-même appartenant à la machine en contexte t3,....,elle même appartenant à la machine en contexte tn.

Dans cet espace de comparaison, il est possible d'appliquer la programmation dynamique . La solution générale s'écrit alors:

$$D(tn,tn-1,...,t0,i) = D(t'n,t'n-1,...,t'0,i') + dp((t'n,t'n-1,...,t'0,i'),(tn,tn-1,...,i))$$

pour $(t'n,t'n-1,...t'0,i') \in V(tn,tn-1,...,t0,i)$.
où $V(tn,tn-1,....,t0,i)$ est un voisinage du point $(tn,tn-1,...,t0,i)$ dépendant de la structure de l'automate et des contraintes sémantiques.

-NB: Ces notions de contextes et de voisinages de machines multi-niveaux nécessitent, pour être définies précisément, des considérations qui débordent largement du cadre de cet article. Celles-ci seront décrites dans des papiers à venir-

4. CONCLUSION

Je propose dans cet article une approche qui permet d'unifier les algorithmes de reconnaissance de la parole utilisant la programmation dynamique. Cette approche unificatrice m'a conduit de façon naturelle à introduire le concept de machines multi-niveaux. Ces machines sont très bien adaptées à la reconnaissance de la parole continue car elles maintiennent la structure du langage et permettent l'introduction de contraintes sémantiques.

5. REFERENCES

[1] A T.K. VINTSYUK: "Speech Discrimination by Dynamic Programming",Kibernetica (Cybernetics), Vol.4, No.1, pp. 81-88, January-february 1968.

[2] V.M. VELICHKO,N.G.ZAGORUYKO: "Automatic Recognition of 200 Words", Int. J. Man-Machine Studies, 2, pp. 223-234, 1970.

[3] D.H. KLATT: "Review of the Arpa Speech Understanding Project", J.Acoust.Soc.Am. Vol. 62, No. 6, December 1977.

[4] J. di MARTINO: " Dynamic Time Warping Algorithms for Isolated and Connected Word Recognition", in Nato Asi Series, Vol. 16, New Systems and Architectures for Automatic Speech Recognition, R. DE MORI, C.Y. SUEN Editors, Springer-Verlag Berlin Heidelberg,1985.

[5] J.M. PIERREL: "Etudes et Mise en Oeuvre de Contraintes Linguistiques en Compréhension de la Parole Continue", Thèse d'Etat, Université de Nancy I.

EXPERIENCES EN RECONNAISSANCE DES PARTIES NON STATIONNAIRES DE LA PAROLE

Chen-Guang Wang, Jean-Pierre Tubach

Département SYC (CNRS, UA 820)
E.N.S.T. , 46, Rue Barrault, 75013, PARIS

ABSTRACT

We have conducted experiments, using vector quantization for the recognition of transient parts of the speech signal (threshold clustering algorithm, nearest neighbour recognition method). Different distances have been used (euclidian distance on LPC cepstral coefficients gives the best results).

Recognition scores are poor on (consonant cluster + any vowel) segments, and fair on (consonant + /e/)short word. This suggests that transitions should be considered after steady-state (mainly vowels) segments have been recognized.

1.INTRODUCTION

La quantification vectorielle (QV) est récemment utilisée avec succès en codage de parole pour la transmission, ainsi qu'en reconnaissance de mots isolés. Dans cet article nous présenterons des expériences de reconnaissance phonétique utilisant cette méthode, appliquée surtout aux parties transitoires du signal de parole.

Ces expériences constitueront une évaluation de la méthode de QV pour la reconnaissance phonétique.

2.DESCRIPTION DU SYSTEME

La fonction du système est de transformer le signal acoustique en une suite de symboles équivalents de taille phonémique. L'unité élémentaire à reconnaitre est l'élément " centiseconde " . Nous considérons que c'est un bon compromis entre la précision et la stationnarité.

Apprentissage

Nous pouvons le décomposer en deux parties : extraction du dictionnaire et étiquetage du dictionnaire.

Le signal de parole est analysé par LPC (ordre 14) , pondéré par une fenêtre de Hamming, la longueur étant 10ms. On a donc un vecteur de paramètres pour chaque centiseconde de signal, que l'on appelle un point. On utilise la QV pour construire le dictionnaire à l'aide d'un algorithme à seuil [1], Le principe est :

-le premier point crée la première classe ; pour classer le n ième point, on procède de la façon suivante : on calcule la distance entre ce point et chacune des m classes précédemment crées, la valeur minimale de ces distances sera comparée à un seuil. Si celle-ci est inférieure au seuil, alors le nouveau point sera intégré à la classe, sinon on créera une nouvelle classe m+1. On précise que l'on entend par distance entre un point et une classe la distance de ce point au représentant de la classe, celui-ci étant le centre de gravité. La valeur choisie pour le seuil est un paramètre important de la méthode. Les différentes distances essayées sont détaillées au paragraphe 3 .

Une fois le dictionnaire construit, on l'étiquette, c'est-à-dire que l'on va donner une interpretation phonétique à chaque classe. En procédant d'abord par l'étiquetage de chaque élément d'une classe on aboutira à l'étiquetage global de la classe. L'étiquetage d'un élément d'une classe est fait au moyen d'un tableau construit à partir des programmes de segmentation, par alignement, d'une part de la transcription phonétique du signal, d'autre part d'une suite d'étiquettes du type " voyelle ", " fricative " , " silence " , et "autre" . Ensuite pour chaque classe du dictionnaire on calcule pour chaque étiquette son pourcentage de présence dans la classe, on sélectionera comme étiquettes de la classe les trois étiquettes de plus fort pourcentage.

Reconnaissance

En reconnaissance, il s'agit d'abord d'identifier chaque centiseconde de signal, et ensuite de structurer le signal en phonèmes à partir des hypothèses faites sur chaque centiseconde. Nous utilisons l'algorithme des plus proches voisins pour l'accès au dictionnaire [1], dont le principe consiste à chercher dans le dictionnaire les k points de référence, qui minimisent leur distance au signal à reconnaitre

(on sélectionnera k=3). Comme le dictionnaire est étiqueté, on trouvera une étiquette phonétique pour le signal à reconnaitre. celui-ci est caracterisé par trois références qui elles mêmes possèdent trois étiquettes. On décidera donc la plus probable parmi ces neufs étiquettes.

3.EXPERIENCE ET DISCUSSION

D.Vicard et L.Miclet ont montré que cette méthode est satisfaisante pour les parties stables, en particulier les noyaux vocaliques [6]. Nous avons fait des expériences pour les parties transitoires, avec plusieurs ensembles de données, et plusieurs distances.

Expérience sur la base de donnée polysons

La base de données utilisée est formée de mots artificiels (par exemple A PL A TA), où il est facile d'isoler des segments du type groupe de consonne + voyelle (ici : PLA). Les mots sont prononcés par une voix masculine et le signal est échantillonné à 8 kHz.

La distance entre centisecondes utilisés est la distance euclidienne entre coefficients cepstraux provenant de l'analyse LPC (LPCC). On s'aperçoit très vite que même pour la reconnaissance sur le corpus d'apprentissage lui même, le résultat est médiocre pour les parties transitoires. En fait pour les noyaux vocaliques, on a plus de 80% de cas bien reconnus, par contre pour les plosives le pourcentage est faible, pour les liquides, il est inferieur à 40%, et pour les semi-voyelles autour de 75%. On peut déjà voir que quand le spectre n'est pas très stable, on obtient d'assez mauvais résultats.

L'interpretation détaillée des resultats sur ces segments particulièrement difficiles (groupe de consonne + différentes voyelles) étant complexe, nous avons procédé à la deuxième expérience sur des données plus simples.

Experience sur mots monosyllabiques

La base de données, est formée de mots monosyllabiques, de type consonne-voyelle, dont la voyelle est toujours /e/ :

ke ge re fe je te se ve we le de ne pe me ye be ʃe ue

On a dix séries de 18 mots, la seule différence entre elles étant l'ordre d'apparition. Ce corpus est prononcé par une voix masculine, échantillonné à 8 kHz, quantifié sur 12 bits. On ne s'interresse donc qu'ici la partie transitoire, le segment transitoire est extrait par le programme de segmentation, qui utilise l'énergie et le nombre de passages par zéro [4].

Les différents types de distances utilisées :

1)distance Tchebychev sur les a_i

$$d_{L_\infty} = \max_{i=1,p} |a_i - \hat{a}_i|$$

où a_i est le i ième coefficient de prédiction

2)distance Minkowski sur les a_i

$$d_{L_1} = \sum_{i=1}^{p} |a_i - \hat{a}_i|$$

3)distance Euclidienne sur les a_i

$$d_{L_2} = \left[\sum_{i=1}^{p} |a_i - \hat{a}_i|^2 \right]^{1/2}$$

4)distance Cepstrale

$$d_{LPCC} = 2 \sum_{i=1}^{p} (c_i - \hat{c}_i)^2$$

où c_i est le i ième coefficient du cepstre, obtenu par analyse LPC.

5)distance d'Itakura [2]

$$d_{Itakura} = \ln \frac{\hat{a} R \hat{a}^t}{a R a^t} = \ln \frac{\hat{a} R \hat{a}^t}{\alpha}$$

où $a=(1, a_1, \cdots a_p)$
α = residu
R = matrice d'autocorrelation du signal

6)rapport de vraisemblance [3]

$$d_{LR} = \frac{\hat{a} R \hat{a}^t}{a R a^t} - 1 = \frac{\hat{a} R \hat{a}^t}{\alpha} - 1$$

a)Résultats d'expérience sur 5 séries en apprentissage et 5 autres séries en reconnaissance : en fonction des différentes distances nous indiquons dans le tableau suivant les taux de reconnaissance :

Taux de Reconnaissance			
Distance	rang No.		
	1	2	3
d_{L_∞}	47%	52%	53%
d_{L_1}	45%	50%	54%
d_{L_2}	56%	60%	64%
d_{LPCC}	61%	66%	67%
$d_{Itakura}$	60%	63%	65%
d_{LR}	57%	63%	66%
d_{RC}	52%	53%	56%

b)Résultats en fonction de nombres de séries

d'apprentissage, pour la distance LPCC. (reconnaissance au premier rang)

nb app.	3	5	7
taux	55%	61%	65%

On constate que la distance cepstrale, puis la distance d'Itakura nous donnent meilleurs résultats. Ceci confirme ce qui avait trouvé par Davis et Mermelstein en reconnaissance de mots [5] . En examinant la matrice de confusion, on voit que les semi-voyelles, les nasales sont bien reconnues, au contraire des plosives et plus spécialement des plosives sourdes, qui ont de mauvais taux de reconnaissance. Ceci revient à dire que si les variations du spectre sont importantes, le résultat est mediocre. Quand on fait croitre la taille de l'ensemble d'apprentissage, le taux de reconnaissance augmente.

Notons aussi que le taux de reconnaissance est aussi sensible au seuil de la classification ; en effet, il existe un seuil optimal, tel que quand on fait augmenter la valeur du seuil, le taux de reconnaissance descend très vite, tandis que quand on baisse cette valeur, le taux de reconnaissance augmente très peu.

Les taux obtenus dans cette deuxième expérience sont bien meilleurs que dans la première, principalement parce que la voyelle est toujours la même. cela suggère l'idée que le décodage acoustique-phonétique doit être fait d'abord en reconnaissant les parties stables, ensuite les parties transitoires en tenant compte parties stables déterminées précédemment.

4.CONCLUSION

La quantification vectorielle peut nous donner de bons résultats pour les noyaux vocaliques, mais ce n'est pas une méthode convenable pour les parties transitoires, surtout pour celles qui ont des spectres très instables. Plus le spectre est instable, plus mauvais est le résultat. Par ailleurs, il est possible que le problème vienne du niveau de modélisation du signal. Les recherches sur la modélisation des parties transitoires du signal peuvent apporter de nouvelles lumières sur ce problème.

5.BIBLOGRAPHIE

[1] M.DABOUZ, "Transmission de la parole à faible debit par vocodeur à classification ", Thèse DDI, ENST, 1984

[2] F.ITAKURA, "Minimum Prediction Residual Principle Applied to Speech Recognition " , IEEE, Trans. ASSP, Vol. ASSP-23, No.1, Feb. 1975

[3] A.H.GRAY, "Distance measures for speech processing" , IEEE trans. ASSP, Vol ASSP-28, No 4, August,1980

[4] BENEDETTO, J.P.TUBACH, "Two Coopeative Methodes for the segmentation of running speech" , Congrès FASE/DAGA 82, Goettingen, RFA, Septembre, 1982

[5] S.B.DAVIS, P.MERMELSTEIN, " Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences " , IEEE, Trans. ASSP, Vol. ASSP-28, No.4, August, 1980

[6] D.VICARD, L.MICLET, " Steady Part Recognition of Continuous Speech for Acoustic-phonetic Decoding " , ICASSP 86, Tokyo

CONSTRUCTION DE REFERENCES-MOTS A L'AIDE DE DEMI-SYLLABES

APPLICATION A L'ITALIEN

Brigitte Flocon*, Louis Sauter*, Maura Colombo**

* Laboratores de Marcoussis, Route de Nozay, F-91460 Marcoussis
** CSELT, 274 Via Romoli, I-10148 Turin

ABSTRACT

Speaker-independent demi-syllable based recognizer has been developed for a French minimal pair vocabulary of 18 words. In the framework of an ESPRIT project, the algorithms have been extended to the Italian language. This paper reports the modifications that have been introduced, in order to perform this experiment and presents the first results. It also points out the kind of problems that may arise when one system has been developed for one language and is extended to another one.

INTRODUCTION

Un système de reconnaissance de mots isolés multi-locuteur utilisant des références globales obtenues par concaténation de demi-syllabes a été développé aux Laboratoires de Marcoussis. Ce système a été testé sur un vocabulaire de 18 mots formé de paires minimales. Cette étude a été réalisée dans le cadre d'un projet Européen du programme ESPRIT, le projet SPIN : Speech Interface at Office Workstation. En collaboration avec le CSELT à Turin, nous étudions l'extension de ce système de reconnaissance à la langue Italienne. Bien que toutes deux langues latines, le Français et l'Italien présente certaines disparités qui ont amenés des modifications dans les algorithmes de segmentation. Les problèmes de co-articulation ont également été examinés. Les premiers résultats des tests mettent en évidence certaines différences dues aux caractères particuliers de chacune des deux langues. Les études sont actuellement poursuivies afin de mieux tenir compte des éléments mis en évidence lors de cette première partie de l'étude.

LE SYSTEME DE RECONNAISSANCE

Le système de reconnaissance décrit ici a fait l'objet de plusieurs présentations par le passé [1], [2]. Nous reprendrons ici uniquement les points principaux, afin de situer précisement les travaux effectués sur l'Italien.

Le système comprend les modules suivants :

Analyse du signal de parole

Le signal est échantilloné à 8 kHz après filtrage passe-bas (fréquence de coupure : 3400Hz) et fenêtre de Hamming. Un jeu de 9 paramètres représentant les coefficients du cepstre (Mel Frequencies Cepstrum Coefficients) est déterminé toutes les 16 millisecondes.

Création des références

Un corpus de 35 mots [3] prononcés par 20 locuteurs est segmenté en demi-syllabes à l'aide d'un algorithme semi-automatique (supervisé par l'opérateur). L'algorithme de segmentation utilise les extrema de l'énergie du signal [4]. Il permet d'obtenir un ensemble de 87 demi-syllabes provenant de 20 locuteurs différents.

Dans une première phase, on recrée par concaténation et lissage chaque mot du vocabulaire de test à partir des demi-syllabes correspondantes pour chacun des 20 locuteurs.

Dans une deuxième phase, un algorithme de classification, utilisant la méthode des nuées dynamiques, détermine 5 classes pour chaque mot-test.

Dans une dernière phase, une référence est obtenue pour chaque classe, grâce à un algorithme de moyennage [3].

Algorithme de reconnaissance

L'ensemble de références est formé de 5 représentants par mot de vocabulaire. Chaque mot inconnu est comparé à toutes les références à l'aide d'un algorithme de comparaison dynamique (Sakoe et Chiba type 1). Le mot reconnu est celui dont la distance au mot inconnu est la plus petite.

Performances

Le vocabulaire de test qui a été choisi comprend des paires minimales. Il est donné dans la table I. Un corpus de test a été enregistré avec 10 nouveaux locuteurs. Le taux de reconnaissance du système est de 92%.

ADAPTATION A L'ITALIEN

Analyse

La chaîne d'enregistrement de parole disponible dans le laboratoire Italien comporte un filtrage passe-bande (fréquences de coupure : 300 et 3400 Hz). Afin de tenir compte de cet élément, il a été décidé de forcer à zéro les sorties des deux premiers filtres utilisés dans le calcul des MFCC.

Création des références

Un corpus de 22 mots prononcés par 20 locuteurs a été segmenté en demi-syllabes avec l'algorithme semi-automatique déja' employé sur le matérial vocal français ; il en résulte un ensemble de 81 demi-syllabes différentes. 44 demi-syllabes appartenant à cet ensemble ont été combinées de facon à former 21 nouveaux mots différents de ceux qu'on avait au départ et dans lesquels chaque demi-syllabe se trouve placée dans un contexte phonétique différent de celui d'où elle a été extraite.

Comme dans le cas de la langue française, la procédure de concaténation est appliquée séparement sur chaque locuteur de référence , en donnant naissance pour chaque mot du vocabulaire à 20 prototypes, qui sont réduits à 5 après classification automatique et moyennage.

Un deuxième corpus, composé des 21 nouveaux mots prononcés par 10 nouveaux locuteurs, a été utilisé lors de la phase de reconnaissance.

Programme de segmentation

Pendant la segmentation semi-automatique du corpus de référence on a remarqué un taux d'erreurs élevé sur les frontières entre les unités déterminées par l'algorithme (omission, adjonction, déplacement par rapport au point exact d'occurence). Cela dépend principalement des différentes situations prosodiques, morphologiques et phonétiques de la langue italienne par rapport à la langue française.

Ainsi, on remarque fréquemment :

1- *l'omission de la frontière qui sépare la dernière partie d'un mot en 2 demi-syllabes de type CV V- .*

Elle est due à la moindre énergie de la dernière voyelle par rapport aux précédentes (on rappelle que dans la langue italienne presque tous le mots se terminent par une voyelle qui n'est que très rarement accentuée, alors que dans la langue française les mots peuvent se terminer par une consonne aussi bien que par une voyelle, qui est en général accentuée.

On remédie à ce type d'erreur en introduisant de facon manuelle la frontière sautée sur le maximum d'énergie présente dans la partie du signal vocal considéré).

2 - *la confusion dans la segmentation des sons liquides ou "glides" intervocaliques, notamment /l/ et /r/ dans le cas du corpus considéré.*

La segmentation automatique a tendance, dans cette situation, soit à omettre la frontière soit à la mettre en position déplacée par rapport au centre du son consonantique.

Les difficultés surgissent du fait que les sons de ce type ont une courte durée, et en même temps que la forte coarticulation empêche de les isoler de façon fiable.

3 - *la confusion entre diphtongues et hiatus.*

On rappelle qu'une diphtongue (CSVC) est composée de deux unités du type : *CSV VC* alors qu'un hiatus (CVVC) est composé de 3 unités : *CV VV VC.*

En réalité les prononciations des deux types de rapprochements sont fréquemment échangées entre elles, et souvent on n'arrive pas à isoler la deuxième voyelle ; de plus, on observe des traits dépendants du locuteur.

4 - *la segmentation incorrecte de sons de type plosif non voisé, ou fricatif non voisé.*

Pour ces sons, la frontière devrait être placée immédiatement avant l'explosion, alors que l'algorithme a plutôt tendance à la positionner quelques millisecondes avant l'explosion. Une telle sorte d'erreur donne naissance a des problèmes dans la phase de programmation dynamique

5 - *Quelques doutes existent sur la segmentation des unités suivantes : as, da, fe, so*

Ces unités ont été extraites de contextes particuliers VCCV :

t r a s f e r i s c i
a dZ e n d a
K u r s o r e

où la forte coarticulation entre les consonnes rend difficile l'isolement des demi-syllabes.

TESTS COMPARATIFS

Les tables I et II donnent la liste des vocabulaires utilisés pour tester le système dans les deux langues.

Vocabulaire de test Français		
Cassette	Paté	Passoire
Termite	Traire	Traitre
Divin	Vitre	Mité
Mitre	Titre	Cité
Douter	Sente	Santé
Quantité	Fête	Fez

Table I : vocabulaire de test Français

Vocabulaire de test Italien		
ACCETTA	CALCE	CORSO
DANZANTE	DATA	FETTA
FISSO	LANCI	MELA
PASSO	PIATTA	QUELLA
RICCA	RINCORRERE	RINUNCIARE
SOLA	TANTE	TORRE
TRATTA	VELA	VETRI

Table II : vocabulaire de test Italien

Le taux de reconnaissance obtenu lors d'une première expérience sur l'Italien est de 75 %. Cette médiocre performance est à attribuer à 3 facteurs majeurs :

- l'insuffisance des règles de segmentation qui sont disponibles pour la langue italienne,
- la ressemblance phonétique entre certaines couples de mots,
- la plus grande longueur moyenne de presque tous les mots-test effectivement prononcés, par rapport aux prototypes correspondants obtenus par concaténation.

Les problèmes de segmentation ont été exposés dans le paragraphe précédent.

Une seconde expérience effectuée après une nouvelle segmentation (manuelle) qui tient compte des problèmes décrits au point 4 améliore le taux de reconnaissance du système qui passe de 75% à 80%.

La ressemblance phonétique est cause d'un deuxième ensemble d'erreurs : notamment il y fréquemment des confusions entre les couples ou triplets *MELA / VELA, DATA / TRATTA / PIATTA, CALCE / TANTE.*

Le troisième type d'erreurs est du à des problèmes de rapport de longueur entre le mot inconnu et les mots de références. En effet, lors de la création de l'ensemble de références et afin d'éliminer tout effet de co-articulation d'une demi-syllabe sur l'autre, certaines demi-syllabes ont été fortement écourtées. Il en résulte, lors de la concaténation, que plusieurs références sont anormalement courtes, comparées à des prononciations réelles des mêmes mots. Ce point est en particulier frappant pour les demi-syllabes de fin de mots, qui, pour l'Italien sont essentiellement des demi-syllabes de type CV.

Une analyse statistique de la longueur des demi-syllabes a été réalisée ; elle met en évidence l'influence de l'accentuation sur la longueur des demi-syllabes. Des extraits de cette analyse sont présentés dans la table III. Les mots choisis sont ceux pour lesquels les phénomènes de rapport de longueur sont les plus marquants. On remarquera également, pour ces mots, les différences d'accentuation entre les demi-syllabes utilisées pour créer les références et celles provenant du corpus de mots test.

CONCLUSION

L'étude est arrivée à un point où les premiers résultats sont disponibles. Ces résultats sont intéréssants, même s'ils ne permettent pas encore de donner les performances finales du système de reconnaissance pour l'Italien. Une étude plus approfondie de l'Italien est en cours afin de pouvoir tenir compte des caractéristiques phonétiques particulières de cette langue par rapport au Français.

REFERENCES

[1] L. SAUTER, "Isolated word recognition using a segmental approach", *ICASSP*, Tampa, March 1985.

[2] L. SAUTER, "Reconnaissance segmentale multilocuteur de mots isolés", *Congrès AFCET-Matériels et logiciels pour la 5ème génération*, Paris Mars 85.

[3] B. FLOCON, N. BRIANT, "SYRIL : Système de reconnaissance de mots indépendant du locuteur", *4ème Congrès AFCET-RFIA*, Paris, Janvier 84.

[4] D. FOHR, J.P. HATON, F. LONGCHAMP, L. SAUTER, "Methodes de segmentation syllabique en reconnaissance de la parole", *14èmes journées d'études sur la parole (GALF)*, Paris, Juin 1985.

mot : **RICCA**

	n^0	l_{min}	l_{max}	l_{moy}	acc.	rap.
ri	13	7	21	12	(1)	1.5
	13	12	24	18		
ik	13	5	17	10	(3)	3.7
	13	27	45	37		
ka	13	5	17	8	(1)	1.6
	13	8	18	13		
a-	13	5	18	9	(1)	2.3
	13	13	31	21		

mot : **SOLA**

	n^0	l_{min}	l_{max}	l_{moy}	acc.	rap.
so	16	22	66	33	(1)	1.0
	16	26	41	33		
ol	16	6	15	8	(3)	2.8
	16	9	32	23		
la	16	6	34	17	(4)	.6
	16	6	31	11		
a-	16	5	18	9	(1)	2.3
	16	10	28	21		

mot : **VELA**

	n^0	l_{min}	l_{max}	l_{moy}	acc.	rap.
ve	20	7	13	9	(3)	3.0
	20	19	37	27		
el	20	5	16	8	(3)	2.6
	20	9	37	21		
la	20	6	34	17	(4)	.8
	20	9	25	14		
a-	20	5	18	9	(1)	2.5
	20	14	31	23		

mot : **FISSO**

	n^0	l_{min}	l_{max}	l_{moy}	acc.	rap.
fi	7	9	18	11	(3)	2.5
	7	21	40	28		
is	7	5	18	11	(3)	2.4
	7	13	42	27		
so	7	22	66	33	(2)	.7
	7	20	34	25		
o-	7	4	20	9	(1)	2.5
	7	10	57	23		

mot : **MELA**

	n^0	l_{min}	l_{max}	l_{moy}	acc.	rap.
me	9	7	19	10	(3)	2.5
	9	16	38	25		
el	9	5	16	8	(3)	2.6
	9	10	35	21		
la	9	6	34	17	(4)	.5
	9	6	15	10		
a-	9	5	18	9	(1)	2.6
	9	12	36	24		

mot : **QUELLA**

	n^0	l_{min}	l_{max}	l_{moy}	acc.	rap.
kwe	12	10	29	16	(3)	1.0
	12	11	23	16		
el	12	5	16	8	(3)	2.8
	12	13	38	23		
la	12	6	34	17	(4)	1.2
	12	9	34	21		
a-	12	5	18	9	(1)	2.6
	12	11	35	24		

pour chaque demi-syllabe :

$1^{ère}$ ligne	mot référence
$2^{ème}$ ligne	mot test
n^0	numéro du mot
l_{min}	longueur minimum
l_{max}	longueur maximum
l_{moy}	longueur moyenne

acc.		accentuation
	(1)	identique unités test et référence
	(2)	unité test non accentuée, référence accentuée
	(3)	unité test accentuée, référence non accentuée
	(4)	variable pour unité référence
rap.		rapport longueurs moyennes test/référence

Table III : Extrait de la statistique
sur les longueurs des unités en Italien

SIBYLLE : UN OUTIL D'AIDE A LA CONSTRUCTION D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE

S. MINAULT, B. DUPEYRAT, M. INVERNIZZI

Centre d'Etudes Nucléaires de Saclay

Abstract :

This paper presents a knowledge based software tool for vocal signal study, based on an amplitude-time representation.

The system includes three main parts :

1) An AI Expert System
2) A rule compiler
3) A real time procedural system.

The set of these modules makes up a knowledge based software tool, which can be very helpful in the construction of a speech recognition system. It has been proved successfully for the sound-silence segmentation, voiced-unvoiced decision and pitch detection on continuous speech signals for several speaker.

INTRODUCTION

Les systèmes experts semblent être une voie prometteuse pour la reconnaissance de la parole. Ils facilitent en effet la représentation et la manipulation d'un grand nombre de connaissances qui peuvent être issues de différents domaines comme l'acoustique, la phonétique, la linguistique... Un certain nombre de systèmes ont été développés suivant cette approche [1,2] .

Cependant, si les systèmes experts permettent une grande souplesse, ils demeurent lents à l'exécution.

C'est pourquoi nous avons adopté la notion de système de développement [3] . Le système de développement se compose d'un système expert explicatif et d'un système procédural équivalent fonctionnant en temps réel. Il présente la particularité d'être simple et d'assurer un compromis entre le système expert souple mais lent et le système procédural classique, rapide mais figé.

L'outil que nous avons développé autour de cette notion a permis d'ébaucher la construction d'un module de reconnaissance acoustico-phonétique, à partir du signal vocal dans sa représentation amplitude-temps.

Nous aborderons tout d'abord le principe de la méthode de reconnaissance que nous avons adoptée. Nous présenterons ensuite la structure générale de notre outil. Puis nous décrirons le contenu de la base de connaissances qui constitue le résultat de l'expertise basée sur l'observation du signal vocal. Nous donnerons enfin les résultats obtenus pour la segmentation son-silence, la segmentation voisé-non-voisé et pour la détection du début du cycle de voisement.

LA METHODE DE RECONNAISSANCE

L'entrée du système est le signal vocal issu du microphone. Il est échantillonné à une fréquence f, choisie entre 8 kHz et 20 kHz, puis codé sous forme d'extrema. La suite de ces extremas, accompagnés de leurs paramètres caractéristiques (amplitude et durée), constitue notre représentation de base du signal vocal. Le codage des extrema permet de diminuer la quantité d'informations à traiter tout en conservant l'essentiel. Le signal reconstitué par interpolation linéaire des extrema reste parfaitement intelligible. La méthode de reconnaissance est basée sur l'observation du signal vocal dans sa représentation amplitude-temps. Un expert humain est en effet capable de reconnaître la suite des phonèmes prononcés : pour ce faire, il pose un certain nombre de diagnostics ou de critères basés sur la forme du signal temporel. Cette méthode avait été initialement développée par M. BAUDRY [4] qui avait formalisé son expertise sous forme d'un ensemble modulaire de règles de récriture. Ces règles sont regroupées en modules hiérarchisés assurant des fonctions de segmentation de plus en plus précises :

. segmentation silence-parole
. segmentation voisé non-voisé
. recherche du cycle de voisement
. segmentation en phonèmes
. identification des phonèmes.

STRUCTURE GENERALE DU SYSTEME SIBYLLE

Dans une première approche [3] , le système se composait d'un système expert et d'un système procédural généré automatiquement à partir de la base de connaissance.

Afin d'améliorer la rapidité du système procédural, nous avons conçu un module intermédiaire qui effectue une compilation préalable de l'ensemble des règles de la base de connaissances.

Le système expert

Il permet d'introduire les diagnostics de l'expert, formalisés sous la forme de règles de récriture. Chaque règle transforme une séquence d'éléments en une autre :

$$\left[nom_o, P_o\right]_{t-n}, \left[nom_1, P_1\right]_{t-(n-1)}, \ldots, \left[nom_n, P_n\right]_{t}$$

$$\longrightarrow \left[nom'_o, P'_o\right]_{t-p}, \ldots, \left[nom'_p, P'_p\right]_{t}$$

Chaque élément est caractérisé par un nom : nom_i, et un certain nombre de paramètres P_i.

La récriture peut être conditionnée par un ensemble de conditions sur les paramètres. D'autre part, les paramètres des éléments récrits sont déduits par l'intermédiaire d'opérations sur les paramètres des éléments d'entrée.

Ces conditions et opérations peuvent requérir l'appel d'une procédure externe (par exemple, le calcul d'histogrammes sur une fenêtre de temps) qui devra avoir été écrite préalablement.

La base de faits initiale, formée de la suite des extremas et de leurs paramètres caractéristiques, est transformée de proche en proche par l'application des règles de récriture. La base de faits finale sera composée à terme de la suite des phonèmes prononcés.

Le compilateur

Il est chargé d'effectuer la construction d'un graphe regroupant les sous-expressions similaires dans les règles, afin de tester l'applicabilité des règles en évitant les redondances dans les tests.

C'est à partir de ce graphe que s'effectue la génération automatique d'un programme équivalent au système expert : elle consiste à générer le code de programme correspondant au parcours du graphe.

Cette approche s'inspire de l'algorithme RETE [5] qui effectue la compilation d'un ensemble de règles de production. La forme en est cependant différente puisque les règles que nous traitons ne sont pas des "if-then rules" mais des règles de récriture.

Le gain de temps par rapport à la version précédente qui n'effectuait pas de compilation préalable, mais simplement la génération automatique d'un programme dans lequel les règles étaient traduites sous forme d'instructions FORTRAN est de 5 environ toutes choses étant égales par ailleurs.

Le système procédural

Il est généré automatiquement par le compilateur à partir du système expert.

Il comporte une partie fixe correspondant à l'acquisition du signal et une partie variable liée à la base de connaissances. Cette dernière partie est la programmation du parcours du graphe. Elle varie en fonction des modifications apportées aux règles.

Actuellement, la détection son-silence qui comprend 25 règles est un peu plus rapide que le temps réel. Par exemple, pour la phrase "pour se protéger du froid il s'est couché près de ma porte" le temps de traitement en segmentation son-silence (détection des silences occlusifs, des silences courts et des portions de parole) est de 2s C.P.U. sur un microprocesseur 68000.

LE CONTENU DE LA BASE DE CONNAISSANCES

L'ensemble de la base de connaissances est composée de différents modules hiérarchisés assurant des fonctions de segmentation de plus en plus précises.

Segmentation silence-parole

Le principe consiste à segmenter le signal de façon grossière en zones de faible amplitude et zones de plus forte amplitude. Les zones sont ensuite regroupées dans des segments de parole ou des segments de silence, suivant leur durée. Dans les segments de silence, on distinguera les silences courts des silences occlusifs et les silences plus longs (correspondant à une pause), suivant leur durée [4] .

Segmentation voisé-non voisé

Cette segmentation n'est effectuée que sur les zones détectées comme étant des segments de parole. Il s'agit ici de différencier les zones bruitées (sons non-voisés) des zones où le signal est quasi périodique (sons voisés). Cette différentiation se base sur deux critères :

- la durée des intervalles entre les passages par zéro du signal dérivé
- la durée des intervalles entre les passages par zéro du signal temporel.

On construit les histogrammes normalisés comprenant des classes de différentes durées. La décision voisé-non-voisé s'effectue d'après la forme de ces deux histogrammes [4] .

Détection du début du cycle de voisement

La détection du début du cycle de voisement (DCV) est effectuée sur les parties du signal détectées comme étant des segments voisés. Lorsqu'on examine visuellement les parties voisées du signal, on s'aperçoit que l'oeil détecte chaque cycle de voisement avec une grande sûreté. Il apparait que cette détection utilise deux critères principaux : régularité du cycle de voisement et maximum d'amplitude du DCV [4] . Lorsqu'il y a hésitation, on se base sur une partie stable du signal, à partir de laquelle on détermine chaque DCV.

Le principe de notre méthode de détection est de se baser sur un point d'ancrage dans une partie stable du signal dans laquelle les débuts du cycle de voisement sont suffisamment marqués : pratiquement ce point d'ancrage est l'extrema d'amplitude maximum et il correspond à un milieu de voyelle. Ensuite, à partir de ce point d'ancrage on effectue une recherche en arrière, puis en avant. Cette recherche est basée sur la combinaison de plusieurs critères : amplitude, variation d'amplitude, durée et régularité.

Les segments pour lesquels il n'a pas été détecté de début de cycle de voisement sont déclarés indéterminés.

La figure 1 représente les résultats de la détection du début du cycle de voisement pour le son "a". Une ligne verticale surmontée d'un "C" marque le début d'un cycle de voisement. "SI" signifie segment indéterminé.

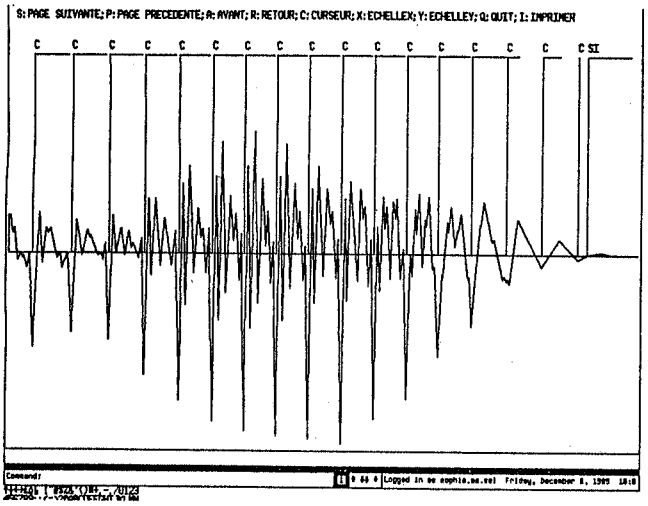


Figure 1 : Détection du début du cycle de voisement pour le son "a"

RESULTATS

L'ensemble du système a été utilisé pour la segmentation son-silence, la segmentation voisé-non-voisé et pour la détection du début du cycle de voisement.

Dans la phase de mise au point, le système expert a permis d'élaborer l'ensemble de la base de connaissances.

Le système procédural a ensuite été testé en parole continue sur un ensemble de phrases phonétiquement équilibrées (phrases de Combescure [6]) prononcées par un locuteur ; il a été testé en multilocuteur (1 femme- 1 homme) sur un vocabulaire de 16 mots prononcés chacun 4 fois.

On obtient de bons résultats pour la segmentation son-silence. Pour la segmentation voisé-non-voisé, il arrive que l'on ait des segments bruités qui soient déclarés voisés. Dans ce cas, il y aura échec au niveau de la détection du début du cycle de voisement : ce segment sera généralement déclaré comme indéterminé. Il arrive aussi, mais rarement, que l'on ait l'inverse : un segment voisé déclaré comme non-voisé (cas d'un "i" très bruité par exemple). Pour ce segment, il n'y aura donc pas de détection du début du cycle de voisement mais cette erreur pourra peut-être être relevée à un niveau supérieur.

En ce qui concerne la détection du début du cycle de voisement, nous avons effectué une série de tests afin d'évaluer les performances de notre jeu de règles. La définition du taux d'erreur s'est basée sur un critère déjà utilisé pour un autre algorithme de détection du DCV [7] : le taux d'erreur est le nombre de DCV mal étiquetés sur le nombre total de DCV. Dans les DCV mal étiquetés, sont inclus les DCV omis (DCV non trouvé alors qu'il y en avait un), les fausses alarmes (DCV trouvé alors qu'il n'y en avait pas), ainsi que les DCV mal placés (erreur de positionnement supérieur à 200 μs, par comparaison au positionnement que l'on effectuerait manuellement).

Les premiers résultats obtenus sont les suivants :

		taux d'erreur
Corpus 1	locuteur (homme) (phrases de Combescure)	1 %
Corpus 2	Locuteur 1 (femme)	3,4 %
	Locuteur 2 (homme) (vocabulaire 4 x 16 mots)	4,8 %

Le corpus 2 n'est qu'une faible partie d'un corpus enregistré beaucoup plus vaste : ces résultats ne constituent donc que des résultats préliminaires. Les résultats montrent un meilleur taux pour le corpus 1. Ceci est dû en partie au fait que la base de connaissances a été initialement élaborée sur ce corpus, mais aussi au fait que le signal est

échantillonné à 8 kHz alors que l'échantillonnage est de 10 kHz pour le deuxième corpus. Si l'on effectue une comparaison avec l'algorithme cité plus haut [7] , utilisant les mêmes critères de mesure, on s'aperçoit que le taux d'erreur obtenu ici est moins bon. Il s'avère cependant que l'algorithme cité procède en 3 passes (une première détection, puis deux passages supplémentaires de corrections des erreurs), alors que l'algorithme utilisé ici est direct. D'autre part, la base de connaissances est actuellement en cours de modification et les résultats énoncés plus haut ne sauraient être définitifs.

CONCLUSION

Le système SIBYLLE se compose d'un système expert, d'un compilateur et d'un système procédural. Le système expert permet l'introduction de connaissances sous forme déclaratives (règles). Le compilateur effectue la compilation de l'ensemble de la base de connaissances, puis génère automatiquement un système procédural équivalent au système expert. Le système procédural, rapide, peut être testé sur un grand corpus de données, en phase de développement. Il peut être alors nécessaire de revenir au système expert, pour modifier la base de connaissance.

Cet outil nous a permis d'élaborer plusieurs modules de connaissances : segmentation son-silence, segmentation voisé-non-voisé et détection du début de cycle de voisement.

L'élaboration de la base de connaissances se poursuit avec la segmentation voyelle-consonne, puis la segmentation en phonèmes avant d'aborder l'identification des phonèmes. Jusqu'à présent, la méthode de reconnaissance s'est basée sur la représentation temporelle du signal vocal, mais il n'est pas exclu de faire appel à une exploration fréquentielle, au niveau de l'identification des phonèmes en particulier. En effet, la structure des règles permettant l'appel à des fonctions externes, il est éventuellement possible d'y intégrer des procédures de calcul des coefficients de prédiction linéaire.

BIBLIOGRAPHIE

[1] GILLOUX M., MERCIER G., TANIDEC C. : Un système expert pour la reconnaissance de parole - CNET Lannion Colloque International d'I.A. Marseille 24/26.10.84

[2] FOHR D., CARBONELL N., HATON J.P. : SYSTEXP, un système expert pour le décodage acoustico-phonétique de la parole. 5èmes Journées Internationales sur les systèmes experts et leur application Avignon 13/14/15.5.85 P. 121-1238

[3] MINAULT S., INVERNIZZI M., DUPEYRAT B. : Système expert pour la reconnaissance de la parole par segmentation du signal - 5èmes Journées internationales sur les systèmes experts et leurs applications - Avignon 13/14/15.5.85 p. 1251-1266

[4] BAUDRY M. : Etude du signal vocal dans sa représentation amplitude-temps. Algorithme de segmentation et de reconnaissance de la parole - Centre d'Etudes Nucléaires de Saclay - Thèse d'Etat 15.6.78

[5] L. FORGY C. : RETE : A fast algorithm for the Many pattern/ Many objet Pattern Match problem - Artificial Intelligence - 19.9.82

[6] P. COMBESCURE : 20 lists of 10 phonetically balanced sentences - Revue d'Acoustique 1981, 14 (56) 34-38

[7] SPECKER P. : A powerful post-processing algorithm for time-domain pitch tracker -ICASSP 84 Proceedings, IEEE, vol 2, 18B.2.1 18B.2.4

ADAPTATION DE SYSTEME DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE A DE NOUVEAUX LOCUTEURS

Application des techniques d'analyse des Correlations Canoniques.

K.CHOUKRI[*,**] , G.CHOLLET[**] , Y.GRENIER[**]

[*] Laboratoires de Marcoussis, CRCGE. Route de NOZAY, 91460 Marcoussis.

[**] ENST-SYC, CNRS UA 820, 46 rue Barrault, 75013 Paris, France.

Abstract:

This paper describes various speaker normalization and adaptation techniques of a knowledge data base or reference templates to new speakers in automatic speech recognition (ASR). It focusses on a technique for learning spectral transformations, based on a statistical analysis tool (Canonical correlation analysis), to adapt a standard dictionary to arbitrary speakers which does not require prior knowledge about them. The proposed method should permit to improve speaker independance in Large vocabulary ASR. Application to an isolated digits recognizer illustrates the expected performances.

1-Introduction

La représentation mathématique du signal de parole est déduite de l'onde acoustique acquise dans différents environnements (microphone, bruit ambiant, ...). La production de la parole (vibrations des cordes vocales et transmission par le conduit vocal) dépend fondamentalement des caractéristiques physiologiques et articulatoires des locuteurs, de l'influence des contraintes sémantiques, syntaxiques et lexicales (compétence et aptitude linguistiques), de l'état physique du locuteur (fatigue, émotion, ...) ainsi que d'autres facteurs paralinguistiques.

Ces différences expliquent la variabilité inter-locuteur observée dans le signal de parole. On observe aussi une variabilité intra-locuteur, mais beaucoup moins importante, ce qui explique les meilleures performances des systèmes dépendents du locuteur par rapport aux systèmes indépendents des locuteurs. Cela explique aussi le biais introduit dans les mesures de distance spectrale.

Pour réaliser des systèmes de reconnaissance indépendents du locuteur, plusieurs axes de recherche sont actuellement explorés. On distingue trois grandes directions. La première, technique dite multi-références, tente d'atténuer l'influence de la variabilité inter-locuteur en augmentant le nombre d'archétypes associés à chaque son dans le dictionnaire de référence, de telle sorte que tous les locuteurs représentatifs de la population d'utilisateurs fassent partie des locuteurs d'apprentissage. Pour y parvenir on utilise différents artifices tels que chaînes de Markov, analyse discriminante, Clustering,...

La seconde méthode cherche des traits invariants tant au niveaux acoustique qu'articulatoire, voire même au niveau perceptuel et ne garde que ces paramètres pour la représentation de la parole.

La première technique est telle que l'acquisition, la sélection et le codage des références deviennent vite une longue et coûteuse procédure. En outre le dictionnaire de références résultant occupe une place mémoire substantielle et on ne fait pas appel à des données spécifiques à la parole. La seconde technique, quoique plus attrayante, a encore besoin de quelques années de recherche avant d'être operationnelle, en élaborant un modèle de l'influence des caractéristiques du locuteur et de ses habitudes articulatoires sur le signal observé.

Ce papier concerne la troisième technique qui est **l'adaptation d'un Système de Reconnaissance Automatique de la Parole (SRAP) de base à chaque nouvel utilisateur.**

Chaque individu écoutant pour la première fois la parole d'un locuteur inconnu a souvent besoin de s'adapter à la nouvelle voix (ou d'adapter son apparail de perception), et les premiers mots d'un dialogue n'apportent guère d'informations que celles nécessaires à cette adaptation. D'une façon similaire on peut envisager une adaptation de SRAP basé sur le dictionnaire spécifique à un locuteur, à d'autres locuteurs sans acquérir leur dictionnaire spécifique respectif. Ceci permettra d'utiliser des algorithmes qui ont fait leurs preuves et dont on connait les performances. Par ailleurs cette procédure peut être accomplie d'une façon dynamique [1], c'est à dire incorporée dans un système en configuration réelle d'exploitation ou d'utilisation.

2- Quelques approches de l'adaptation de SRAP au locuteur

Beaucoup d'auteurs qui s'intéressent aux problèmes dûs à la variabilité du signal de parole, en vue de réaliser une meilleure independence du SRAP vis a vis des utilisateurs, incorporent dans leurs systèmes une phase (ou un étage) d'adaptation au nouveaux locuteurs.

Cette variabilité semble avoir deux causes prédominantes, à savoir la structure anatomique de l'appareil phonatoire (cordes vocales, conduit vocal, ...) et les manières de mouvoir les organes articulatoires.

Quelques techniques de normalisation des paramètres représentant le signal de parole sont décrites si-dessous.

2.1 Normalisation de paramètres articulatoires:

Plusieurs auteurs ont tenté de minimiser l'influence du conduit vocal en procédant à une normalisation de paramètres liés à sa géometrie.

Kasuya & Wakita [2] ainsi que Wakita [3] déduisent, de quelques phrases, une longueur moyenne (l_r) du conduit vocal. Les paramètres considérés sont les quatres premiers formants et la longueur instantannée l du conduit vocal. Les fréquences formantiques sont alors normalisées par rapport à ces deux longueurs par :

$$F_i = \frac{l}{l_r} F_i$$

Cette technique réduit le taux de confusion dans l'identification de voyelles mais se heurte au problème d'extraction des formants dans le cas général.

2.2 Normalisation par ajustements fréquentiels:

Golibersuch [4] , essaye d'adapter son dictionnaire par un ajustement des formants pour une normalisation de leurs positions relatives. Il calcule à cet effet les corrélations entre les histogrammes (représentant les 3 premiers pics spectraux) du locuteur standard avec ceux du nouveau locuteur. Ensuite il réalise un ajustement linéaire par interpolation sur les fréquences centrales. Il se heurte aussi au problèmes d'extraction des formants .

Matsumoto & Wakita [5] proposent une méthode d'ajustement non linéaire, basée sur des algorithmes de programmation dynamique, pour trouver une transformation (non linéaire) de l'espace spectrale de référence sur l'espace utilisateur. Il proposent une distance normalisée en fréquence à partir d'une distance minimale parmi tous les choix possibles d'ajustements fréquentiels entre les spectres de références et le spectre de test.

Paliwal & Ainsworth [6] proposent une extention pure et simple de l'alignement temporel par programmation dynamique (DTW) au domaine fréquentiel. Cette méthode réduit aussi bien la distance spectrale entre des spectres équivalents qu'entre des spectres distincts.

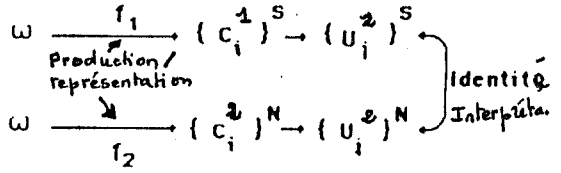
2.3 Normalisation et Quantification vectorielle:

La Quantification vectorielle (QV) est de plus en plus utilisée dans les SRAP [11]. Shikano & Raj Reddy [12] proposent une technique d'adaptation au locuteur par QV. L'idée de base est de trouver une correspondance entre deux dictionnaires de vecteurs (celui du locuteur standard et celui du nouveau locuteur). Cette correspondance peut ce faire par une substitution de table (vecteur A_i --> vecteur B_j) ou une combinaison linéaire avec un poids déterminé par un algorithme de programmation dynamique (A_i --> $\sum_j w_j B_j$).

3- Principe de la méthode proposée:

Le principe de la méthode proposée dans cet article est basé sur le constat qu'un même "son", produit par différents locuteurs, est interprété de manière identique par les personnes qui l'entendent malgré la variabilité inter-locuteur. On peut donc envisager un espace où des sons phonétiquement identiques seront représentés par des modèles identiques [1].

Si on considère des cepstres sur une échelle Mel (MFCC) comme paramètres représentant la manifestation acoustique de chaque mot, l'espace associé à chaque locuteur est donc, dans un premier temps, un espace cepstral où la variabilité inter-locuteur s'exprime pleinement. Si on considère un son ω (mot, syllabe, polyson, ...), on peut schématiser ces constats par la figure suivante où $\{C_i^j\}$, représente une succession de vecteurs cepstraux associée au locuteur j [7,8].



Production/Perception de la parole

Le problème de l'adaptation sera résolu si on arrive à déterminer les références $\{C_i^2\}$, associés au nouveau locuteur (2) à partir de celles associées à un locuteur standard (1). Il va de soi que nous ne connaîtrons jamais - à moins de refaire un apprentissage sur le locuteur 2 - les références exactes mais uniquement une estimation de celles-ci.

Au lieu de chercher des transformations directes $C_2^i = \Phi(C_1^i)$, on se propose de chercher des transformations qui permettent de définir l'espace commun U. Pour cela on va partir d'un échantillon représentatif des espaces paramètriques C_1 et C_2 , par exemple une phrase code ou un nombre limité de mots. Ensuite on va déterminer les projecteurs P_N et P_S tels que les projections soient identiques.

On se contentera dans un premier temps de transformations linéaires qui donnent des spectres projetés aussi proches que possible au sens d'un critère d'erreur. Si on choisit le critère des moindres carrés l'erreur de projection se traduit par l'équation (1):

$$J = \sum_i (u_i^1 - u_i^2)^T (u_i^1 - u_i^2) \qquad (1)$$

Il est facile de montrer à partir de cette équation qu'on peut minimiser l'écart entre les spectres projetés si et seulement si la corrélation entre les spectres associés est maximale, ce que réalise l'Analyse des Corrélations Canoniques [9], en fournissant les projecteurs P_N et P_S en question [1].

L'analyse canonique a pour but d'étudier la position relative d'un nuage de points par rapport à un autre (dans notre cas chaque nuage représentera l'échantillon d'un espace spectral d'un locuteur). Elle recherche des couples de variables, formés d'une combinaison des variables du premier nuage et d'une combinaison du second, les plus corréllés possible. Elle permet ainsi de définir un espace paramètrique où les projections de ces nuages coincident au mieux (au sens d'un critère d'erreur), qui sera alors une sorte d'espace "**typologique**" des deux locuteurs. On parle alors d'invariance des spectres par analyse des corrélations canoniques.

4- Procédure d'adaptation:

Pour valider notre propos on se propose d'appliquer cette méthode dans le cadre d'un système de reconnaissance de mots isolés avec un vocabulaire des **dix chiffres.**

Le spectre est paramétrisé avec 6 coefficients MFCC par trame. Durant la phase d'apprentissage chaque chiffre est prononcé une fois par un locuteur standard pour obtenir le dictionnaire de référence. La reconnaissance se fera grâce à des algorithmes de comparaison dynamique classiques, la détection de début et fin de mot est réalisée manuellement pour éviter toute erreur de détection pendant l'évaluation de cette méthode.

La première phase de la procédure d'adaptation consiste à acquérir et à aligner temporellement un échantillon représentatif de l'espace spectral associé à chacun des deux locuteurs. Il se pose alors le problème du choix de cet échantillon: que doit-on faire prononcer au nouveau locuteur comme "phrase code"?.

Dans une évaluation préliminaire cet échantillon sera réduit à un mot (le dixième du vocabulaire). les meilleurs mots semblent ceux qui reflètent le mieux la structure de l'espace phonétique (meilleure distribution dans le plan des premiers axes canoniques). Un logiciel d'analyse des corrélations canoniques permet alors de définir le nouvel espace de projection.

Grâce à ce logiciel on détermine la base génératrice du nouvel espace sur laquelle on projette le dictionnaire associé au locuteur standard pour obtenir le nouveau dictionnaire. On se retrouve dans le cas d'un "système monolocuteur" et on reprendra les algorithmes du système de base.

5- Evaluation:

Pour l'évaluation de cette méthode on dispose d'un corpus de 130 mots (comprenant les dix chiffres) prononcés par 100 locuteurs une seule fois. On cherche à évaluer la méthode dans le cadre d'un système monoréférence en insistant sur la variabilité inter-locuteur.

Des tests préliminaires ont pour but d'évaluer le système **non-adapté en mono-locuteur croisé**: le dictionnaire est obtenu grâce à un locuteur standard et on le teste sur des locuteurs choisis parmi les autres. Ensuite avec les mêmes données on évalue le système **après adaptation.**

Les taux de reconnaissance sont présentés en donnant les "bons" candidats qui sont reconnus en première position ou dans les deux premières positions avec l'intervalle de confiance correspondant à une probabilité d'erreur de 5%). Le taux de reconnaissance d'un système multi-référence utilisant les techniques de clustering (Syril) est de 93% en première position **[10]**.

Score avec intervalle de confiance		
	première position	deux premières
non adapté	70% (68,73)	84% (81,86)
adapté	87% (84,89)	92% (91,94)

6- Illustrations:

Une manière d'illustrer le passage d'un échantilon de l'espace spectral des deux locuteurs à leur espace commun est la synthèse des "mots" de ce dernier espace (les projections des mots enregistrés par le locuteur standard). Une synthèse intelligible (!) serait une bonne confirmation des resultats chiffrés ci-dessus. Une étude de la synthèse à partir de coefficients MFCC est en cours. Une autre façon d'illustrer la modification (ou l'amelioration) de la représentation paramètrique des données originales est l'évolution des distances locales sur les chemins optimaux lors de l'alignement temporel.

Les figures suivantes montrent le cas du mot "**deux**" (par une locutrice) qui a été reconnu comme étant le mot "**quatre**" (d'un locuteur) avant la phase d'adaptation et qui a été bien reconnu ensuite (la représentation de l'espace est réalisée par le mot "**profil**").

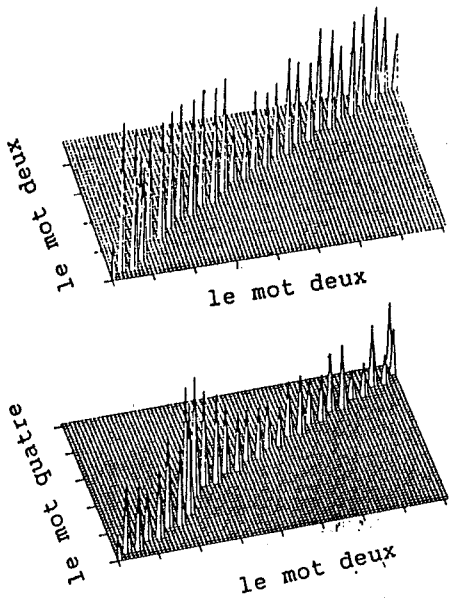


Figure de dtw avec distance locale avant adaptation.

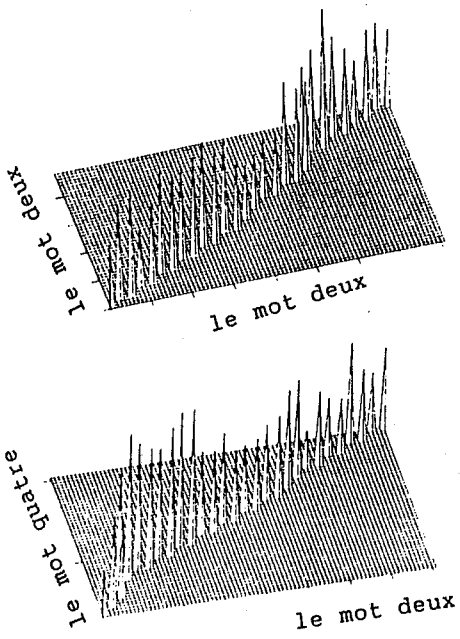


Figure de dtw avec distance locale après adaptation.

7- Conclusion:

Ce papier montre une adaptation de dictionnaires de formes à de nouveaux locuteurs. Une application à des systèmes mono-référence montre que les taux de reconnaissance sont améliorés de quelque 17%. Ce résultat reste à confirmer dans le cadre des systèmes Multi-références et de vocabulaire plus grands (130 mots).

Dans la suite des travaux on s'intéressera tout particulièrement à la représentation de l'espace cepstral de départ et au choix des phrases codes.

Une procédure dynamique supervisée par l'utilisateur, permettra de partir d'un seul mot, de mettre en oeuvre la phase d'adaptation et en fonction des resultats de parfaire la représentation de l'espace cepstral (en concaténant des mots, polysons,...) et donc l'espace de projection sans reprendre tous les calculs.

8- Références:

[1] **Choukri, K. , Chollet, G. & Grenier, Y. (1986),** Spectral transformations through canonical correlation analysis for speaker adaptation. in Proc. ICASSP, April 1986, Tokyo (to be published).

[2] **Kasuya, H. & Wakita, H. (1976)** Speech segmentation and feature normalization based on area functions. in proc ICASSP, Philadelphia, pp.29-32.

[3] **Wakita, H. (1977)** Normalization of vowels by Vocal-Tract Length and its application to vowel Identification. in IEEE ASSP, April, ASSP-25, no 2, pp 566-569.

[4] **Golibersuch, R.J. (1983)** Automatic prediction of linear frequency warp for speech recognition. in proc ICASSP, Boston, pp.769-772.

[5] **Matsumoto, H. & Wakita, H. (1979)** Frequency warping for nonuniform talker normalization. in proc ICASSP, Washington, pp.566-569.

[6] **Paliwal, K.K. & Ainsworth, W.A. (1985)** Dynamic frequency warping for speaker adaptation in automatic speech recognition. in Journal of Phonetics 13, pp.123-134.

[7] **Grenier, Y. (1980),** Speaker adaptation through canonical correlation analysis. in Proc. ICASSP, Denver, pp.888-891.

[8] **Grenier, Y., Miclet, L., Maurin, J.C. & Michel, H. (1981),** Speaker adaptation for phoneme recognition. in Proc. ICASSP, Atlanta, pp.1273-1275.

[9] **Golub, G.H. (1970),** Matrix decomposition and statistical calculations. in Statistical computation, Edited by Milton, R.C. & Nelder, Y.A. (Academic press), PP. 365-397.

[10] **Flocon, B.** and **Briant, N. (1984),** SYRIL: système temps réel de reconnaissance de mots isolés indépendant du locuteur, 4ème congrés AFCET RFIA, Paris.

[11] **Gray, M. R. (1984),** Vector Quantization. IEEE, ASSP, April 1984.

[12] **Shikano, K.** and **Raj Reddy, K. L., (1986).** Speaker Adaptation through Vector Quantization, in Proc. ICASSP, April 1986, Tokyo (to be published).

EVALUATION AUTOMATIQUE DES SEUILS DE REJET

O.DIOURI, JL.GAUVAIN

L.I.M.S.I. - CNRS, B.P.30 91406 ORSAY Cedex

Résumé

Le but de ce travail est de pouvoir rejeter des mots prononcés n'appartenant pas au lexique appris. Pour cela, des valeurs des seuils de rejet sont ajustées de façon à minimiser le taux de fausses acceptations sans pour autant augmenter celui de faux rejets (mots du vocabulaire rejetés à tort). Lorsqu'il s'agit de traiter des mots acoustiquement voisins, ce problème de rejet devient plus ardu. L'ajustement des valeurs des seuils représente une tâche très délicate et nécessite une certaine expérience de la part de l'utilisateur, car les seuils dépendent à la fois du vocabulaire et du locuteur. Nous avons utilisé un modèle stastistique afin d'évaluer automatiquement les seuils de rejet en fonction du locuteur et du contenu acoustique de chaque mot du vocabulaire. Les résultats obtenus sont concluants et montrent comment le taux d'erreurs peut être réduit de façon considérable par rapport aux méthodes basées sur un seuil évalué par locuteur ou pour tout le vocabulaire et tous les locuteurs. Une comparaison entre les résultats de l'évaluation automatique et ceux obtenus avec un seuil évalué a posteriori, montre que ne nous sommes pas loin du maximum qu'on peut atteindre.

1. Introduction

L'aptitude d'un système de reconnaissance, de séparer entre mots hors vocabulaire et mots du vocabulaire, dépend du corpus de test. Lorsqu'il s'agit de traiter un corpus de test contenant des mots voisins avec les mots du vocabulaire de références, un ajustement expérimental, ne suffit pas pour pouvoir rejeter des mots hors vocabulaire, car les valeurs des seuils sont choisies en général, de façon à minimiser le taux des faux rejets. Le choix du vocabulaire a donc un effet très important sur les performances d'un système de reconnaissance. Dans tous les cas, l'ajustement des seuils nécessite une certaine expérience de la part de l'utilisateur, et devient ardu lorsqu'il s'agit de traiter des mots voisins. Afin de baisser le taux d'erreurs, à l'intérieur des vocabulaires difficiles, il est nécessaire que l'évaluation des seuils soit effectuée automatiquement en fonction du locuteur et du contenu acoustique de chaque mot du vocabulaire. Pour cela, nous avons utilisé un modèle statistique[1].

2. Approche statistique: description de la méthode

Etant donné un vocabulaire de références V, constitué de mots isolés, $V = (R_1, R_2, ..., R_N)$, lorsqu'un mot test T est comparé avec toutes les références, une note globale est donnée pour chaque mot R_i, et la référence dont la note D_i est minimale, est identifiée, soit R_m. Un seuil $\gamma(m)$ est choisi de façon que, si D_m est inférieure ou égale à $\gamma(m)$, alors le test T est reconnu mot du vocabulaire. Sinon, T est rejeté et considéré comme mot hors vocabulaire. A partir de ces deux décisions, on peut définir une probabilité d'avoir un *faux rejet* P_{fr} (dans le cas où T est rejeté alors qu'il appartient au vocabulaire), et une probabilité d'avoir une *fausse acceptation* P_{fa} (dans le cas où T est reconnu alors qu'il est hors vocabulaire). Si le seuil $\gamma(m)$ est choisi trop grand, la probabilité P_{fa} augmentera alors que P_{fr} baissera. Par contre, si le seuil est choisi trop petit, la probabilité P_{fa} baissera alors que P_{fr} augmentera. Notre critère consiste à faire un compromis entre les deux valeurs du seuil, en choisissant un seuil qui égalise les deux probabilités P_{fr} et P_{fa}, soit un seuil à erreurs égales. On note par M_1 et σ_1, la moyenne et l'écart type entre les différents énoncés du même mot R_i (intraclasse) respectivement. De même, on note par M_2 et σ_2, la moyenne et l'écart type entre le mot R_i et tous les les mots $R_j, j \neq i$ du vocabulaire (interclasse) respectivement.

La probabilité P_{fr} peut être évaluée en fonction des distributions intraclasses comme suit (équation(2.1)):

$$P_{fr} = \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x-M_1)^2/2\sigma^2 1}\, dx = erfc\left(\frac{\gamma - M_1}{\sigma_1}\right)$$

où γ est le seuil choisi pour le mot R_i et $erfc$ est la fonction erreur standard.

De la même façon, la probabilité P_{fa} peut être évaluée en fonction des distributions interclasses (équation(2.2)) :

$$P_{fa} = \int_{-\infty}^{\gamma} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(x-M_2)^2/2\sigma^2 2}\, dx = erfc\left(\frac{M_2 - \gamma}{\sigma_2}\right)$$

Par définition du seuil à erreurs égales on a : $P_{fr} = P_{fa}$

Des équations (2.1) et (2.2), on déduit le résultat cherché :

$$\gamma = \frac{M_1 + M_2 \frac{\sigma_1}{\sigma_2}}{1 + \frac{\sigma_1}{\sigma_2}} \qquad (2.3)$$

A l'aide de l'équation (2.3), les seuils de rejet sont évalués automatiquement pour chaque mot du vocabulaire.

3. Méthode retenue

Pour utiliser un tel modèle, il est nécessaire que le corpus d'apprentissage soit très large (plusieurs énoncés pour chaque mot), afin que le calcul des distributions soit significatif. Notre corpus ne comporte que cinq énoncés pour chaque mot du vocabulaire, il est donc impossible d'utiliser le modèle statistique tel qu'il a été decrit plus haut. Pour le calcul des distributions interclasses, le problème ne se pose pas, car le corpus intermot est assez large pour chaque mot du vocabulaire. Pour les distributions intraclasses, l'évaluation de la moyenne intraclasse peut être significative avec dix valeurs de distances pour chaque classe, mais l'évaluation de l'écart type intraclasse avec dix valeurs n'est vraiment pas envisageable. Une solution à ce problème est de moyenner les différentes distributions intraclasses en déterminant ainsi une moyenne et un écart type intraclasse communs à toutes les classes de mots. Soient, pour chaque mot R_i du vocabulaire, $(M_1(i),\sigma_1(i))$ et $(M_2(i),\sigma_2(i))$ les distributions intraclasse et interclasse respectivement. Les distributions moyennes intraclasse sont alors:

$$M_1 = \frac{1}{N}\sum_{i=1}^{N} M_1(i) \qquad \text{et} \qquad \sigma_1 = \frac{1}{N}\sum_{i=1}^{N} \sigma_1(i)$$

L'équation (2.3) devient :

$$\gamma(i) = \frac{M_1 + M_2(i) \frac{\sigma_1}{\sigma_2(i)}}{1 + \frac{\sigma_1}{\sigma_2(i)}} \qquad (3.1)$$

dans le cas où les moyennes et les écarts types intraclasse sont moyennés, ou bien :

$$\gamma(i) = \frac{M_1(i) + M_2(i) \frac{\sigma_1}{\sigma_2(i)}}{1 + \frac{\sigma_1}{\sigma_2(i)}} \qquad (3.2)$$

dans le cas où, sont moyennés seulement les écarts types intraclasses. Toutes ces techniques sont expérimentées dans le paragraphe suivant, en retenant comme modèle, les deux équations (3.1) et (3.2).

4. Expériences et résultats

4.1. Conditions expérimentales

Pour les différentes expériences réalisées au cours de cette étude, l'analyse acoustique est réalisée par un banc de 16 filtres couvrant la bande de fréquence 100-5000Hz et dont les fréquences centrales sont réparties selon une échelle de Bark. L'algorithme de comparaison dynamique est utilisé pour calculer les distances entre les mots et les ajuster temporellement [2] à l'aide d'une équation récursive symétrique. Nous avons retenu un vocabulaire de références V_1, contenant 16 mots dont les chiffres et les six mots d'un langage de commande:

je-répète, écoute, exécute, menu, départ, arrêt,

et un corpus de test équilibré, comprenant 80 mots du vocabulaire appris V_1, et 80 mots hors vocabulaire V_2. Quatre locuteurs ont participé aux tests, chacun a prononcé 10 fois le vocabulaire V_1 et une fois le vocabulaire V_2. Les cinq premières séries de V_1, sont utilisées pour l'apprentissage et toutes les autres de V_1 et V_2, sont utilisées pour les tests de reconnaissance. L'ensemble de références est obtenu par un apprentissage robuste [3][4], qui consiste à moyenner ensemble les deux énoncés du même mot, les plus proches au sens d'une distance. Les tests sont réalisés dans un contexte monolocuteur et en mots isolés.

4.2. Seuil absolu a postériori

Dans cette expérience, un seuil de rejet est fixé pour tout le vocabulaire et tous les locuteurs. En faisant varier ce seuil, nous avons pu déterminer l'optimum pour lequel le taux fausses acceptations est égale à celui de faux rejets. La figure (1a) montre la variation de ces deux taux en fonction du seuil absolu évalué pour tous les locuteurs. La variation du seuil optimum est de 72.5 à 95 selon le locuteur. En moyennant les deux courbes, faux rejet et fausse acceptation, la figure (1b) montre la variation du taux d'erreur en fonction du seuil absolu (de 2 à 10% selon le locuteur). On remarque sur ces courbes que, la plage moyenne du seuil à taux d'erreurs stable est de $7db$ (plus ou moins large selon le locuteur). Cette notion de plage du seuil est intéressante car, elle permet à l'utilisateur de choisir un seuil en fonction de son vocabulaire et de l'application qu'il veut réaliser, selon qu'il veut moins de faux rejets ou moins de fausses acceptations. En conclusion à cette expérience, les résultats obtenus montrent en moyenne,

qu'avec un seuil absolu évalué a postériori pour chaque locuteur, le taux d'erreur est de 6% [4], alors qu'avec un seuil évalué a postériori pour tous les locuteurs, le taux d'erreur est de 7% comme le montre la figure (1b).

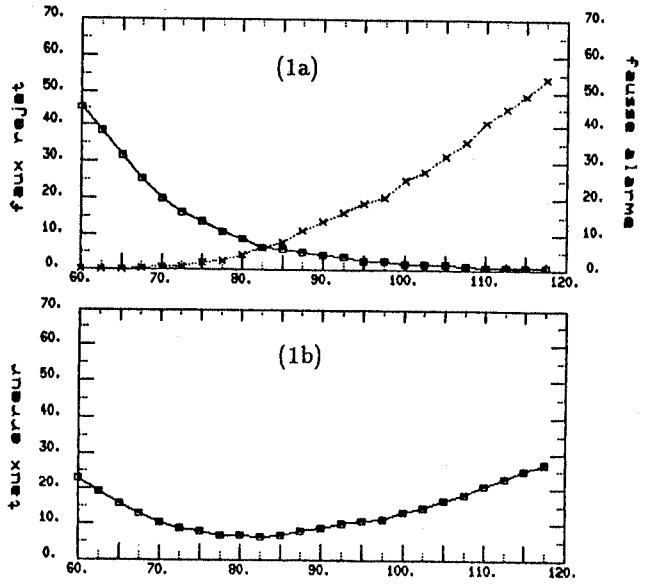


Figure 1: Taux d'erreurs en fonction du seuil de rejet absolu évalué pour tous les locuteurs.

4.3. Seuil de rejet a priori

Pour cette expérience, l'évaluation des seuils de rejet est effectuée automatiquement. Quatre types de tests ont été réalisés avec:

- un seuil évalué a priori par mot et par locuteur par l'équation (3.1)
- un seuil évalué a priori par mot et par locuteur par l'équation (3.2)
- un seuil par locuteur mais pour tout le vocabulaire, évalué a priori en moyennant pour chaque locuteur, toutes les distributions intraclasses entre elles et les distributions interclasses entre elles.
- un seuil pour tout le vocabulaire et pour tous les locuteurs, évalué a priori en fonction des distributions interlocuteurs (en moyennant les distributions obtenues pour chaque locuteur). Les tableaux (4.1), (4.2), (4.3) et (4.4) illustrent ces quatre tests:

	JLG	*JM*	*DB*	*ME*	*global*
F.R.	0	1	12	2	3.7
F.A.	13	22	6	13	13.5
$\frac{FR+FA}{2}$	6.5	11.5	9	7.5	8.6

Tableau 4.1: taux d'erreur avec un seuil a priori pour tous les locuteurs.

	JLG	*JM*	*DB*	*ME*	*global*
F.R.	7	3	2	1	3.2
F.A.	1	8	30	20	14.7
$\frac{FR+FA}{2}$	4	5.5	16	10.5	9

Tableau 4.2: taux d'erreur avec un seuil a priori par locuteur.

	JLG	*JM*	*DB*	*ME*	*global*
F.R.	8	7	1	1	4.2
F.A.	3	15	30	17	16.2
$\frac{FR+FA}{2}$	5.5	11	15.5	9	10.2

Tableau 4.3: taux d'erreur avec un seuil a priori par mot et par locuteur : équation (3.1).

	JLG	*JM*	*DB*	*ME*	*global*
F.R.	1	0	1	0	0.5
F.A.	0	6	20	11	9.2
$\frac{FR+FA}{2}$	0.5	3	10.5	5.5	4.8

Tableau 4.4: taux d'erreur avec un seuil a priori par mot et par locuteur : équation (3.2).

Le tableau (4.4) présente les meilleurs résultats obtenus parmis les différentes expériences avec un taux d'erreurs de 5% en utilisant l'équation (3.2). Les résultats obtenus montrent l'interêt de l'évaluation automatique des seuils de rejet en fonction du contenu acoustique de chaque mot du vocabulaire. Le taux d'erreurs a baissé de 4% par rapport à la méthode utilisant un seuil moyen pour tous les mots et pour chaque locuteur, et de 3.5% par rapport à la méthode utilisant un seuil commun à tous les locuteurs. En utilisant l'équation (3.1), le tableau (4.3) montre des résultats moins performants que ceux du tableau (4.4). Un tel résultat explique tout à fait le problème posé lorsqu'on moyenne des distributions moyennes intraclasses qui diffèrent de manière significative.

En effet, l'évaluation d'un seuil par mot n'a plus aucun sens puisque la distribution moyenne commune à toutes les

classes, ne dépend plus du contenu de la classe considérée. Si la moyenne commune est supérieure à la moyenne de la classe, le seuil de cette classe sera plus large, ainsi le taux de fausses acceptations augmentera. Si la moyenne commune est inférieure à la moyenne de la classe, le seuil de la classe sera plus petit, ainsi le taux de faux rejets augmentera. Dans les deux cas, le taux d'erreurs augmente comme le montre le tableau (4.3). Ce résultat montre que le fait d'utiliser une moyenne pour chaque mot est primordiale pour le modèle. D'autre part, ces résultats dépendent énormèment du locuteur. Le tableau (4.2) montre des variations considérables d'un locuteur à un autre. Par exemple, le locuteur DB pour lequel le taux d'erreurs est le plus élevé, présente un cas particulier car son apprentissage est incomplet : les énoncés du même mot dans le corpus de tests ayant des durées différentes de manière significative. Un processus de compression temporelle serait nécessaire, pour réduire les parties stables du mot [5].

4.4. Comparaisons expérimentales

Mettant à profit les différents résultats obtenus dans les deux expériences, il est clair que, l'évaluation automatique des seuils de rejet a priori en fonction du contenu acoustique de chaque mot du vocabulaire, permet de réduire au mieux le taux d'erreur par rapport aux méthodes utilisant un seuil par locuteur ou un seuil pour tous les locuteurs. Le test avec un seuil évalué a postériori par mot ne peut être réalisé, car avec cinq énoncés pour chaque mot, on ne peut pas fixer un seuil expérimentalement d'une part, et d'autre part, l'évaluation d'un taux d'erreurs optimal avec dix tests intramots n'est pas significative. Nous ne pouvons donc comparer que les résultats obtenus avec un seuil par locuteur, et ceux obtenus avec un seuil pour tous les locuteurs. Ces deux comparaisons sont illustrées par les tableaux (4.5) et (4.6):

seuil	FR	FA	$(FR+FA)/2$
a postériori	7	7	7
a priori	3.7	13.5	8.6

Tableau 4.5: taux d'erreur en fonction d'un seuil évalué pour tous les locuteurs.

seuil	FR	FA	$(FR+FA)/2$
a postériori	6	6	6
a priori	3.2	14.7	9

Tableau 4.6: taux d'erreur en fonction d'un seuil évalué par locuteur.

Conclusion

En évaluant automatiquement les seuils de rejet en fonction du locuteur et du contenu acoustique de chaque mot du vocabulaire, nous avons montré comment le taux d'erreurs peut être réduit d'une façon considérable par rapport aux méthodes basées sur un seuil évalué par locuteur ou pour tout le vocabulaire et tous les locuteurs. Les résultats obtenus avec l'évaluation automatique, ne sont pas loin du maximum qu'on peut atteindre avec un seuil évalué a postériori et la méthode est facilement implantable, avec une procédure qui diffère selon la taille du vocabulaire.

Références

[1] N.Sirnov, "Table for estimating the goodness of fit of empirical distributions", *Ann. Math. Stat. Vol 19 1948.*

[2] H.Sakoe et S.Chiba, "Dynamic programming optimisation for spoken word recogntion", *IEEE Trans. Acoust., Speech, and Signal Proc., ASSP-26, 1978.*

[3] L.R.Rabiner, A.E.rosemberg, S.E.Levinson et J.G.Wilpon, "Speaker independant recognition of isolated words using clustering techniques", *IEEE Trans. Acoust., Speech, and Signal Proc., ASSP 79.*

[4] O.Diouri, "Reconnaissance globale discriminante", thèse 3ème cycle - *Septembre 1985.*

[5] J.L.Gauvain, J.J.Mariani, J.S.Lienard, "On the use of time compression for word-based recognition", *Proc. IEEE ICASSP-83.*

ANALYSE

Président

Jean CAELEN

C.E.R.F.I.A. de Toulouse

DETECTION DU FONDAMENTAL PAR PRETRAITEMENT AMDF ET PROGRAMMATION DYNAMIQUE

Gérard BAILLY

INRS-Télécommunications
Université du Québec
Ile-des-Soeurs, Québec, Canada, H3E 1H6

Abstract

We present here a pitch detection algorithm using AMDF (Average Magnitude Difference Function) preprocessing. Classically a certain number of candidates are selected by peak detection for every analysis frame. The originality of the proposed method lies in the strategy of selection of the final hypothesis: a plausibility is associated to each candidate by a local DP (Dynamic Programming) applied to the AMDF. A global DP is then applied to the lattice of candidates after voicing decision. Tested on noisy speech with spontenous recordings, this algorithm achieves less than 1 % gross pitch errors.

1. Introduction

De nombreuses études ont été entreprises afin d'effectuer une comparaison objective entre les divers algorithmes de détection de fondamental existants [1, 2, 3]. Les résultats obtenus semblent privilégier les méthodes de détection temporelles [4, 5] aux méthodes fréquentielles [6, 7] qui ont grande difficulté à suivre les variations rapides du fondamental et n'ont qu'une faible immunité au bruit: de nombreuses publications récentes privilégient la détection par fonction AMDF [8, 2] pour sa simplicité d'implémentation sur microprocesseur par rapport à la méthode d'autocorrélation [9]. De plus, le prétraitement du signal par "clippage" ou filtrage inverse LPC permet d'obtenir une plus grande immunité au bruit [2] et aux variations brutales de la structure formantique. L'algorithme présenté ici utilise le prétraitement par clippage et AMDF, mais la stratégie de décision pourrait s'appliquer à d'autres types de prétraitement.

L'essentiel de notre travail a porté sur la stratégie de décision : alors que la plupart des algorithmes de décision mis en oeuvre par les systèmes pré-cités sont contextuels et déterministes, la stratégie que nous proposons est basée sur une programmation dynamique évoluant sur un treillis de candidats.

Cette stratégie permet d'ancrer la détection du fondamental sur les parties très résonnantes du signal, laissant à la programmation dynamique globale le soin d'assurer les contraintes de continuité sur la courbe. Ceci était auparavant réalisé au moyen d'algorithmes de lissage [10], en espérant que les erreurs de détection étaient suffisamment "dispersées" [8]. Mais ces algorithmes de lissage détruisent en même temps les effets microprosodiques (ex: hausse de F0 après une plosive sourde, baisse lors d'une occlusion ou constriction sonore du conduit vocal...) indices extrêmement importants en reconnaissance aussi bien qu'en analyse de la parole.

2. Prétraitement

Le prétraitement du signal d'entrée $e(t)$ consiste d'une part à fournir au détecteur un signal où la structure formantique a été minimisée et les variations d'amplitude à l'intérieur de la fenêtre d'analyse ont été corrigées, d'autre part à fournir deux indices à l'algorithme de décision final : EnrBF (énergie inférieure à 1 KHz) et Friction (quotient de l'énergie supérieure à 1 KHz et de EnrBF). La correction d'énergie consiste à calculer la pente d'énergie P à l'intérieur de la fenêtre de largeur L_f. Le signal d'entrée corrigé $e'(t)$ sera alors égal à :

Pour $0 \leq a \leq L_f$,

$$e'(t+a) = e(t+a).\begin{cases} |1 - \frac{(1-1/P)}{L_f}.a| & \text{si } P \leq 1 \\ |P - \frac{(P-1)}{L_f}.a| & \text{si } P \geq 1 \end{cases} \quad (1)$$

Le signal ainsi obtenu est ensuite clippé à un rapport de 0.3. Le signal prétraité $s(t)$ sera donc égal à :

$$s(t+a) = \begin{cases} e'(t+a) - S_{max} & \text{si } e'(t+a) \geq S_{max} \\ 0 & \text{si } S_{min} \leq e'(t+a) \leq S_{max} \\ e'(t+a) - S_{min} & \text{si } e'(t+a) \leq S_{min} \end{cases} \quad (2)$$

avec:

$$\begin{cases} S_{max} = 0.3. \max\limits_{0 \leq a \leq L_f} e'(t+a) \\ S_{min} = 0.3. \min\limits_{0 \leq a \leq L_f} e'(t+a) \end{cases}$$

3. Détection et Quantification

Le but de cette étape est de fournir un treillis de possibilités à l'algorithme de décision final. Les valeurs de la fonction AMDF pour des décalages correspondant aux fréquences fondamentales de $F0_{min} = 60$ à $F0_{max} = 500$Hz, sont calculées afin d'obtenir un certain nombre N_t de candidats $F0_{t,i}$, $1 \leq i \leq N$) par détection de minima pour la fenêtre d'analyse $[t, t+L_f]$. A chaque pic $T0_{t,i} = \frac{Fe}{F0_{t,i}}$ est associé sa largeur $[T1_{t,i}, T2_{t,i}]$ (cf.Fig. 1). La fréquence fondamentale est supposée pour cette première sélection stationnaire sur la fenêtre d'analyse. Cette hypothèse ne tient pas compte de la possibilité de variation de F0 et de distorsion de phase à l'intérieur de la fenêtre [8]. Une programmation dynamique locale est pour ceci appliquée au voisinage de chaque candidat dans la limite de sa largeur de pic. On considère pour ceci l'AMDF comme la distance cumulée d'un parcours de la fenêtre d'analyse à décalage constant $T0_{t,i}$ pour une fréquence d'échantillonnage Fe :

$$\text{AMDF}(T0_{t,i}) = \frac{\sum\limits_{a=0}^{L_f - T0_{t,i}} |s(t+a) - s(t+a+T0_{t,i})|}{L_f - T0_{t,i}} \quad (3)$$

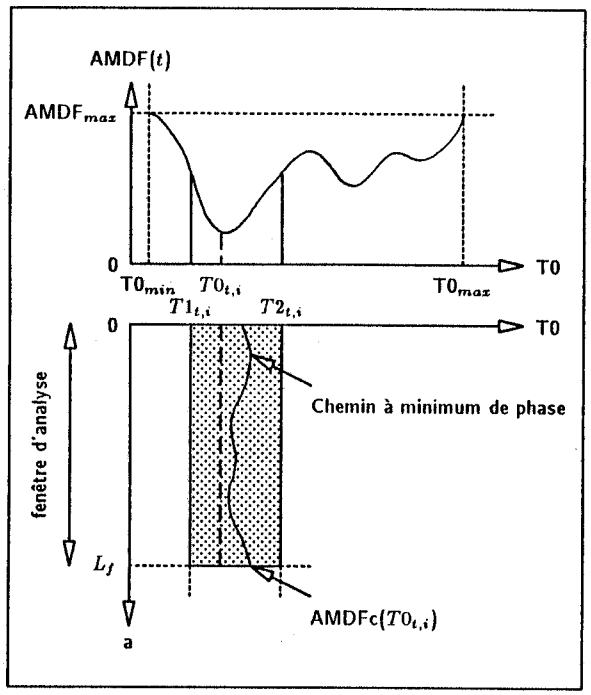


Fig. 1 AMDF et programmation dynamique locale à l'instant t.

L'AMDF modifiée prend en considération les variations de phase du signal afin d'optimiser globalement la somme des distances entre le signal et le signal décalé de $T0_{t,i}$. Une programmation dynamique locale (cf.Fig. 1) est effectuée au voisinage de chaque candidat et dans la limite de sa largeur de pic :

Pour $T1_{t,i} \leq T \leq T2_{t,i}$.

$$\text{AMDF}'(T,a) = \begin{cases} |s(t+a) - s(t+T+a)| & \text{si } a = 0 \\ \min\limits_{T-1 \leq T' \leq T+1} \left(\dfrac{\text{AMDF}'(T',a-1)+}{|s(t+a)-s(t+T'+a)|} \right) & \text{si } 0 < a \leq L_f \end{cases} \quad (4)$$

Finalement, on calcule une AMDF corrigée AMDFc pour le candidat $T0_{t,i}$ par la formule (5):

$$\text{AMDFc}(T0_{t,i}) = \frac{\min\limits_{T1_{t,i} \leq T \leq T2_{t,i}} \text{AMDF}'(T, L_f)}{L_f - T2_{t,i}} \quad (5)$$

La plausibilité $Pb_{t,i}$ du candidat $T0_{t,i}$ sera alors égale à :

$$Pb_{t,i} = log\left[\frac{\max\limits_{T0_{min} \leq T0 \leq T0_{max}} \text{AMDF}(T0)}{\text{AMDFc}(T0_{t,i})} \right] \quad (6)$$

4. Stratégie

La détection du fondamental peut être considérée comme un processus de reconnaissance cherchant à minimiser la somme des erreurs de détection sur les segments voisés de la parole. Dans cette optique, nous avons adopté comme stratégie de décision une programmation dynamique globale qui se charge, dans le treillis de candidats proposés par la détection précédente, de trouver la courbe de fondamental la plus plausible, en fonction de certaines contraintes de continuité de la courbe mélodique. La distance locale est égale à la plausibilité du candidat considéré. Une plausibilité de transition $Pt(F1,F2)$ entre deux candidats est calculée suivant la formule (7):

$$Pt(F1,F2) = Pta.\frac{|F1 - F2|}{F1} \text{ avec } Pta = -2.5 \quad (7)$$

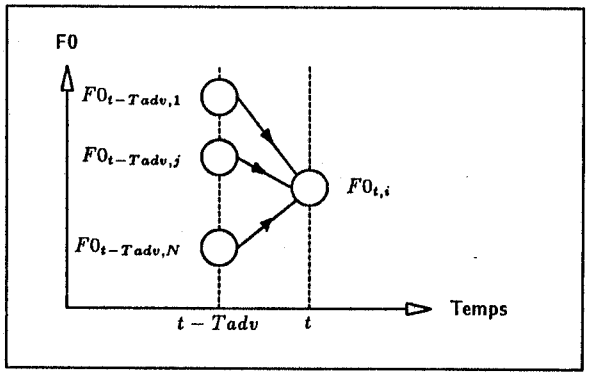


Fig. 2 Programmation dynamique globale.

La courbe mélodique proposée dans l'intervalle de voisement $[t1, t2]$ sera obtenue en tracant le meilleur chemin obtenu par les équations de la programmation dynamique globale données ci-dessous:

$$Cumul_{t,i} = Pb_{t,i} + \begin{cases} 0 & \text{si } t = t1 \\ \min\limits_{j} \left(\dfrac{Cumul_{t-Tadv,j}+}{Pt(F0_{t-Tadv,j}, F0_{t,i})} \right) & \text{pour } t1 < t \leq t2 \end{cases} \quad (8.a)$$

Avec la condition de continuite de voisement:

$$j / Pt|F0_{t-Tadv,j}, F0_{t,i}| > Pta \quad (8.b)$$

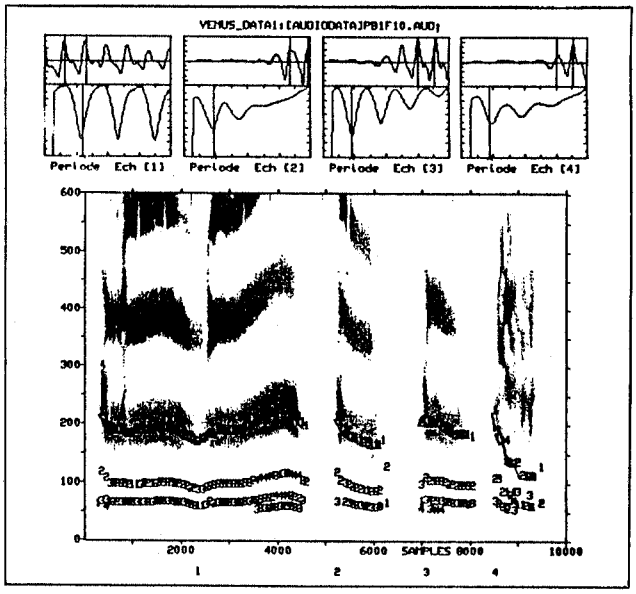


Fig. 3 Exemple de détection et de résultat donné par la programmation dynamique globale sur la phrase : "mes gants sont usés" échantillonée à 10 KHz. Les chiffres de 1 à 5 superposés au spectrogramme à bande étroite (70Hz) représentent les candidats proposés par la détection. La courbe est l'hypothèse finale retenue par la programmation dynamique globale. 4 coupes du signal ont été réalisées et sont numérotées de 1 á 4 en bas. L'AMDF ainsi que la fenêtre du signal sont présentées en haut. Les périodes sont marquées par des traits verticaux.

Un premier découpage du signal en parties voisées est effectué grâce aux indices EnrBF et Friction ainsi que certains seuils de plausibilité. Une première hypothèse est émise sur la partie voisée précédant la fenêtre non-voisée en remontant la programmation dynamique réalisée. On autorise la possibilité d'erreurs de voisement en limitant l'amplitude des transitions. L'intervalle de recherche d'une transition est limitée par un seuil de plausibilité de transition égal à $Ptmin = Pta$. L'algorithme de programmation dynamique autorise donc un choix multiple de points d'arrêts qui corresponderaient à des candidats abherrants proposés par la détection.

Cet algorithme segmente donc la partie voisée en plusieurs segments voisés. Les segments retenus seront ceux qui assureront une somme des distances cumulées aux extrémités minimale et un nombre minimal de segments.

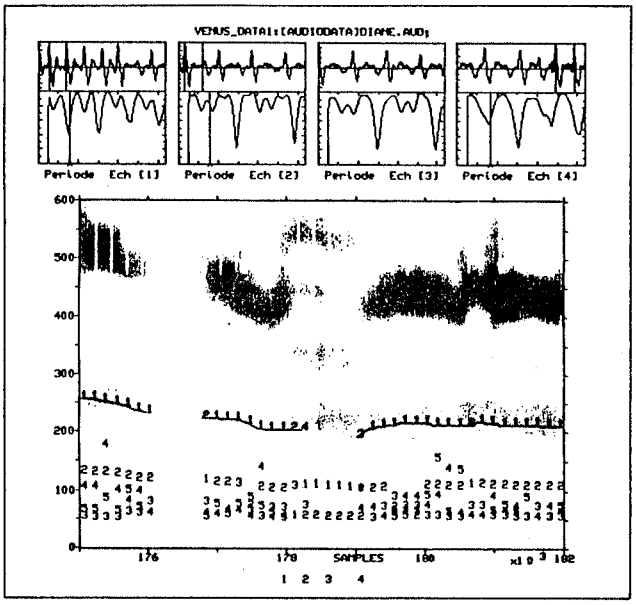


Fig. 4 Exemple de détection pour locuteur diplophonique extrait de DIANE locutrice anglaise :"..drivers are one of the..".

L'hypothèse finale est émise après une détection de pause. On reconsidère alors les premières hypothèses segmentales afin d'assurer la continuité de la courbe mélodique de part et d'autre de segments non-voisés en s'appuyant sur les segments les plus longs. Cette stratégie permet de corriger les erreurs dues à des segments voisés brefs où les effets microprosodiques et l'imprécision de la détection de voisement conduisent à une première hypothèse erronée.

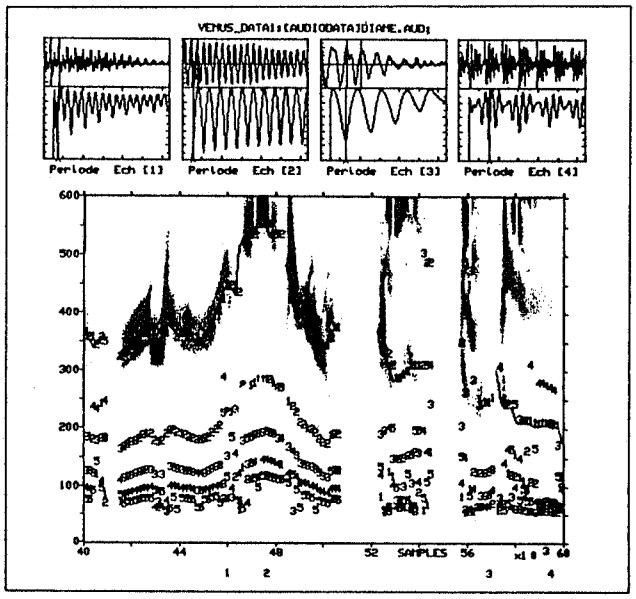


Fig. 5 Dynamique réalisée par la locutrice VICTOIRE dans la phrase: "...malade, où j'irai chercher.."

Nous n'avons retenu aucune contrainte de continuité entre segments voisés espacés de plus de 30ms, la courbe mélodique pouvant évoluer du simple au double de part et d'autre d'un segment non voisé comme illustré par le figure Fig 5.

Un exemple de treillis de décision est donné Fig. 3. La solution donnée par la stratégie globale est représentée par la courbe reliant les hypothèses retenues. 4 coupes du signal sont réalisées reproduisant en haut de la figure la fenêtre d'analyse et son AMDF ainsi que la période retenue dans chaque cas.

5. Résultats

Une série de tests préliminaires entrepris sur de la parole non bruitée en contexte de lecture n'a pas produit d'erreur globale supérieure à 0.5 % (cf.5 : locutrice PB1) . Nous avons donc choisi d'évaluer la robustesse de l'algorithme sur de la parole bruitée en conversation spontanée de sujets non avertis. Chaque fichier de parole dépasse la minute. Les résultats obtenus sont présentés Table 5 pour un signal à large bande et à bande étroite. Les erreurs grossières sur F0 avoisinent les 1 % alors que l'erreur globale est d'environ 3 %. La locutrice DIANE présente un taux d'erreurs grossières sur F0 plus important de part une tendance diplophonique marquée (cf.Fig. 4) représentant un cas trés difficile pour une détection du fondamental [1]. L'accroissement d'erreurs entre les deux bandes s'explique principalement par la dégénérescence de la qualité de la détection de voisement. On peut remarquer en outre que les erreurs de voisement (NV⇒V,V⇒NV) sont équilibrées. La plupart des erreurs de dévoisement (V⇒NV) sont dues à l'imprécision de la détection de fin de mot: Beaucoup de voyelles en fin de phrase sont écourtés et certains schwas en fin de mot sont étiquetés non-voisés à cause de leur énergie trop

faible (ce phénomène est trés important pour la locutrice VICTOIRE).

LOCUTEURS			LARGE BANDE (0-900 Hz)			
Nom (Langue)	Fe KHz	F0moy Hz	NV⇒V (%)	V⇒NV (%)	F0 (%)	Total (%)
BERNARD (F)	16	111 ± 27	0.86	1.53	0.78	3.17
GINETTE (F)	14	212 ± 42	1.53	1.22	0.45	3.20
VICTOIRE (F)	14	217 ± 47	0.76	4.09	1.02	5.87
HENRY (A)	16	174 ± 32	1.67	0.36	1.02	3.05
GORDY (A)	16	157 ± 27	1.57	1.02	0.90	3.48
DIANE (A)	16	243 ± 52	2.99	0.24	3.98	7.21
PB1 (A)	8	185 ± 23	-	0.02	-	0.02

Tableau (a) Détection du fondamental : Résultats large bande.

LOCUTEURS			BANDE ETROITE (300-900 Hz)			
Nom (Langue)	Fe KHz	F0moy Hz	NV⇒V (%)	V⇒NV (%)	F0 (%)	Total (%)
BERNARD (F)	16	111 ± 27	0.87	1.57	0.79	3.33
GINETTE (F)	14	212 ± 42	1.07	1.81	1.17	4.05
VICTOIRE (F)	14	217 ± 47	1.15	4.44	1.75	7.34
HENRY (A)	16	174 ± 32	2.76	0.22	1.09	4.06
GORDY (A)	16	157 ± 27	2.71	1.00	0.17	3.89
DIANE (A)	16	243 ± 52	2.30	2.09	5.48	9.87
PB1 (A)	8	185 ± 23	-	0.02	-	0.02

Tableau (b) Détection du fondamental : Résultats bande étroite.

6. Conclusions

Cet algorithme basé sur une double programmation dynamique locale et globale permet d'assurer la cohérence des résultats sur l'ensemble des parties voisées du signal. Certaines améliorations restent possibles notamment en ce qui concerne la décision de voisement et certaines anomalies du signal (ex: diplophonie) où il est nécessaire de conserver le signal original sans détruire sa structure formantique.

Remerciements

Je remercie Mr. Matt Lennig pour avoir conseillé et stimulé ce travail.

Références

1. Rabiner L.R. et al., "A comparative study of several pitch detection algorithms", *IEEE Trans. on Acous., Speech, and Sig. Proc.*, 24, pp.399-418, 1976.
2. Oh K.A., Un C.K., "A performance comparaison of pitch extraction algorithms for noisy speech", *IEEE Int. Conf. on Acous, Speech, and Sig. Proc.*, 2, pp.18B.4.1-18b.4.4, 1984.
3. Abry C., Boe L.J., Zurcher J.F., "La détection de voisement par les propriétés physiques résultant de l'excitation périodique du conduit vocal: comparaison statistique de trois procédés", *6èmes Journées d'Etudes sur la Parole*, pp.228-245, 1975.
4. Dubnovski J.J., Schafer R.W., Rabiner L.R., "Real-time digital hardware pitch detector", *IEEE Trans. on Acous, Speech, and Sig. Proc.*, 24, pp.2-8, 1976.
5. Ross M.J., Shaffer H.L., Cohen A., Feudberg R., Manley H.J., "Average magnitude difference function pitch extractor", *IEEE Trans. on Acous, Speech, and Sig. Proc.*, 22, pp.353-362.
6. Markel J.D., "The SIFT algorithm for fundamental frequency estimation", *IEEE Trans. on Audio. Electroacoust.*, 20, pp.367-377, 1972.
7. Schafer R.W., Rabiner L.R., "System for automatic formant analysis of voiced speech", *J. Acoust. Soc. Amer.*, 47, pp.634-648, 1970.
8. Tuffelli D., "A pitch detection algorithm with hypothesis and test strategy by means of fast surface AMDF", *IEEE Int. Conf. on Acous, Speech, and Sig. Proc.*, 2, pp.18B.3.1-18B.3.4, 1984.
9. Boulogne M., *Détection de voisement et de la fréquence fondamentale d'un signal de parole*, Thèse de Docteur-Ingénieur, Grenoble, 1979.
10. Rabiner L.R. et al., "Applications of a nonlinear smoothing algorithm to speech processing", *IEEE Trans. on Acous., Speech, and Sig. Proc.*, 23, pp.552-557, 1975.

Eléments de caractérisation des formes spectrales

Hélène CERF-DANON, Anne-marie DEROUAULT, Marc EL BEZE, Bernard MERIALDO, Serge SOUDOPLATOFF

Centre Scientifique IBM France
36 avenue Raymond Poincaré, 75116 Paris FRANCE
Tel: (1) 45 05 14 00

Abstract

This paper summarizes the results of our experiments on a distance measure for speech signals. This measure aims at characterizing the shape of the signal so as to take account of such physical phenomenons as the dilatation, retraction or frequency shift in a speech spectrum.

After a short overview of the context in which this work was done, the first part describes the measure itself. In the second part, we explain the test we chose to evaluate the performance and then give the results compared with those for some usual measures.

Introduction

Ce travail a été effectué dans le cadre d'un projet pour la dictée automatique du français avec un très grand vocabulaire . Dans notre système, une première étape de quantification vectorielle consiste à extraire d'un corpus d'apprentissage 200 vecteurs représentatifs du signal de parole, afin de ramener ensuite l'acoustique observée à une suite de nombres compris entre 1 et 200, chaque vecteur étant affecté à la classe du vecteur représentatif dont il est le plus 'proche'. Le problème est donc d'évaluer quantitativement cette proximité entre deux fenêtres de parole.

Plusieurs types de paramètres peuvent être calculés pour définir les vecteurs acoustiques (coefficients d'autocorrélation, coefficients de prédiction linéaire, spectre...). Pour évaluer la proximité entre deux tels vecteurs, de nombreuses mesures ont été proposées. La mesure euclidienne, est utilisée très fréquemment à cause de sa facilité d'emploi et aussi parce qu'elle fournit des résultats relativement corrects.

La mesure de proximité sur laquelle nous avons travaillé a pour but d'exploiter la ressemblance visuelle entre deux spectres d'un même phonème. Le concept de forme que nous allons maintenant exposer a déjà été formalisé et utilisé avec succès pour des spectres physiques, en chromatographie et en géologie ([3], [2]).

Définition d'une mesure de comparaison des formes.

Similitude de deux spectres théoriques.

Nous travaillons maintenant sur des fonctions d'une variable réelle, continues, strictement positives, intégrables et d'intégrale finie. Pour une telle fonction, la notion courante de forme visuelle est caractérisée par un changement d'échelle sur les ordonnées, et une transformation affine sur les abscisses. Autrement dit, deux spectres u et v sont de même forme si et seulement si:

(1) $u = k \times v o \phi$

avec k réel positif, et ϕ affine croissante.
Cette similitude se traduit simplement à l'aide des fonctions de répartition de u et v: Pour une fonction u vérifiant les hypothèses ci-dessus, on peut définir sa fonction de répartition U par la formule:

$$U(x) = \frac{\int_{-\infty}^{x} u(t)\,dt}{\int_{-\infty}^{+\infty} u(t)\,dt}$$

D'après les hypothèses sur u, U est à valeurs dans [0 , 1], strictement croissante et continue donc inversible. Sa réciproque U^{-1} est définie de [0 , 1] dans l'ensemble des réels.

Soit u et v deux fonctions vérifiant les hypothèses ci-dessus, elles sont dites semblables au sens de la similitude des formes si et seulement si il existe (a,b) deux constantes (a > 0) telles que, quelque soit x,

(2) $V(x) = U(ax + b)$

En effet, ceci entraîne, en dérivant les deux membres:

$$\frac{1}{\int_{-\infty}^{+\infty} v(t)\,dt} \times v(x) =$$

$$\frac{1}{\int_{-\infty}^{+\infty} u(t)\,dt} \times \frac{d}{dx}\int_{-\infty}^{ax+b} u(t)\,dt$$

soit:

$$\frac{1}{\int_{-\infty}^{+\infty} v(t)\,dt} \times v(x) = \frac{1}{\int_{-\infty}^{+\infty} u(t)\,dt} \times u\left(\frac{x-b}{a}\right)$$

Réciproquement, u et $v = k\,uo\phi$ ont même fonction de répartition, et (2) est vérifiée.

Ceci s'écrit encore, grâce à la bijectivité de U^{-1}

$$U^{-1}(V(x)) = ax + b.$$

D'où le critère de similitude suivant : u et v sont semblables au sens de la similitude des formes si et seulement si il existe une fonction affine ϕ strictement croissante telle que

(3) $\quad U^{-1}oV = \phi.$

Propriétés: la similitude ainsi définie est une relation d'équivalence. Ceci se démontre aisément grâce au fait que l'ensemble de fonctions affines croissantes est un groupe pour la loi de composition.

Il reste maintenant à définir une mesure de la distance entre deux spectres grâce à ce critère de similitude. Remarquons qu'il a une interprétation graphique agréable: il s'écrit simplement " le graphe (U^{-1}, V^{-1}) est une droite". Cette remarque justifie l'idée suivante: on va mesurer la similitude au sens des formes par l'écart de ce graphe à la droite 'la plus proche'. Pour préciser cet écart nous allons maintenant nous placer dans le cadre qui nous intéresse, c'est à dire le cas discret.

Soit u_1^n une suite réelle strictement positive, on définit

$$U(i) = \frac{\sum_{j=1}^{i} u_j}{\sum_{j=1}^{n} u_j}$$

pour i = 1 à n.

On peut alors inverser la fonction obtenue par interpolation linéaire ce qui permet de définir n valeurs, $U_k = U^{-1}(\frac{k}{n})$ pour k = 1 à n.

Pour comparer deux suites u_1^n et v_1^n, on va définir la distance de similitude entre u et v, E(u,v) comme étant la somme des distances dans le plan euclidien des points (U_i, V_i) à leur projection sur la droite d'équation y = ax + b qui minimise cette somme.

Application aux spectres de parole.

Données vocales.

Le signal est digitalisé à 10 khz. La transformée de Fourier est calculée sur chaque fenêtre de 12,8 ms, après lissage en énergie, et les paramètres acoustiques retenus sont les valeurs de énergies dans 20 bandes de fréquence, suivant une échelle MEL. Ces spectres sont inversés par interpolation linéaire, qui fournit les valeurs de l'inverse pour 1/20, 2/20... 1.
Remarque:
L'écart de forme ne sera pas très sensible aux petits décalages de formants (translation en abscisse). Par contre, la valeur de b^2 apportant sa contribution à l'écart quadratique calculé, tout grand décalage sera pénalisé.

Quantification vectorielle.

Un ensemble de 30000 vecteurs correspondant à quelques minutes de parole, est classifié automatiquement avec l'algorithme de K-Means. Les centres de classe seront pris comme les 200 vecteurs représentatifs. Ce corpus d'apprentissage a été préalablement segmenté en phonèmes. L'algorithme de K-Means avec une "distance" d donnée, se déroule de la façon suivante:

1. Les centres sont initialisés à des vecteurs pris dans la portion de signal correspondant à chaque début, milieu et fin de phonèmes.

2. Chaque vecteur est alors affecté à la classe du centre le plus proche, au sens de la distance d.

3. Les "centres de gravité" au sens de d des nouvelles classes sont calculés: On prend comme nouveau centre de chaque classe un point qui minimise la somme des distances aux points de la classe. Dans le cas de la distance euclidienne, il s'agit du centre de gravité qui est donc calculé de manière explicite, et n'est a priori pas un point "réel" de la classe. Pour notre distance de forme, on ne dispose pas de formule explicite pour calculer un tel point, nous prenons donc le point "réel" de la classe qui minimise cette somme de distances.

4. On itère le procédé à partir de 2.

L'algorithme s'arrête lorsque le pourcentage de migration (nombre de points qui changent de classe après une itération) est inférieur à un seuil. Expérimentalement, il converge plus rapidement pour la similitude de forme (7-8 itérations), que pour la distance euclidienne.

Etiquetage du signal et évaluation.

La suite de vecteurs acoustiques est remplacée, pour la suite du décodage, par la suite des numéros (ou "labels") des centres de classes les plus proches.

Pour comparer différentes distances spectrales, il serait très utile d'avoir un critère qui ne suppose pas de recalculer tous les paramètres du décodeur acoustique (actuellement notre système est basé sur des modèles de Markov cachés, cf [1]). C'est pourquoi nous proposons comme critère de qualité d'un étiquetage pour une distance donnée, la **quantité d'information mutuelle** entre la suite de phonèmes prononcée et la suite de labels obtenue.

Rappelons que l'entropie d'une variable aléatoire X est égale à

$$-\sum_{i=1}^{n} p(a_i)\log(p(a_i))$$

Pour deux variables aléatoires X et Y, la quantité d'information mutuelle entre deux valeurs respectives a_k et b_j est définie par:

$$I_{X;Y}(a_k, b_j) = Log\frac{P_{X/Y}(a_k / b_j)}{P_X(a_k)}$$

Cette quantité est symétrique en X et Y, et représente l'information fournie par la survenue de a_k au sujet de b_j. La quantité d'information mutuelle entre les deux variables X et Y, est définie comme la moyenne de la variable aléatoire $I_{X;Y}$:

$$\sum_{k=1}^{K}\sum_{j=1}^{J} P_{XY}(a_k, b_j) \times Log\frac{P_{X/Y}(a_k / b_j)}{P_X(a_k)}$$

Cette quantité est toujours positive et majorée par l'entropie de chacune des variables X et Y.

Résultats.

Nous avons effectué un K-Means, et un étiquetage du corpus, pour trois distances:

1. la distance euclidienne entre spectres MEL ("EU-MEL"),
2. La distance euclidienne entre coefficients PARCOR ("PARCOR")
3. La distance de forme ("FORME").

Quantité d'information.

Nous avons calculé la quantité d'information mutuelle entre chacune des suites de labels obtenues, et la suite de phonèmes prononcée. Les résultats sont présentés dans le tableau suivant. D désigne la distance, H_l l'entropie des labels (les probabilités sont simplement les fréquences d'apparition), H_p l'entropie des phonèmes, $Q_{l,p}$ la quantité d'information mutuelle entre les labels et les phonèmes.

D	H_l	H_p	$Q_{l,p}$
EU-MEL	6.04	3.43	2.20
PARCOR	6.33	3.43	2.10
FORME	6.86	3.43	1.91

D'un point de vue global, (en moyenne sur tous les phonèmes), l'étiquetage avec la distance de forme fournit donc moins d'information sur la suite de phonèmes que la distance euclidienne. Nous allons maintenant analyser ces suites de labels phonème par phonème.

Matrice de confusion.

Grâce à l'initialisation particulière du K-Mean, (les classes de départ correspondent à des segments de début, milieu et fin de chaque phonème), chaque label est en rapport avec un certain phonème. De cette correspondance on déduit une matrice 36 × 36, où l'élément (i,j) est la somme des fréquences d'apparition, dans l'étiquetage du phonème i, des labels initialement en correspondance avec le phonème j.

Le tableau suivant montrent quelques extraits intéressants de cette matrice. La première colonne indique le phonème prononcé i, la deuxième colonne la famille phonétique de labels dont la fréquence d'apparition est la plus grande, et la seconde plus grande, puis les deux dernières colonnes indiquent ces fréquences (en pourcentage) successivement pour l'étiquetage issu de la distance euclidienne, et de la similitude de forme.

Phonème	Label	Eu-mel	Forme
S	S	55	73
	Z	15	14
CH	CH	64	70
	JJ	9	5
IN	IN	27	40
	UN	33	23
W	W	23	34
	AU	36	20
UU	UU	41	43
	U	30	19
GN	GN	60	80
	N	20	20

Les "confusions" entre phonèmes pour les sons CH, S, IN, UU (nuage), W (loi), GN (agneau), sont donc nettement améliorées par la similitude de forme. Les labels obtenus dans ces cas montrent d'ailleurs plus de répétition et de structure. Ce phénomène se produit surtout pour des semi-voyelles (trois derniers cas), qui sont en général difficiles à reconnaître. La similitude de forme est donc intéressante pour les sons qui présentent une migration de formants importante, avec éventuellement une grande largeur de bande pour les deux premiers formants, (ce qui est également vrai pour la voyelle nasale IN). Ceci est à rapprocher du fait que les utilisations de cette notion faites jusqu'à maintenant en physique concernaient des décompositions de spectres en pics ([3], [2]).

Conclusion

Nous avons cherché une mesure de similitude entre spectres de parole qui ne dépende pas tellement des valeurs brutes des paramètres acoustiques, mais qui vise plutôt à caractériser des informations invariantes par des transformations naturelles sur le signal telles que changement d'échelle ou affinité. La similitude introduite formalise la notion courante de forme visuelle. A l'aide d'une mesure d'information mutuelle, nous avons comparé ses performances avec celles de la distance euclidienne.

D'une part, il reste à étudier si la notion de forme choisie ici serait mieux adaptée à d'autres paramètres acoustiques. D'autre part, cette équivalence de forme peut être généralisée en remplaçant, dans la relation (3), l'ensemble des fonctions affines croissantes décrit par ϕ, par un autre groupe de fonctions continûment dérivables et croissantes.

Bibliographie

[1] H. Cerf, A.M. Derouault, M. Elbeze, B. Mérialdo, S. Soudoplatoff: Reconnaissance de la parole par des modèles markoviens: application aux grands vocabulaires. 15èmes Journées d'Etude sur la Parole, 1986, Aix en Provence.

[2] D. Massiot: "Shape comparison of physical spectra: application to Mössbauer spectra of silicate glasses", Journal of Non-Cristalline solids, 69 (1985), 371-380, North Holland, Amsterdam.

[3] H. Rix and J.P. Malenge: (1980), "Detecting small variations in shape." IEEE Transactions on Systems, Man and Cybernetics, vol SMC10, 2, 90-96.

EFFETS DE LA DUREE D'EMISSION VOCALE ET DU CONTENU PHONEMIQUE SUR LE SPECTRE MOYEN A LONG TERME

B. Harmegnies [1]

Laboratoire de Phonétique - Université de Mons - Hainaut [2]

ABSTRACT

400 - Channels constant bandwidth (12.5 Hz) Long Term Spectra (LTS) delivered by a BK 2033 analyser have been drawn from French utterances produced by 5 speakers. The cross-correlation coefficient was used to investigate LTS variability under changes in phonemic contexts and in speech samples durations. For all the investigated durations (10, 20, 30, 45 and 60 sec.), the LTS exhibited more variability for inter- than for intra-speakers comparisons. This differences nevertheless decreased with increasing duration. For intra- as well as for inter-text comparisons, the correlations increased with increasing speech sample duration.

INTRODUCTION

Ainsi que l'ont souligné Harmegnies et Landercy [1], la notion de variabilité intra-sujet du spectre moyen à long terme (SMLT) est souvent définie de manière imprécise dans la littérature. En particulier, certains auteurs [2-4] adoptent des plans expérimentaux ne permettant pas de dissocier la part des variations dues au sujet de celle des variations dues à la variabilité du contenu phonémique des productions. Sans doute faut-il en chercher la raison dans l'idée, généralement admise, que le SMLT est indépendant du contenu phonémique. Nolan [5], dans sa définition du SMLT reflète cette conception, mais il insiste toutefois sur la nécessité d'utiliser des échantillons vocaux suffisamment longs[3] pour annihiler l'effet des segments linguistiques individuels. Or, cette notion de "durée suffisante" n'a jamais fait l'objet d'études systématiques. Des durées de productions vocales aussi diverses que 10 secondes [6], 12 secondes [7], 27 secondes [8], 32 secondes [2-4], 45 secondes [9] sont ainsi utilisées sans guère de justification. Tout au plus, les auteurs se contentent-ils, au vu de résultats estimés probants, de déduire que la durée retenue était suffisante (par exemple, FURUI et al. [10], cités par NOLAN [5]).

Dans cet article, nous tenterons de jeter les bases d'une étude systématique des effets combinés de la distribution phonémique et de la durée d'émission vocale sur le SMLT. Nous étudierons, dans ce but, les variations de la similarité spectrale lors de comparaisons intra-locuteur impliquant des SMLT provenant de productions de durées variables et de contenus phonémiques soit constants, soit variables.

EXPERIMENTATION

Recueil des données

Les sujets étaient cinq adultes francophones de sexe masculin, âgés de 19 à 21 ans et exempts de toute pathologie vocale. Chacun fut invité, d'une part, à lire cinq fois un texte français équilibré d'une durée approximative d'une minute trente secondes ("texte court") et, d'autre part, à lire une fois un passage d'un roman français d'une durée approximative de sept minutes trente secondes ("texte long"). Les productions furent enregistrées selon la procédure utilisée lors de nos travaux antérieurs [1, 11, 12].

Constitution des échantillons de parole

A partir des enregistrements recueillis, 50 échantillons de parole furent extraits des productions vocales de chaque sujet. D'une part, de chacune des 5 productions du texte court furent extraits 5 échantillons correspondant respectivement à ses 10, 20, 30, 45 et 60 premières secondes. D'autre part, le texte long fut scindé en cinq extraits de durée égale; de chacun de ces derniers, furent également extraits 5 échantillons correspondant respectivement à ses 10, 20, 30, 45 et 60 premières secondes. Pour chaque durée d'émission retenue (10, 20, 30, 45 et 60 secondes), chaque sujet a donc fourni, d'une part, 5 échantillons provenant de 5 productions d'un corpus invariant (texte court) et, d'autre part, 5 échantillons provenant de la production de 5 corpus différents (extraits du texte long).

Traitement acoustique

Les analyses acoustiques furent réalisées lors d'une phase ultérieure au moyen de l'analyseur spectral BRÜEL KJAER 2033. Celui-ci, échantillonnant le signal d'entrée à 12.8 kHz, produisit des spectres définis sur 400 canaux dans la bande 0-5 kHz avec une résolution

constante de 12.5 Hz. Ces spectres instantanés furent appliqués à l'entrée de l'algorithme de calcul de moyenne à pondération linéaire qui fournit, pour chaque échantillon de parole retenu, un SMLT. Les 250 (5 sujets x 5 durées d'émission x 2 textes x 5 productions) SMLT obtenus furent transmis à un micro-ordinateur Apple II via une interface GPIB en vue du traitement statistique.

Traitement statistique

Afin de rendre les résultats de la présente recherche comparables avec nos publications antérieures [11, 12], nous avons recouru au coefficient de corrélation interspectrale pour procéder à la mesure de la similarité des SMLT. Un coefficient de corrélation a ainsi été produit pour chaque comparaison d'un SMLT à un autre.

Pour chaque locuteur et chaque durée d'émission vocale retenue, il fut procédé à 20 comparaisons interspectrales intra-sujet. Celles-ci se répartissent en 10 comparaisons intra-texte (toutes les comparaisons possibles de 2 SMLT provenant de productions différentes du texte court) et 10 comparaisons inter-texte (toutes les comparaisons possibles de 2 SMLT provenant de productions de deux extraits différents du texte long). Au total, 500 valeurs de corrélation intra-sujet représentant chacune une comparaison interspectrale ont donc été calculées lors de cette étude (10 comparaisons possibles x 2 types de corpus x 5 sujets x 5 durées d'émission).

RESULTATS

L'ensemble des valeurs calculées ne pouvant être reproduit ici, nous en fournissons des résumés statistiques aux tableaux 1 (comparaisons intra-texte) et 2 (comparaisons inter-texte). Dans chacun de ces tableaux figure, pour chaque sujet et chaque durée d'émission, la moyenne et l'écart type des 10 coefficients de corrélation issus de toutes les comparaisons possibles des 5 SMLT concernés entr'eux.

	10 s	20 s	30 s	45 s	60 s
S_1	.904	.942	.950	.963	.964
	.031	.012	.009	.008	.011
S_2	.894	.921	.952	.959	.964
	.019	.020	.007	.007	.008
S_3	.895	.958	.968	.967	.982
	.045	.008	.004	.005	.003
S_4	.904	.927	.948	.956	.967
	.019	.008	.009	.007	.006
S_5	.876	.913	.940	.948	.959
	.013	.016	.012	.010	.004
$S_{1\,5}$	.895	.932	.952	.958	.967
	.030	.021	.013	.010	.011

Tableau 1 : Moyennes (soulignées) et écarts-types par sujet des coefficients de corrélation interspectrale obtenus lors de comparaisons intra-texte, en fonction de la durée d'émission vocale.

Ainsi que le montre le tableau 1, la corrélation moyenne intra-texte augmente lorsque la durée d'émission augmente; parallèlement la variabilité de la corrélation tend à diminuer (baisse de l'écart-type). De plus, l'augmentation de la corrélation moyenne est, dans l'ensemble, régulière : toute moyenne générale du tableau 1 est plus grande que celle située à sa gauche dans le tableau et plus petite que celle située à sa droite. En outre, ces constatations (accroissement de la corrélation et régularité de l'accroissement) restent d'application à l'examen de chaque sujet considéré isolément[4].

Enfin, si, en général, l'accroissement des corrélations est assez important aux passages de 10 s à 20 s et de 20 s à 30 s (respectivement .037 et .020 en moyenne), il est en revanche plutôt faible aux passages de 30 s à 45 s et de 45 s à 60 s (respectivement .016 et .007, en moyenne).

	10 s	20 s	30 s	45 s	60 s
S_1	.862	.896	.923	.943	.946
	.038	.024	.015	.013	.012
S_2	.771	.839	.918	.933	.937
	.111	.055	.017	.013	.020
S_3	.856	.928	.940	.953	.959
	.049	.020	.013	.014	.006
S_4	.854	.896	.917	.934	.951
	.035	.022	.026	.011	.007
S_5	.826	.869	.913	.925	.934
	.032	.034	.009	.020	.020
$S_{1\,5}$	.834	.886	.922	.938	.945
	.069	.045	.019	.017	.017

Tableau 2 : Moyennes (soulignées) et écarts-types par sujet des coefficients de corrélation interspectrale obtenus lors de comparaisons inter-texte, en fonction de la durée d'émission vocale.

Les constatations opérées à la faveur des comparaisons intra-texte sont en tous points applicables aux comparaisons inter-texte (cfr. tableau 2). Il faut cependant remarquer que les coefficients de corrélation inter-texte sont systématiquement plus faibles que les coefficients de corrélation intra-texte. Cette tendance d'ensemble, déjà remarquée lors de nos études antérieures [12], est si marquée qu'on peut même l'observer dans le chef de chacun des locuteurs, quelle que soit la durée d'émission : on constatera ainsi aisément que toute corrélation moyenne du tableau 1 est plus grande que la corrélation moyenne correspondant, dans le tableau 2, aux mêmes sujet et durée d'émission.

DISCUSSION

Comme le montre la figure 1, qui résume les données de l'expérience, il est clair que les deux sources de variations à l'étude - contenu

phonémique et durée d'émission - influent sur le spectre moyen à long terme. Les recherches antérieures en matière de SMLT n'avaient guère permis la mise en lumière de ces phénomènes.

Au vu de la figure 1, il semble assez évident que la durée d'émission 30 secondes joue un rôle de pivot. On constate aisément, en effet, ainsi que nous l'avons souligné au chapitre précédent, que l'accroissement de corrélation dû à l'allongement du corpus est surtout sensible en-dessous de cette durée d'émission. Si la présente recherche ne suffit pas à mettre en évidence une durée d'émission optimale pour le SMLT, elle semble en tous cas indiquer que 30 secondes constituent un bon compromis entre longueur de l'échantillon vocal et représentativité du SMLT.

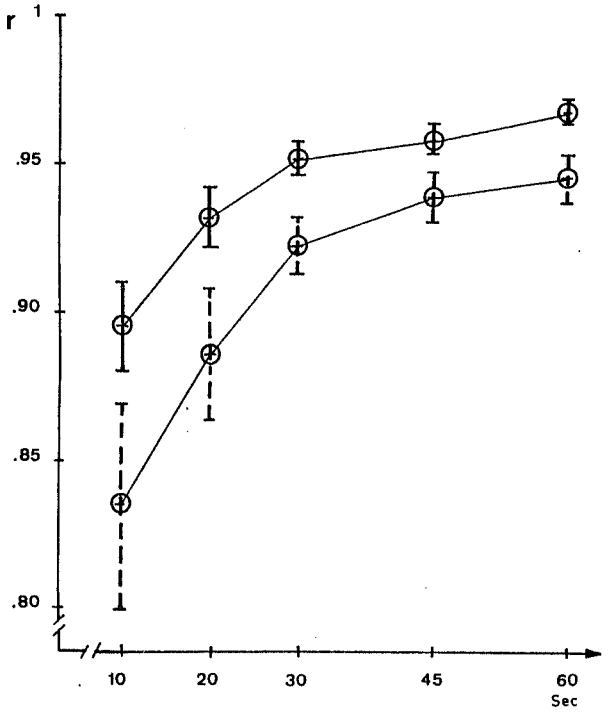


Figure 1 : Moyennes (cerclées) et écarts-types (longueur des segments) des coefficients de corrélation interspectrales intra-sujet (r) obtenus, tous sujets confondus, lors de comparaisons inter- (lignes brisées) et intra- (lignes pleines) texte, en fonction de la durée d'émission vocale.

En ce qui concerne le contenu phonémique, il est manifeste que sa variation entraine une chute de la similarité spectrale intra-sujet, et ce, quelle que soit la durée d'émission retenue parmi celles que nous avons investiguées. Il semble donc que, même pour une durée d'une minute d'émission vocale, le SMLT n'est pas strictement indépendant du texte.

CONCLUSION

Nous formulerons, au terme de cette brève étude, deux remarques à propos de la validité de nos constatations. D'une part, le nombre relativement réduit de sujets et d'échantillons vocaux peut certes paraître amoindrir la portée de nos conclusions; le fait que les constatations opérées s'appuient toutes sur des variations observables sur chague sujet et dans chague condition nous conduit néanmoins à une confiance renforcée dans le caractère systématique des phénomènes mis au jour. D'autre part, si les mesures effectuées ne peuvent s'appliquer qu'au type de dispositif que nous avons utilisé, elles nous semblent cependant avoir le mérite d'établir de manière patente l'existence de phénomènes habituellement ignorés.

Finalement, cette étude apporte une confirmation de plus à la nécessité de recherches méthodologiques à propos du SMLT qui, avant qu'en en saisît toutes les nuances, s'est vu prématurément accaparé par la recherche appliquée.

REFERENCES

[1] HARMEGNIES, B. et A. LANDERCY, "Statistical study of speaker-dependant variability of the long-term speech spectrum", Speech Communication, soumis, 03-86.

[2] BYRNE, D., "The speech spectrum-some aspects of its significance for hearing aid selection and evaluation", Br. J. Audiol, 11, 40-46, 1977.

[3] ZALEWSKI, J., W. MAJEWSKI et H. HOLLIEN, "Cross-correlation of long-term speech spectra as a speaker identification technique", Acustica, 34, 20-24, 1975.

[4] HOLLIEN, H. et W. MAJEWSKI, "Speaker identification by long-term spectra under normal and distorted speech conditions", J. Acoust. Soc. Am., 62, 4, 975-80, 1977.

[5] NOLAN, F., The phonetic bases of speaker recognition, Cambridge, Cambridge University Press, 1983.

[6] FURUI, S., F. ITAKURA, et S. SAITO, "Personal information in the long-time averaged speech spectrum", Rev. of the Electrical Communication Laboratories, 23, 9-10, 1133-1141, 1975.

[7] BUNGE, E., "Herkenning van sprekers door een computer", Philips techn., 37, 179-192, 1977, 2.

[8] FORMBY, C. et R.B. MONSEN, "Long-term average speech spectra for normal and hearing-impaired adolescents", J. Acoust. Soc. Am. 71 (1), Jan. 1982.

[9] FROKJAER-JENSEN, B., et S. PRYTZ, "Registration of Voice Quality", Brüel Kjaer Techn. Rev., 3, 3-17, 1976.

[10] FURUI, S., F. ITAKURA et S. SAITO, "Talker recognition by long time averaged speech spectrum", Electronics and Communications in Japan, 55-A, 10, 54-61, 1972.

[11] HARMEGNIES, B., et A. LANDERCY, "Language features in the long-term average spectrum", Rev. Phon. Appl., 73-5, 69-80, 1985.

[12] HARMEGNIES, B., "Traitements et utilisations des spectres vocaux moyens", Actes des 13èmes JEP, Bruxelles, GALF, 115-117, 1984.

NOTES

1. Cette recherche a été supervisée par le Professeur A. Landercy, à qui va toute notre chaleureuse gratitude.
2. Cette recherche est partiellement subsidiée par le Fonds National de la Recherche Scientifique belge, sous le numéro 2.4450.86F.
3. "[Long-term spectra] are overall spectra obtained by averaging the short term power spectrum over a sample of speech long enough for the effect of the spectral characteristics of individual linguistic segments no longer to be significant" (NOLAN, [5], p. 130).
4. Avec cependant une légère exception dans le chef du sujet 2, qui présente une diminution de .001 (!) au passage de 30'' à 45'' (.967 < .968).

ETUDE COMPARATIVE DE MESURES DE DISTORSION SPECTRALE

M.J. CARATY X. RODET

UNIVERSITE PIERRE ET MARIE CURIE - INSTITUT DE PROGRAMMATION
Laboratoire de Reconnaissance de Formes et de Parole
4, place Jussieu - 75230 PARIS CEDEX 05

ABSTRACT

The purpose of this study is to compare the performance of a new perceptually based measure, to classical spectral distortion measures performances.
Two usual distances have been selected for their known performances in various recognition systems : the log likelihood ratio measure (LLR) and the mel frequency cepstrum coefficients measure (MFCC).
The new one, spectral peaks adjustment measure (mesure par ajustement de pics spectraux : APS), is based on perceptual criterions relative to spectral maxima.
Several recognition experiments on a data base of vowel steady states (extracted from continuous speech) are considered for the evaluation of protocoles and performances.

I . INTRODUCTION

Déterminer une distance entre deux spectres instantanés est un problème fondamental dans l'analyse de la parole.
Si la distance spectrale est un opérateur priviligié pour les systèmes de reconnaissance automatique, elle est utile dans des applications aussi variées que la vérification du locuteur, le codage de la parole par compression d'états stables, l'aide à la segmentation.
La distance est fonction de la représentation paramétrique du signal acoustique. L'objet de la paramétrisation est de réduire les données du signal de parole, elle se doit donc de conserver l'"information utile" à l'analyse phonétique, voire même d'amplifier les aspects du signal permettant la mesure de "différences perceptives".
Ainsi motivées par la perception sonore, on note dans les paramétrisations récemment proposées l'intégration de connaissances sur la perception, par exemple l'utilisation d'échelle de fréquences Mel (3), Bark (12), de fonction de poids (10), (14).
Deux groupes de paramétrisations sont distingués dans notre étude :
- paramétrisation par transformée de Fourier (MFCC),
- paramétrisation par prédiction linéaire (LLR, APS).
Les distances résultantes sont soit euclidiennes (MFCC), soit spécifiquement liées à la modélisation (LLR, APS).

II . MESURES DE DISTANCE INTERSPECTRALE

2.1. Logarithme du Rapport de Vraisemblance (LLR)

Cette mesure, également connue sous le nom de distance d'Itakura, (9), ((3), (7), (8), (12)), compare les modélisations par prédiction linéaire de deux séquences d'échantillons en terme de logarithme du rapport des vraisemblances de deux observations.
Soit s(n) la séquence test à comparer à s'(n) la séquence de référence ;

$$a^T = (1, a_1, \ldots a_p) \text{ et } a'^T = (1, a'_1, \ldots, a'_p),$$

les coefficents de prédiction relatifs à ces séquences, et R la matrice d'autocorrélation de s(n).
La distance LLR est alors définie par :

$$d_{LLR}(s', s) = \text{Log} \left(\frac{a'^T R a'}{a^T R a} \right)$$

et peut être calculée plus rapidement sous la forme :

$$d_{LLR}(S', S) = \text{Log} \left(\frac{b'(0)\, r(0) + 2 \sum_{k=1}^{P} b'(k)\, r(k)}{\alpha} \right)$$

où $b'(k)$ est la séquence d'autocorrélation de a',

$$r(k) = R_{(k,k)},$$

$$\alpha = a^T R a .$$

2.2. Distance Euclidienne sur les coefficients cepstraux en échelle MEL (MFCC)

Le spectre de Fourier est calculé sur une fenêtre d'analyse. Ce spectre est filtré au moyen d'un banc de filtres régulièrement répartis sur une échelle MEL : par exemple un banc de 22 filtres triangulaires espacés de 100 Mels de 0 à 5000 Hz. Les coefficients cepstraux sont calculés sur la séquence $\{E(k)\}$, où E(k) est l'énergie dans le filtre k (3), ((2), (6), (11)) :

$$CC_j = \sum_{k=1}^{22} \log (Ek) \cos \left(j \left(k - \frac{1}{2} \right) \frac{\pi}{22} \right)$$

2.3. Mesure par ajustement de pics spectraux

La fenêtre d'analyse est paramétrisée sous forme de maxima locaux (P) du spectre LPC (S). Ces maxima (représentant en général les formants) sont caractérisés par leur fréquence centrale (f en Hz), largueur de bande à 3 dB (l en Hz) et amplitude (a en dB).
On note alors un spectre :

$$S = \{ P^i(f^i, l^i, a^i) , i \in N^* \}$$

Soient les deux spectres à comparer S_r et S_t, spectre référence et spectre test.
On mesure la déformation globale de S_r relativement à S_t, en établissant la meilleure correspondance des pics de S_r aux pics de S_t, au sens d'une déformation minimale.
Une distance "inter-pic" (1) $d(P^j{}_t, P^i{}_r)$ est introduite pour valuer ces déformations partielles de S_r, où interviennent des mesures de déformation liées aux écarts, en fréquence, largueur de bande et amplitude, avec des coefficients de pondération associés.
Des connaissances inhérentes aux domaines de fréquence, de largeur de bande et d'amplitude, sont quantifiées par classes de fréquences (1).
Les coefficients de pondération sont alors déterminés en faisant l'hypothèse d'égale contribution à la distance inter-pic, des variations égales aux seuils différentiels perceptifs (5), de fréquence, de largeur de bande ou d'amplitude (normalisation de la quantification).
On suppose que tout spectre référence S_r admet une "forme invariante" dans la classe qu'il représente, définie par la présence des n premiers pics de S_r, et à partir duquel seront pénalisées les apparitions ou les disparitions de pics.
La distance interspectrale finale est alors la moyenne d'une sélection de distances inter-pic, basée sur la "forme invariante" (1).

III . EXPERIMENTATIONS DE RECONNAISSANCE

3.1. Base de données

Le signal de parole provient de l'enregistrement d'un article lu continûment par un locuteur-homme dans un environnement peu bruyant. Sa durée est de 3'30, il est échantillonné à 10 KHz.
Un algorithme de segmentation automatique fondé sur des critères de variation d'énergie du signal et de durée d'états stables, permet d'extraire des candidats noyaux stationnaires (4).
Les candidats (860) sont étiquetés après plusieurs procédures d'écoute (11), (13). Seuls sont sélectionnés pour la base de données les noyaux vocaliques (524) d'étiquettes phonétiques :

/ɛ/ /e/ /ã/ /ɔ/ /õ/ /ɛ̃/ /a/ /oe/ /i/ /o/ /y/ /u/

980 voyelles (accentuées ou non) sont répertoriées dans le texte.
Les fenêtres d'écoute sont cadrées sur les fenêtres d'analyse.
La décision de reconnaissance est basée sur l'étiquetage résultant de la fenêtre d'écoute considérée.
Deux procédures d'étiquetage liées à la durée des fenêtres d'écoute ont été expérimentées.
Elles sont motivées par l'idée d'adéquation entre la durée de la fenêtre de travail (22ms), porteuse de l'information analysée et la durée de la fenêtre d'écoute, porteuse de l'information utilisée par l'auditeur pour le décodage acoustico-phonétique, conditionnant la décision de reconnaissance.
a - "Contexte court" (étiquetage C) : 110ms.
b - "Contexte long" (étiquetage L) : 440ms.
Si l'étiquetage C présente la meilleure adéquation, taille de fenêtre d'analyse / taille de fenêtre d'écoute, l'étiquetage L correspond à la transcription phonétique de l'enregistrement (*). C'est l'étiquetage traditionnellement considéré pour les applications de reconnaissance de type décodage acoustico-phonétique.
Soit (C^C_p) (resp. C^L_p) la classe des noyaux étiquetés phonétiquement par p en contexte d'écoute court C (resp. long L).
L'étiquetage de la base de données à l'issue des deux procédures d'écoute fait apparaître des modifications du cardinal des classes phonétiques (11), (13).

p	ɛ	e	ã	ɔ	õ	ɛ̃	a	oe	i	o	y	u
C^C_p	79	65	53	15	19	20	79	93	49	23	18	11
C^L_p	46	73	55	17	18	20	118	83	41	20	17	16

Les différences d'étiquetage enregistrées (environ 17 %) affectent plus particulièrement les classes /ɛ/, /ɔ/ et /a/. 46 % des éléments étiquetés /ɛ/ en contexte c, se retrouvent étiquetés /a/ en contexte L, ce "glissement" a des répercussions comme nous le verrons sur l'homogénéité de la classe /a/ (cf.3.2). Le cardinal de la classe /ɔ/ ne subit pas de modification notable, cependant seul la moitié des éléments admettant cette étiquette dans l'étiquetage C le conservent dans l'étiquetage L.

3.2. Contextes d'analyse

Les fenêtres d'analyse de 22 ms (échantillonnage : 10 KHz) sont prétraitées par préemphase (filtre P(z) = 1-15/16 z) et fenêtrage de Hamming.
La prédiction linéaire (autocorrélation) est d'ordre 16 (LLR-APS).
Les spectres FFT (MFCC) et LPC (APS) sont obtenus par transformation à 512 points.
Le nombre de coefficients MFCC fixé à 8 .
Pour déterminer les caractéristiques des maxima spectraux (APS), on choisit la méthode qui consiste à détecter les maxima locaux par un balayage du spectre LPC :
- la fréquence centrale est exprimée en Hz,
- la largeur de bande à 3dB en Hz, déterminée suivant la précédence :
  . largeur de bande réelle,
  . demi-largeur de bande,
  . interpolation parabolique passant le maximum et les points d'inflexion gauche et droite à l'enveloppe spectrale,
- l'amplitude en dB, avec le maximorum référencé à 0dB.

(*) Nous remercions M. Eskenazi (LIMSI) de sa collaboration.

3.3. Protocoles de reconnaissance

Un protocole est défini par :
- le type d'étiquetage (C ou L),
- le nombre de références choisies par classe (références uniques ou multiples),
- le principe de définition des références (références intra-classe, extra-classe ou moyenne)
- le choix de la distance (LLR, MFCC ou APS).

Soit au total, 36 expériences de reconnaissance menées successivement sur l'ensemble d'apprentissage et l'ensemble de reconnaissance.

3.3.1. Ensembles d'apprentissage et de reconnaissance

Ces ensembles sont construits par une procédure relative à l'étiquetage de la base de données.
L'ensemble d'apprentissage, à partir duquel sera défini l'ensemble des "formes référence", est construit en répartissant les noyaux vocaliques dans les douze classes d'apprentissage (chacune de cardinal fixé) selon leur ordre d'apparition dans le signal de parole, jusqu'à l'obtention du cardinal voulu.
L'ensemble de reconnaissance, constitué des "formes inconnues" à reconnaître parmi les formes référencées, est déduit par complément de l'ensemble d'apprentissage à la base de données.

3.3.2. Formes référence

1. Protocoles à références uniques et multiples.

Le premier autorise une seule référence par classe, le second, une ou plusieurs références, celles-ci étant décidées par analyse d'histogrammes.
... Distance APS : on analyse les histogrammes de fréquences, largeurs de bande et amplitudes des pics de même numéro.
Pour l'étiquetage C, quatre classes ont été divisées car leur distribution de fréquences ne semblait pas unimodale. Les séparations sont effectuées sur critère de seuil de fréquence :
- du second pic des classes /ã/, /õ/ et /i/
- du premier pic de la classe /oe/.

On retrouve les mêmes critères de séparation de classe dans l'analyse relative à l'étiquetage L, avec un autre dédoublement, celui de la classe /a/ selon un seuil de fréquence du premier pic.
... Distances LLR, MFCC : l'analyse des histogrammes des coefficients LPC et MFCC n'a permis de déterminer aucun critère de séparation de classe. Cependant, afin de ne pas défavoriser ces distances, on utilisera, pour un protocole à références multiples, les dédoublements obtenus grâce aux histogrammes de fréquences.

2. Définition des "formes référence" réelles et moyennes

a. Formes réelles intra-classe.

Le représentant de la classe C_i est le spectre S' ($S' \in C_i$) minimisant la distance intra-classe :

$$\left\{ \frac{1}{|C_i|-1} \sum_{S \in C_i} d(S',S) \right\}$$

C'est le représentant qui reconnait au mieux les éléments de sa classe.

b. Formes réelles extra-classe.

Le représentant de la classe C_i est le spectre S' ($S' \in C_i$) qui minimise :

$$\left\{ \frac{\frac{1}{|C_i|} \sum_{S \in C_i} d(S',S)}{\min_{j \neq i} \frac{1}{|C_j|} \sum_{S \in C_j} d(S',S)} \right\}$$

Remarque : dans le cas de la distance APS, pour les références intra-classe et extra-classe, la "forme invariante" est définie par le nombre de pics du représentant de la classe.

c. Formes moyennes.

Le représentant de la classe est un vecteur moyen (LLR, MFCC) ou un spectre moyen (APS).
Le spectre moyen est défini par moyenne des caractéristiques des pics de même numéro. La forme invariante (i) est définie par le maximum des n, tel que tout spectre de la classe admette au moins n pics.

IV . RESULTATS EXPERIMENTAUX

4.1. Scores de reconnaissance

La reconnaissance est effectuée à l'ordre 1, 2 et 3 sur les ensembles d'apprentissage (E.APP) et de reconnaissance de (E.REC) suivant les différents protocoles.
Les résultats actuels sont les suivants :

Etiquetage C

REF.	INTRA-CLASSE			MOYENNE			EXTRA-CLASSE			
ORDRE	1	2	3	1	2	3	1	2	3	
APS	79.3	93.4	96.2	71.4	83.4	88.3	71	87.6	92.1	E. APP.
MFCC	70.3	86.2	92.8	76.9	89	96.2	71	84.1	92	E. APP.
LLR				73.8	91.7	94.1				E. APP.
APS	73.5	89.7	93.2	70.1	81.2	89.3	69.7	82.9	91.5	E. REC.
MFCC	66.2	81.6	90.6	70.5	83.8	92.3	63.7	79.1	88.5	E. REC.
LLR				67.9	81.6	90.2				E. REC.

a1. protocole à références uniques

REF.	INTRA-CLASSE			MOYENNE			EXTRA-CLASSE			
ORDRE	1	2	3	1	2	3	1	2	3	
APS	79.7	94.1	97.9	87.2	95.5	98.6	77.9	91	96.6	E. APP.
MFCC	71.7	85.2	92.8	80	91	95.5	76.6	87.2	91.7	E. APP.
LLR				78.3	92.1	95.5				E. APP.
APS	76.1	92.3	97	81.6	93.2	98.7	76.9	88.9	95.7	E. REC.
MFCC	66.7	86.7	93.6	75.2	89.7	95.7	78.2	90.2	97	E. REC.
LLR				73.9	87.6	93.6				E. REC.

b1. protocole à références multiples

Etiquetage L

REF.	INTRA-CLASSE			MOYENNE			EXTRA-CLASSE		
ORDRE	1	2	3	1	2	3	1	2	3
APS	70.7	85.2	91	64.1	78.3	85.2	65.9	82.4	86.6
MFCC	68.3	79.7	85.5	69	80.7	85.9	63.1	77.2	84.8
LLR				70.3	83.8	89.3			
APS	62	80.3	87.2	58.1	77.4	84.6	65	81.6	88.5
MFCC	60.7	71.8	82.5	60.7	72.2	80.3	56	69.7	80.8
LLR				58.1	75.6	86.8			

E. APP. / E. REC.

c1. protocole à références uniques

REF.	INTRA-CLASSE			MOYENNE			EXTRA-CLASSE		
ORDRE	1	2	3	1	2	3	1	2	3
APS	75.2	90.3	95.2	79	93.4	95.5	74.1	89	94.5
MFCC	69	85.9	90	74.1	87.9	91.4	71.7	84.8	89.7
LLR				76.2	89.7	94.5			
APS	67.5	87.2	92.7	73.1	90.2	96.2	69.2	88	95.3
MFCC	60.3	80.8	88.9	63.7	83.8	90.6	63.7	81.2	87.2
LLR				62.8	83.8	91			

E. APP. / E. REC.

d1. protocole à références multiples

On récapitulera les performances des différentes distances pour les deux étiquetages, en sélectionnant les meilleurs scores de reconnaissance sur l'ensemble de reconnaissance pour les protocoles à références uniques et multiples.

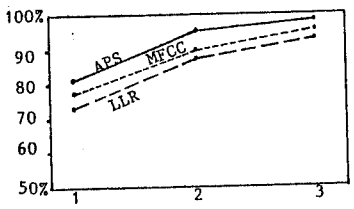
Etiquetage C

ORDRE	1	2	3
APS (intra.)	73.5	89.7	93.2
MFCC (moy.)	70.5	83.8	92.3
LLR (moy.)	67.9	81.6	90.2

a2. protocole à réf. uniques

ORDRE	1	2	3
APS (moy.)	81.6	93.2	98.7
MFCC (extra.)	78.2	90.2	97
LLR (moy.)	73.9	87.6	93.6

b2. protocole à réf. multiples



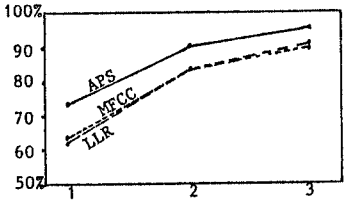
Etiquetage L

ORDRE	1	2	3
APS (extra.)	65	81.6	88.5
MFCC (moy.)	60.7	72.2	80.3
LLR (moy.)	58.1	75.6	86.8

c2. protocole à réf. uniques

ORDRE	1	2	3
APS (moy.)	73.1	90.2	96.2
MFCC (moy.)	63.7	83.8	90.6
LLR (moy.)	62.8	83.8	91

d2. protocole à réf. multiples



4.2. Analyse des résultats

L'étude des minima des courbes de variation spectrale au voisinage des noyaux vocaliques (11) montre qu'ils coincident avec la segmentation automatique dans 88 % des cas, 11 % indiquent un décalage d'une fenêtre, dans 1% des cas, le minimum n'est pas détecté (7).

Le taux d'erreur de reconnaissance mesure une certaine "imperfection" de la mesure de distance. Quel que soit le principe de reconnaissance, on remarque une nette dégradation des performances (particulièrement à l'ordre 1, environ 10%) en passant de l'étiquetage C à l'étiquetage L ; le taux d'erreur est donc lié à la classification perceptive qui conditionne la reconnaissance.

Le choix de la procédure d'étiquetage est donc à considérer, même si l'étiquetage L est le seul à répondre à l'application du décodage acoustico-phonétique.

La différence de taux de reconnaissance observée entre les deux contextes supporte l'idée qu'il convient de considérer un contexte plus "large" qu'une simple fenêtre de 22 ms.

Nous étudions actuellement la reconnaissance des triphonèmes (modèles VCV) où une part plus importante de l'information contextuelle est prise en compte ; les noyaux vocaliques y jouent le rôle de points d'ancrage limitant l'arbre de recherche des triphonèmes.

Les meilleurs scores de reconnaissance résultent des protocoles à références multiples (références moyennes). Ce résultat n'est pas étonnant puisque les références multiples visent à une meilleure représentation des classes. Rappelons néanmoins que les critères de dédoublement de classe sont liés à la paramétrisation APS , aucun dédoublement ne semble fondé sur les paramétrisations LLR ou MFCC.

L'analyse des matrices de confusion montre que la classe des /õ/ pénalise fortement la reconnaissance à l'ordre 1.
L'étude dans la paramétrisation APS des modèles référence (multiples) montre la proximité des différents modèles /õ/ et /ã/. A l'ordre 1, si les éléments de la classe /ã/ sont "bien" reconnus, les éléments de la classe /õ/ sont très souvent reconnus par un modèle de /ã/. Une étude particulière est à envisager pour la classe /õ/, une seule fenêtre ne suffit vraisemblablement pas à la reconnaissance.
Les scores résultant des références intra-classe et extra-classe sont souvent comparables. Pour l'APS, dans un protocole à référence unique, ils sont toujours supérieurs aux scores résultant des références moyennes. Dans ce cas, la référence moyenne est inadaptée : les classes /ã/ et /õ/, par exemple, ont des spectres très différents suivant que le premier pic est dédoublé ou non. Le spectre moyen résultant n'est alors plus représentatif de la classe, et ce qui est plus grave, trop "proche" d'autres modèles.
Si les conditions d'analyse LLR et APS sont identiques, des réserves peuvent être émises quant au choix du nombre de coefficients MFCC.
Les études (3), (2), (6) s'accordent à présenter de bons résultats pour 6, 8 et 10 coefficients, néanmoins, la comparaison avec une paramétrisation LPC d'ordre 16 peut être au désavantage de la MFCC. Nous espérons lever cette ambiguité lors de la présentation.

VI. CONCLUSION

Nous notons pour l'APS, l'importance des coefficients de pondération dans le calcul de la distance.
Ces coefficients sont quantifiés à partir de connaissances sur la perception sonore mal adaptées à notre étude, ils ont donc peu de chance d'être optimaux.
Deux études sont en cours :
- l'une, sur un système de correction d'erreurs de la distance actuelle,
- l'autre, sur l'optimisation par rapport à l'ensemble d'apprentissage de ce type de distance.

Conscients des réserves à émettre sur la taille et la représentativité de la base de données, (sous-représentation des voyelles /ɔ/ et /w/), nous remercions J.P. Tubach et L. Miclet de l'ENST, pour les expériences en cours sur un corpus plus important.

BIBLIOGRAPHIE

(1) M.J. CARATY, X. RODET
"Distance interspectrale à critères perceptifs". JEP, 14ème journées d'études sur la Parole, ENST, Paris 10-13 juin 1985.

(2) G. CHOLLET, C. GAGNOULET
"On the evaluation of speech recognizers and data base using a reference systems".
ICASSP, 1982.

(3) S.B. DAVIS, P. MERMELSTEIN
"Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences".
IEEE, vol. ASSP-28, Aug. 1980.

(4) P. DELEGLISE
"Paramétrisation et détermination des noyaux stationnaires en vue de la reconnaissance de la parole continue".
Thèse de 3ème cycle, PARIS VI, 1983.

(5) J.L. FLANAGAN
"Speech analysis, synthesis and perception".
Springer-Verlag, Berlin 1972.

(6) C. GAGNOULET, M. COUVRAT
"Seraphine : a connected word speech recognition system".
IEEE, vol. ASSP- 36, 1982.

(7) A.H. GRAY, J.D. MARKEL
"Distances measures for speech processing".
IEEE, vol. ASSP- 24, oct. 1976.

(8) R.M. GRAY
"Distortion measures for speech procesing".
IEEE, vol. ASSP-28, 1980.

(9) F. ITAKURA
"Minimum prediction residual principle applied to speech recognition".
IEEE, vol. ASSP -23, feb. 1975.

(10) D.H. KLATT
"Prediction of perceived phonetic distance from critical-band spectra : a first step".
IEEE, ICASSP, Aug. 1982.

(11) L. LEVY, D. DUPHLOUX
"Utilisation de la paramétrisation MFCC dans la reconnaissance des noyaux stationnaires de la parole continue".
Rapport de stage de DEA de reconnaissance des formes, PARIS VI.

(12) N. NOCERINO, F.K. SOONG, L.R. RABINER, D.H. KLATT
"Comparative study of several distortion measures of speech recognition".
IEEE, vol. ASSP- 23, 1985.

(13) H. PIGOT
"Paramétrisation de la parole continue. Apprentissage binaire des noyaux vocaliques".
Thèse de 3ème cycle, PARIS VI, 1985.

(14) R. VISWANATHAN, J. MAKHOUL, W. RUSSEL
"Towards perceptually consistent measures of spectral distance".
ICASSP 1981.

NOMOGRAMMES ET SYSTEMES VOCALIQUES

Louis-Jean BOË & Christian ABRY

Institut de la Communication Parlée
Institut de Phonétique
Grenoble

ABSTRACT

Presented by FANT in 1960, nomograms have not been thoroughly exploited in studying vocalic productions. Only recently were FANT's nomograms confronted with natural vowels by two phoneticians producing a full range of these articulations.

However, in spite of its simplicity, a four-tube articulatory-to-acoustic modeling can be used in studying main vowel types. From the nomograms we shall introduce the notion of focalization to discuss a hypothesis concerning prediction of vocalic systems.

INTRODUCTION

Avec le recul on apprécie d'autant plus la richesse de l'incontournable "Acoustic Theory of Speech Production". A ne s'en tenir qu'à la simulation des voyelles [1], il faut plus que de l'intuition pour modéliser l'essentiel de la substance vocalique, à partir de quatre paramètres "évidents", caractérisant la constriction intra-buccale et le pavillon labial.

Pourtant il ne semble pas que les nomogrammes aient été exploités comme ils auraient pu l'être, pour prédire les universaux des systèmes vocaliques. A la vérité on peut même dire qu'une seule des propriétés de la relation articulatori-acoustique que révèle ce type de document a bénéficié de développements théoriques. L'hypothèse quantique de STEVENS [2] n'exploite en effet que les stabilités relatives entre déplacements articulatoires et variations acoustiques. Pour obtenir du nouveau sur les nomogrammes, il a fallu attendre ... Ce n'est qu'en 1983 que deux phonéticiens [3] - et non des moindres - s'essaient à reproduire des abaques "humains" et les comparent à ceux de 1960.

Les remarques que nous présentons ici prennent place dans ce regain d'intérêt pour un outil d'analyse et de réflexion qu'aucun autre ne semble avoir véritablement remplacé jusqu'à présent. Les fonctions de sensibilité proposées par FANT & PAULI [4] et appliquées au français par MRAYATI & CARRE [5] se sont vite révélées inexploitables, par le fait même de la non linéarité du système et de l'arbitraire des variations de la fonction d'aire.

Pour exploiter stabilités et focalisations ainsi que nous le montrerons, la recherche doit être menée en trois étapes :

1. Tirer les principes généraux de nomogrammes de FANT qui sont issus de modélisation de fonctions d'aire simplifiées;
2. Utiliser les modèles articulatoires disponibles pour limiter les variations arbitraires sur les fonctions d'aire;
3. Rejoindre par cette modélisation les productions humaines qui présentent dans l'état actuel des données encore trop de paramètres incontrôlés pour être directement interprétés [3] et [6].

Nous nous limiterons ici à la première étape en proposant une discussion à partir d'une lecture interprétative des nomogrammes analogiques de FANT que nous avons simulés numériquement.

Pour les prolongements vers la seconde étape, nous renvoyons au travail connexe [7] présenté dans ces mêmes actes.

DES NOMOGRAMMES EN SONAGRAMMES

Parmi les représentations du conduit vocal proposées par FANT nous retiendrons le modèle à 4 tuyaux, d'aire A_i et de longueur l_i, avec les spécifications suivantes :

A_1 = 4 cm² et 0.16 cm² (lèvres)
A_2 = 8 cm² (cavité d'avant)
A_3 = 0,65 cm² (constriction)
A_4 = 8 cm² (cavité d'arrière)
l_1 = 1 cm et $l_2 + l_3 + l_4 = 15$ cm
l_3 = 5 cm

Nous avons utilisé un programme de simulation électrique implanté par CHARPENTIER au CNET [8]. Sont prises en compte les pertes par viscosité, chaleur, vibration des parois et l'effet de rayonnement. La fonction de transfert est calculée de 0 à 5 kHz. Nous avons déplacé la constriction (milieu de A_3 repéré par rapport à la glotte) de

-2.3 cm à 17.3 cm, soit au total 96 simulations pour les deux conditions d'aperture aux lèvres présentées ici. Nous proposons une représentation de type sonagraphique en codant la fonction de transfert sur 8 niveaux par pas de 8 db (FENG [9]). (figures 1).

REMARQUES SUR LES STABILITES FORMANTIQUES

La discussion sur la stabilité des voyelles porte en premier lieu sur la localisation et le degré de constriction.

En utilisant la modélisation à 4 tubes pour toutes les voyelles (STEVENS l'utilise pour [u] mais simplifie à 2 tubes pour [a] et [i]) on peut avancer les constatations suivantes pour F1 et F2 :

1. La voyelle [u] est bien stable pour une grande gamme de variations de ces 2 paramètres. C'est ce que montraient déjà les monogrammes de FANT et ce que nous avons pu simuler [7] avec le modèle de MAEDA [10].
2. Pour [i], la stabilité proposée par STEVENS avec un modèle à 2 tubes ne tient plus, lorsque la simulation est affinée. Avec 4 tuyaux, comme avec une fonction d'aire plus complexe (la "horn-shaped tongue section" de FANT [1] par exemple), cette voyelle révèle une très forte sensibilité aux variations de lieu pour F2 comme à celles de constriction pour F1 et F2. Les variations de la langue modélisée pour ces deux paramètres par LINDBLOM & SUNDBERG [11] , GOLDSTEIN [12] PERKELL & NELSON [6] - et encore plus nettemnt les nôtres [7] - mettent en évidence cette sensibilité extrême du [i] . Ne sont donc contradictoires que les résultats obtenus par LADEFOGED & BLADON [3] en essayant de reproduire les nomogrammes. Pour eux F1 et F2 restent stables, à constriction égale, quel que soit le degré d'avancement de la langue. Précision articulatoire du phonéticien ou justesse du modèle ? Là est la question.

Ceci ne peut remettre en cause, en ce qui concerne la stabilité supposée de cette voyelle, que les arguments basés sur la relation articulatori-acoustique. Mais aucunement les précisions ou imprécisions observées. Celles-ci peuvent dépendre des contrôles articulatoires, éventuellement meilleurs pour cette voyelle [6, 13].
3. En ce qui concerne la voyelle [a] les différences de modélisation ne changent pas les résultats de STEVENS sur l'insensibilité aux variations de lieu. Pour la constriction on peut retenir que c'est surtout F2 qui est sensible aux variations de ce paramètre. En revanche, F1 reste remarquablement stable et c'est, rappelons-le, lui qui confère traditionnellement au [a] son caractère de voyelle basse.

CONVERGENCES FORMANTIQUES REMARQUABLES

L'analyse des différents nomogrammes qui intègrent la complexification des modélisations disponibles (du modèle 4 tubes aux modèles articulatoires) offre d'autres possibilités de lecture que le repérage des stabilités articulatori-acoustiques.

Alors que ces dernières ont fait l'objet d'une réflexion fondamentale (dans le cadre de la théorique quantique), peu d'attention semble avoir été portée jusqu'à présent au phénomène suivant. En déplaçant le lieu de constriction on produit un jeu d'éloignement et de rapprochement des formants et ceci d'autant plus que la constriction est petite (c'est à dire que le couplage entre grandes cavités est faible). A la limite on obtient des "croisements" de formants - déjà relevés par STEVENS [2] - et qui viennent d'être replacés dans la théorie des catastrophes de THOM par PETITOT-COCORDA [14] . Utilisée par STEVENS pour les consonnes, notamment pour expliquer le "hub locus" des vélo-palatales [k, g] , la convergence F2, F3 apparaît ainsi comme un point catastrophique, où se produit un échange d'affiliation entre formants et cavités.

Nous pouvons reprendre à notre compte, pour les voyelles, une telle approche. Le cas du [i] est exemplaire. L'instabilité de F2, F3 est maximale au point de rapprochement. L'échange de résonateur (avant/arrière) va provoquer un basculement des amplitudes et des bandes passantes, visible sur le nomogramme sonagraphique (figures1). Ce qui donne l'occasion d'observer des [i] avec un point de confusion F2, F3. Il est possible de simuler ce phénomène à partir d'une fonction d'aire de [i] , ainsi que nous l'avons fait (figure 2); ou encore en les recherchant dans le dictionnaire des 13^5 configurations que nous avons généré avec le modèle de MAEDA [7] .

Ce phénomène permet sans doute d'expliquer la difficulté apparente de certaines mesures pour le [i] et des résultats peu homogènes sur F2, F3. C'est encore ce qui explique le perplexité de LADEFOGED & BLADON [3] lorsqu'ils perdent le suivi formantique dans leur série de production. La langue s'avançant, et se rapprochant du point d'échange, ils constatent : "... there is only a single formant visible at about 2100 Hz, and F3 would appear to have suddenly assumed a value comparable to that of F4 in the previous vowel [...]. It is difficult to relate acoustic behaviors such as these sudden formant dicontinuities to the articulatory states which produced them".

En réalité F3 ne fait que se confondre avec F2 qui croît avec l'avancée du lieu de constriction. Ce prétendu F3, atteint de discontinuité brutale, n'est bien sûr que F4 qui lui n'a pas changé. Pour qui sait les prévoir ces résultats sont parfaitement visibles sur de la parole naturelle.

VOYELLES FOCALES

En généralisant cette approche que nous avons menée pour [i] , il est possible d'utiliser d'autres convergences que celles de F2, F3 et pour d'autres types vocaliques.

Convergence F1, F2

Le [a] à lèvres ouvertes (fig. 1a) et le [u]

à lèvres fermées (fig. 1b').

Convergence F2, F3

C'est le [i] à lèvres ouvertes (fig. 1a') et le [y] à lèvres fermées (fig. 1b').

Convergences F3, F4, F5

Celles-ci se rencontrent pour le [u] déjà cité mais est peu exploitable à lèvres fermées : les formants élevés sont, dans ce cas, de très faible amplitude (seule la convergence F1, F2 peut caractériser nettement cette voyelle). En revanche à lèvres ouvertes il existe une voyelle qui peut présenter cette convergence c'est le [ɯ] (figure 1a'). D'autres convergences pourraient être exploitées, notamment pour les consonnes, comme l'a proposé STEVENS [2].

En ce qui concerne simplement les voyelles on peut déjà remarquer sur les nomogrammes issus de simulations plus complexes qu'il faut vraisemblablement compter avec des convergences additionnelles F4, F5 pour [a] et [i].

[i, a, u] seraient ainsi 3 voyelles doublement focales, [u] étant celle des trois qui manifesterait le moins cette double polarité.

Les deux autres voyelles sont monofocales : [y] par perte de la convergence F4, F5 que possède [i] et [ɯ] par perte de la convergence F1, F2 de [u].

CONCLUSION

Les critères perceptifs actuellement utilisés - à partir des travaux du groupe de Léningrad - par FANT [15] et SYRDAL [16] pour évaluer la distance critique entre les formants (3.5 barks) et classer les voyelles pourraient être retenus comme seuil de nos convergences.

Dans un espace où la convergence (< 3.5 barks) F1, F2 caractérise les voyelles d'arrière de [u] jusqu'à [a] (selon les auteurs) et la convergence F2, F3 les voyelles d'avant, il reste que les convergences remarquables tirées des nomogrammes pourraient expliquer certains universaux des voyelles [i, a, u] (diversement bifocales) bien sûr et des dérivées monofocales plus rares [y] et [ɯ].

REFERENCES

[1] G. FANT, Acoustic Theory of Speech Production. Mouton, The Hague, 1960.

[2] K.N. STEVENS, The Quantal Nature of Speech : Evidence from Articulatory-Acoustic Data. In Human Communication. Mc Graw Hill, New-York, 1972.

[3] P. LADEFOGED & A. BLADON, Attempts by Human Speakers to Reproduce Fant's Nomograms. Speech Comm. 1, 185-198, 1982.

[4] G. FANT & S. PAULI, Spatial Characteristics of Vocal Tract Resonance Modes. Speech Communication Seminar 2, 121-132, 1974.

[5] M. MRAYATI & R. CARRE, Relations entre la forme du conduit vocal et les caractéristiques acoustiques des voyelles françaises. Phonetica 33, 285-306, 1976.

[6] J.S. PERKELL & W.L. NELSON, Variability in Production of the Vowels /i/ and /a/. J. Acoust. Soc. Am. 77, 1889-1895, 1985.

[7] R. MAJID, L.J. BOË & P. PERRIER, Fonctions de sensibilité, modèle articulatoire et voyelles du français. 15èmes JEP GCP-GALF, 1986.

[8] F. CHARPENTIER, Un logiciel de simulation du conduit vocal. CNET Comm. Personnelle, 1982.

[9] G. FENG, Analyse cepstrale, visualisation sonagraphique et détection des formants. Seminaire Traitement du signal de Parole GALF-GRECO, 207-216, 1984.

[10] S. MAEDA, Un modèle articulatoire de la langue avec des composantes linéaires. 10èmes JEP, GCP-GALF, 152-162, 1979.

[11] B.E.F. LINDBLOM & J.E.F. SUNDBERG, Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement. J. Acoust. Soc. Am. 50, 1166-1179, 1971.

[12] L. GOLDSTEIN, Vowel Shifts and Articulatory-Acoustic Relations. 10th Int. Congr. Phonetic Sci. IIA, 267-273, 1983.

[13] J.S. PERKELL, On the Nature of Distinctive Features : Implications of a Preliminary Vowel Production Study. In : Frontiers of Speech Comm. Research, 365-380. Academ. Press. London 1979.

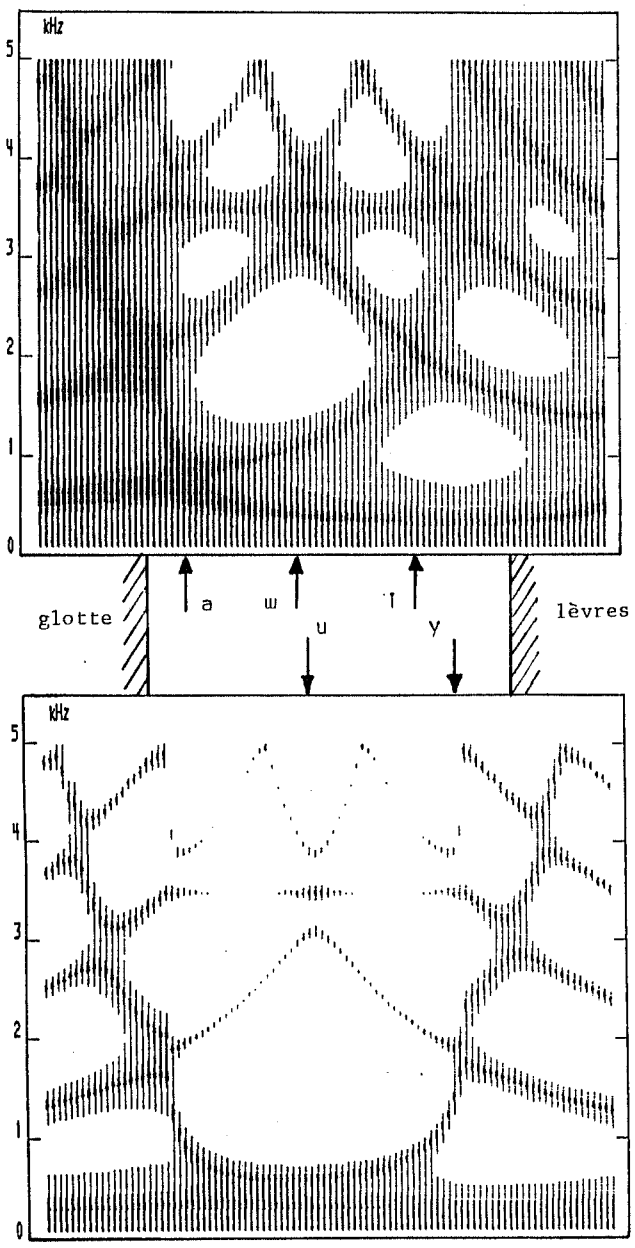
[14] J. PETITOT-COCORDA, Les catastrophes de la parole. Maloine, Paris, 1985.

[15] G. FANT, Feature Analysis of Swedish Vowels. STL-QPSR 2-3, 1-19, 1983.

[16] A.K. SYRDAL, Aspects of a Model of the Auditory Representations of American English Vowels. Speech Comm. 4, 121-135, 1985.

Remerciements

Une partie de ce travail a été discutée avec Shinji MAEDA et Francis CHARPENTIER à l'occasion d'un exposé au CNET Lannion. Nous avons tenu compte de leurs remarques et suggestions.



Figures 1 - Nomogrammes sonagraphiques de type FANT.

a. A_1 = 4 cm² $A_2 = A_4$ = 8 cm² A_3 = 0.65 cm²

b. A_1 = 0.16 cm² $A_2 = A_4$ = 8 cm² A_3 = 0.65 cm²

Dans les limites du lieu de constriction : voyelles focales [i, a, u, y, ɯ].

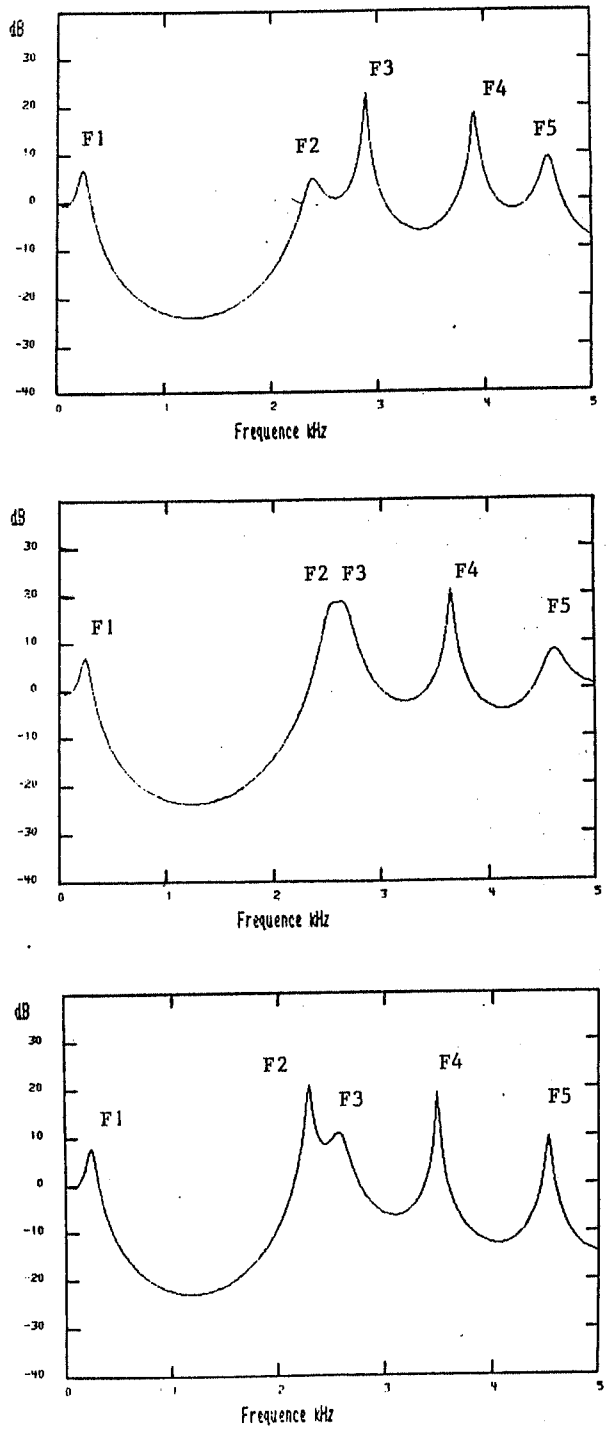


Figure 2 - Fonctions de transfert obtenues à partir d'une fonction d'aire réaliste de [i]lorsque l'on déplace légèrement le lieu de constriction. On notera la réalisation focale et l'effet d'échange de résonateur pour F2, F3 (amplitudes et bandes passantes).

VOTS ET VTT EN FRANCAIS

R. Sock & C. Benoit

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

ABSTRACT

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

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

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

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

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

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

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

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

CORPUS ET MESURES

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

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

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

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

Les VOTs

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

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

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

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

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

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

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

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

Le VTT

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

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

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

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

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

3. Le timing intrasegmental du VTT

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

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

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

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

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

Remerciements à :

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

REFERENCES

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

APPENDICE

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

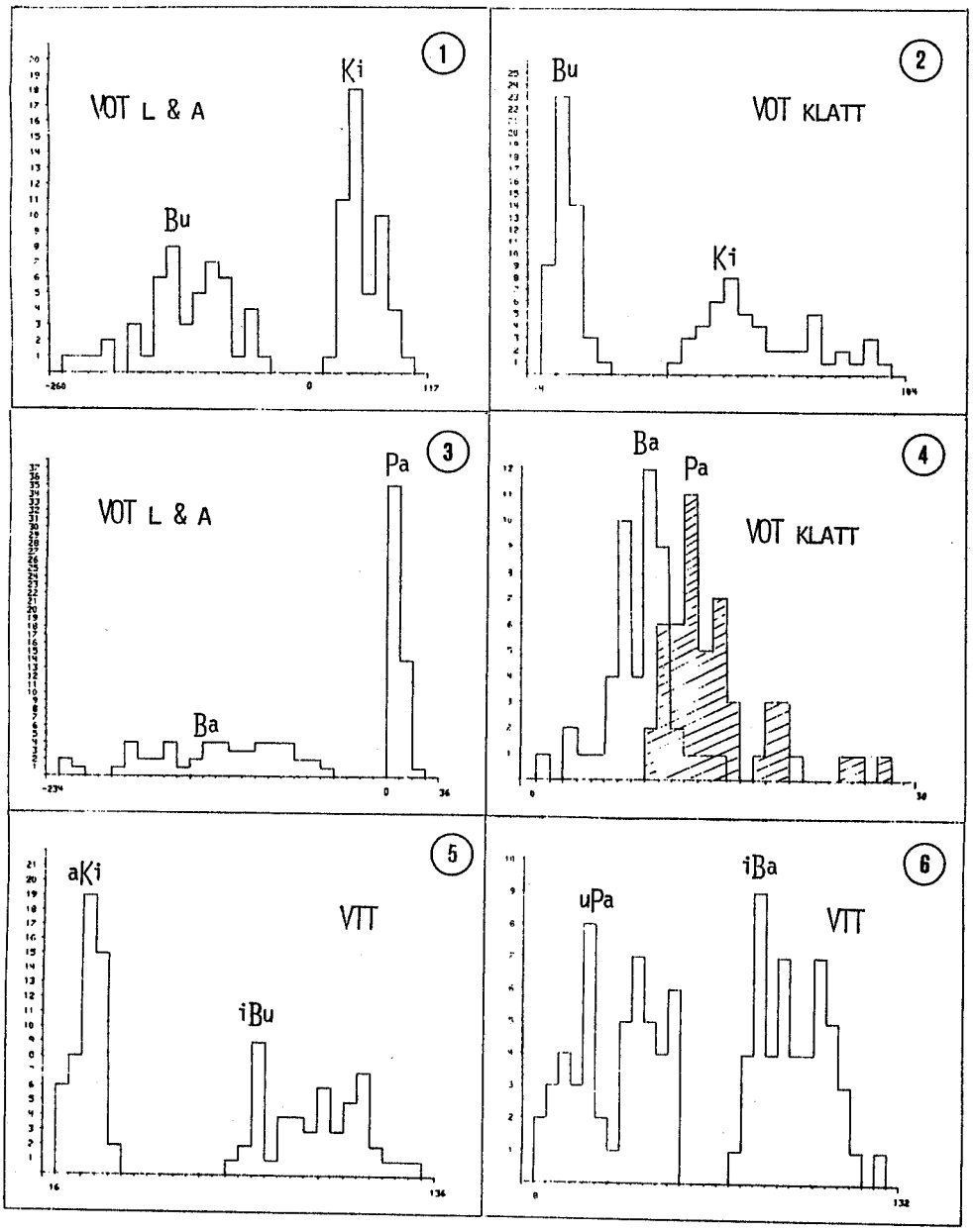


Fig. 1-2, Contraste maximal, VOT LISKER & ABRAMSON vs. VOT KLATT pour /buki/ et /kibu/ (5 loc. x 10).

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

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

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

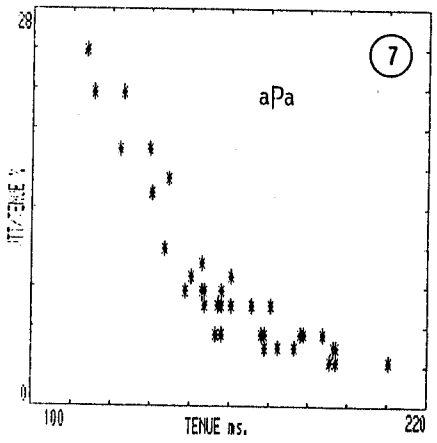


Fig. 7. Le phasage du VTT par rapport à la TENUE pour /kapa/ (4 loc. x 10).

MICRO-INDICES DANS LES OCCLUSIVES SOURDES

C. Barrera, J. Caelen

Laboratoire CERFIA UA-CNRS N°824
Université P. Sabatier, 118 Route de Narbonne
31062 Toulouse CEDEX.

ABSTRACT

Through an ear-model very short term analysis, we study the burst pattern of unvoiced plosive.
To this purpose a set of cues and micro-cues is presented, discussed and verified by statistical analysis. We believe that the use of these cues in the field of acoustic-phonetic decoding will be of a great interest.

1. INTRODUCTION

L'étude de phénomènes brefs dans la parole nécessite des systèmes d'analyse adaptés, à large bande passante. Certaines méthodes FFT, LPC, CEPSTRE utilisent des fenêtres trop longues, même si un recouvrement serré entre elles compense quelque peu ce défaut [1]. Par contre, les bancs de filtres --ainsi que leurs dérivés, les modèles physiologiques-- sont mieux adaptés à une analyse à très court terme [2].

Pour les occlusives, Delgutte [3] utilise un modèle d'oreille qui a la propriété de faire émerger ces phénomènes brefs. Une idée similaire est mise en œuvre dans cet article en tentant de plus d'interpréter les "pattern" des décharges dans toutes les fibres nerveuses pour les sons transitoires.

Nous utilisons un modèle dont les caractéristiques principales [4] sont :

- un filtre passe-bande d'ordre 4 simulant l'adaptation oreille externe-moyenne.
- la membrane basilaire modélisée par 24 filtres couplés du second ordre.
- les cellules ciliées internes et les fibres afférentes constituées par un système redresseur-intégrateur dont la constante de temps (2ms) ne lisse pas trop les impulsions.
- une compression de dynamique logarithmique et une visualisation pseudo-sonagraphique améliorant la lecture des résultats en vue de leur interprétation. Nous appelons cette visualisation un neurogramme (Cf. fig 1).

L'adaptation à court terme et l'inhibition latérale sont absentes du modèle, volontairement simplifié pour cette étude. Par contre le couplage des filtres introduit une composante intéressante dans l'analyse de la parole (lissage et accentuation de la compacification du spectre).

Le signal est analysé par ce modèle, sur une fenêtre à très court terme (moins de 1 ms) dans des conditions précisées ci-après. Les neurogrammes obtenus constituent la base des observations de notre étude.

2. OBSERVATIONS - MODE OPERATOIRE

2.1. Corpus

L'étude porte sur le corpus GRECO de logatomes bi-syllabiques comprenant des occlusives sourdes en contexte CVCV(C). Le signal est échantillonné à 16 kHz et codé sur 16 bits. Les occlusives sont en position initiale, intervocalique et parfois en finale (ex. /kykyk/) et les différentes voyelles sont toutes présentes et répétées dans chaque syllabe. Trois fichiers ont été analysés ce qui constitue un ensemble de 347 occlusives statistiquement équilibrées.

2.2. Description acoustico-phonétique des occlusives à travers le neurogramme

Examinons les neurogrammes (fig. 1) dans le plan temps-fréquence.

a) Sur l'axe fréquentiel :

L'explosion apparaît de façon très nette et se manifeste par une augmentation rapide du taux de décharge dans les fibres dans une zone variant entre 1000 Hz /p/ et 4000 Hz /t/. On sait que les 20 à 100 ms suivant l'explosion sont riches d'informations: notamment la forme globale du spectre serait un indice du lieu d'articulation [5], [6], bien que cet indice ne soit pas toujours considéré comme primaire comparé aux transitions formantiques et à leur évolution temporelle [7], [8]. Nous pouvons faire ici des constatations similaires: compacité et énergie élevée des vélaires, étalement des labiales et décalage des dentales vers les hautes fréquences, non jugées diffuses.

b) Sur l'axe temporel :

L'analyse à très court terme restitue la dimension temporelle, contenue dans la notion de V.O.T. (Voice Onset Time) [9]: sur les neurogrammes, elle est déterminée simplement lorsqu'on a positionné les deux "événements" [10] que sont la

barre d'explosion et l'apparition du voisement. De plus la dilatation des phénomènes obtenue par le neurogramme montre que le "pattern" des labiales s'établit plus lentement.

2.3. Méthodologie

Pour quantifier ces phénomènes il faut définir des fonctions appropriées. Le signal est analysé par modèle d'oreille sur une fenêtre de 8 ms. Un étiquetage manuel [11] détermine la fenêtre --centrée sur l'occlusive et bornée par les contextes vocaliques-- sur laquelle le signal subit une analyse à très court terme (moins de 1 ms)

La procédure automatique **DETEC** détermine deux événements importants:

--> l'explosion de bruit, pour laquelle les fibres sont excitées de façon sensiblement synchrone, est déterminée par une fonction (**FX**) qui mesure le degré de cohérence des canaux [2] en tenant compte des fronts montants (somme des dérivées positives dans chaque canal). De plus cette fonction est pondérée par la somme des énergies des canaux dans la bande (700-5600 Hz). Divers traitements sont alors possibles: algorithmes de détection itératifs à seuil majoritaire, détecteurs d'enveloppe, etc. La première solution a été retenue; elle est pilotée par une stratégie de contrôle vérifiant que l'énergie du signal est suffisamment faible, excluant ainsi les pics dus au contexte vocalique postérieur. Dans le cas où l'explosion n'a pas été trouvée --ce qui peut se produire pour les occlusives labiales, très diffuses et énergétiquement faibles-- la procédure est réitérée en choisissant un nouveau seuil dépendant du seuil initial et de la moyenne de **FX** obtenue au premier essai.

--> l'apparition du voisement est déterminée en considérant que les énergies des canaux basses fréquences (180-320 Hz) donnent un pic temporel pour une zone proche du premier formant de la voyelle (ceci est dû au couplage mentionné plus haut)

A ce stade nous disposons d'une mesure du V.O.T. (**2+**) ainsi que d'une estimation de l'attaque du burst par **FX**. Nous calculons aussi l'énergie locale du signal (**WX**) sur 3 ms au voisinage de l'explosion.

Nous introduisons en outre, un certain nombre de paramètres tels que:

- **HB** rapport des énergies des canaux hautes (1130-4000 Hz) et basses fréquences (180-260 Hz) sur environ 3 ms. Cet indice constitue une variante de l'indice Aigu/Grave mieux adapté au "pattern" explosif puisqu'il tient compte de la bande médiane.

- **AX, MX** temps d'amortissement et de montée des décharges dans la fibre la plus excitée à l'explosion.

- **XG, YG**, coordonnées du centre de gravité et **IX** inertie. Pour expliciter ces paramètres, assimilons la portion de neurogramme comprise entre les 2 événements, à un nuage de points (i,j) pondéré par NEU(i,j), décharge de la fibre numéro i à l'instant j. L'inertie IX (indice Diffus/Compact) mesure la dispersion de ce nuage par rapport au centre de gravité G de coordonnées XG YG .

$$IX = \sum_i \sum_j [(i-XG)^2 + (j-YG)^2] \, NEU(i,j)$$

Cet indice tente de s'affranchir de la notion de contexte car il est peu dépendant de la position relative de G, seuls les écarts et l'amplitude étant retenus. **IX** est normalisé par le nombre de points de la structure pour pouvoir comparer des occlusives dont les V.O.T. sont différents.

Ainsi, le segment compris entre la barre d'explosion et le début du voisement est analysé sous tous ses aspects: temporel / spatial / spatio-temporel. La notion de "pattern" relève de cette dernière catégorie et bien que très séduisante, se montre délicate à maîtriser tant les relations temps-fréquence-amplitude sont complexes. Dans une classification synthétique, la longueur des fenêtres d'analyse distingue les indices des micro-indices (Cf. Table 1).

Type \ Nature	Temporel	Amplitude/Temps	Spatio-Temporel
Indice	2+		IX
Micro-indice	AX MX	WX	FX HB

Table 1
Répartition des indices et micro-indices selon leur nature.

3. RESULTATS ET PERSPECTIVES

3.1. Repérage des événements

Pour l'étude des occlusives, dans le système de paramètres défini ci-dessus, les 3 catégories phonémiques /p/, /t/, /k/ ont été regroupées en classes selon

- la position dans la phrase,
- le contexte vocalique environnant, lui-même découpé en 4 classes :

V1 = [/i/, /y/, /e/] V2 = [/u/, /ɔ/, /O/]
V3 = [/a/, /ɑ̃/, /ɔ̃/, /œ̃/] V4 = [/ø/, /œ/, /ɛ/, /ɛ̃/]

La procédure **DETEC** (Cf. 2.3) ne laisse apparaître dans la détection du "burst" que 2,6% d'erreurs de première espèce c'est à dire des erreurs pour lesquelles la barre d'explosion est placée de façon arbitraire, le plus souvent pour /p/. Dans la suite des traitements (Cf. 3.2.) ces observations sont éliminées du fichier d'observations pour ne pas biaiser les analyses de données ultérieures. Ouant aux erreurs de seconde espèce (décalage de moins de 5 ms), il est difficile de les évaluer en raison de l'imprécision du repérage manuel servant de référence (Cf. les multiples définitions du V.O.T., points d'ancrage sur passages par zéro, etc. [10]).

3.2 Validation des indices et micro-indices

(a) Tendances générales :

Le tableau suivant consigne les résultats obtenus:

	nb. obs.	FX	WX	HB	2+ (ms)	AX (oo)	MX (ms)	IX
P	73	6384 5530	17 5	-11.8 5	26.7 11	3.6 2.6	5.5 5	11400 9174
T	83	17985 10791	15.6 4.8	-19.5 2.9	38.9 13.4	4.9 3.8	4.2 4.0	25114 14577
K	74	21247 14457	22 5	-14.5 7.6	43.4 11.5	5.4 3.9	3.4 3.2	29710 16750

Moyenne Ecart-type

POSITION : INITIALE / INTERVOCALIQUE

Ces résultats appellent quelques commentaires:

-. les valeurs du V.O.T. coïncident avec les données connues [9], [12] (V.O.T.(p) <V.O.T.(t) < V.O.T.(k))

- **FX** , **WX** ont le même sens de variation. /k/ étant généralement plus énergétique que /t/ et /p/ mais la comparaison de /t/ et /p/ montre que si l'énergie locale WX de /p/ est relativement élevée, la dérive spatio-temporelle **FX** de /t/ est nettement supérieure.

- un autre élément paraît intéressant: HB prend la valeur la plus faible pour /t/ et son écart-type réduit montre la résistance de l'indice au contexte. Cette remarque n'est plus vraie pour /k/ dont le pattern se décale vers les hautes fréquences devant les voyelles hautes [13].

- bien que **AX** et **MX** aient des plages de variation assez faibles, il est possible de dégager des tendances: /k/ présente un front très raide à l'attaque et un amortissement plus long que /t/. A l'opposé, le temps de montée de /p/ est le plus long (aspect diffus en temps et en fréquence).

- **IX** bien qu'a priori intéressant, s'est avéré trop corrélé avec le V.O.T. Il conviendra dans une deuxième version de tenir davantage compte de la pondération NEU(i,j).

- En revanche 2+ n'est pas corrélé avec **FX**.

(b) Validation par analyse de données

Deux types d'analyses de données ont été utilisées:

* Analyse discriminante : la discrimination théorique (pour les 7 paramètres) entre /p/ et /t/ à l' initiale et intervocalique est de 81% atteignant 89% en excluant la classe V1; entre /p/ et /k/ 85% (resp. 96%) et 58% seulement entre /t/ et /k/ (resp. 66%)

* Analyse en composantes principales : elle confirme la première analyse. Les 3 premiers axes représentent 80% de l'information, les paramètres les plus significatifs étant 2+, IX, HB (voir figures).

4.2. Perspectives

Si les discriminations /p/-/k/ et /p/-/t/ semblent encourageantes, celle de /t/-/k/ n'est pas encore suffisante. Le modèle ne parvient pas à dissocier les "pattern" respectifs et une étude plus poussée du ou des paramètres les plus pertinents (2+ IX HB) reste à faire .

Ces résultats permettent malgré tout de concevoir une procédure de décodage acoustico-phonétique en deux passes:

a) la première sur une fenêtre de quelques millisecondes permettant de mettre en évidence la phase d'occlusion toute entière, à l'aide d'indices robustes.

b) la deuxième ayant pour but l'extraction de phénomènes plus fins sur une fenêtre à très court terme (Cf. procédure **DETEC**).

Certains problèmes, comme le mode d'articulation (opposition sourd/sonore) semblent avoir été éludés. Cependant ils sont aisément intégrables dans notre modèle (mesures du "buzz" et d'un V.O.T. négatif par exemple).

4. CONCLUSION

A partir d'une analyse à très court terme fondée sur un modèle d'oreille et après une détection --automatique et fiable-- de l'explosion, nous avons voulu montrer qu'il était utile d'étudier la structure fine des "pattern" occlusifs. Quelques problèmes de définition et de normalisation des paramètres subsistent encore: indices et micro-indices se révèlent rarement indépendants du contexte et/ou du locuteur. De plus ils ne sont pas nécessairement synchrones entre eux [14].

Mais le type d'analyse que nous avons utilisé élargit le champ d'investigation des indices. D'ores et déjà, les indices décrits ci-dessus seront prochainement intégrés dans une stratégie de décodage à laquelle sera dévolue une pondération hiérarchique et sélective de tous les indices.

5. REFERENCES

[1] C. Searle, J.Z. Jacobson, S.C. Rayment, "Stop consonant discrimination based on human audition", J.A.S.A. 65-3, pp. 799-809, 1979.

[2] J. S. Liénard, "Quelques aspects de l'analyse à très court terme de la parole", Séminaire traitement du signal de parole, pp. 94-105, 1983.

[3] B. Delgutte ,"Codage de la parole dans le nerf auditif", Thèse de doctorat d'état, 256 p., Juin 1984.

[4] J. Caelen, "Space/time data-information in the A.R.I.A.L. project ear Model", Speech communication, Vol. 4, pp. 163-179, Aug. 1985.

[5] S. Blumstein, E. Isaacs, J. Mertus, "The role of the gross spectral shape to place of articulation in initial stop consonants", J.A.S.A. 72-1, pp. 43-50, 1982.

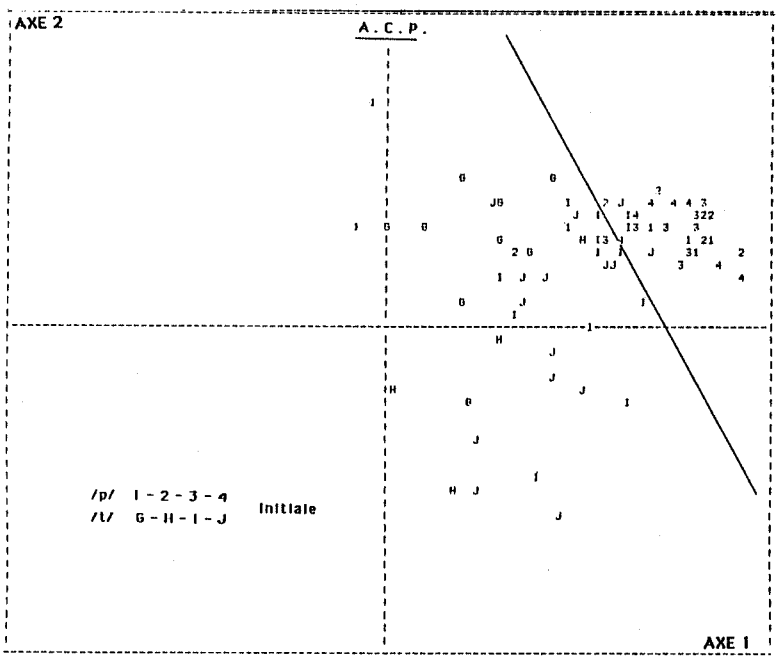
[6] K. Stevens, "Acoustic correlates of some phonetic categories", J.A.S.A. 68-3, pp. 836-842, Sept. 1980.

[7] A. Lahiri, L. Gewirth, S. Blumstein, "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonant: evidence from a cross-language study", J.A.S.A. 76-2, pp. 391-396, April 1984.

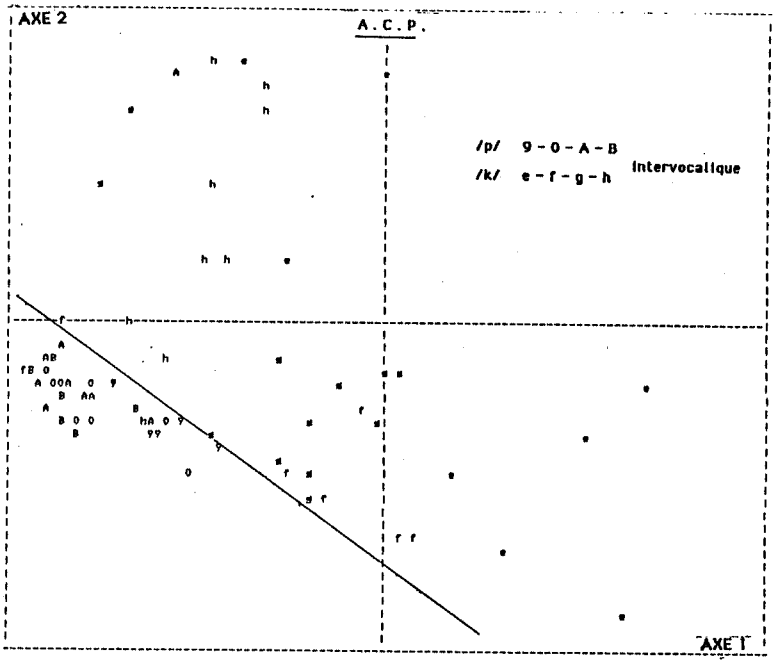
[8] D. Kewley-Port, "Time varying features as correlates of place of articulation in stop consonants", J.A.S.A. 73, pp. 322-335, 1983.

[9] L. Lisker & A. Abramson, "A cross language study of voicing in initial stops : acoustical measurements", Word 20 N°3, pp. 384-422, 1964.

[10] C. Abry, C. Benoît, L.J. Boë, R. Sock, "Un choix d'événements pour l'organisation temporelle du signal de parole", XIV J.E.P. GALF-CNRS, pp. 133-137, Juin 1985.
[11] N. Vigouroux, J. Caelen, "Segmentation en vue de l'organisation d'une base de données acoustiques et phonétiques", XIV J.E.P. GALF-CNRS, pp. 152-155, Juin 1985.
[12] W. Serniclaes, "Fenêtre de prélèvement temporel des indices d'occlusives", XIII J.E.P. GALF-CNRS, pp. 70-78, mai 1984.
[13] C.J. Weinstein, S.S. McCandless, L.F. Mondshein, V.W. Zue, "A system for acoustic-phonetic analysis of continuous speech", I.E.E.E. ASSP Vol. 23, No 1, pp. 54-67, 1975.
[14] F. Lonchamp, "Validation perceptive des indices", I.C.A., XI congrès, Vol. 9, pp. 93-108, Juil. 1983.



Analyse en composantes principales



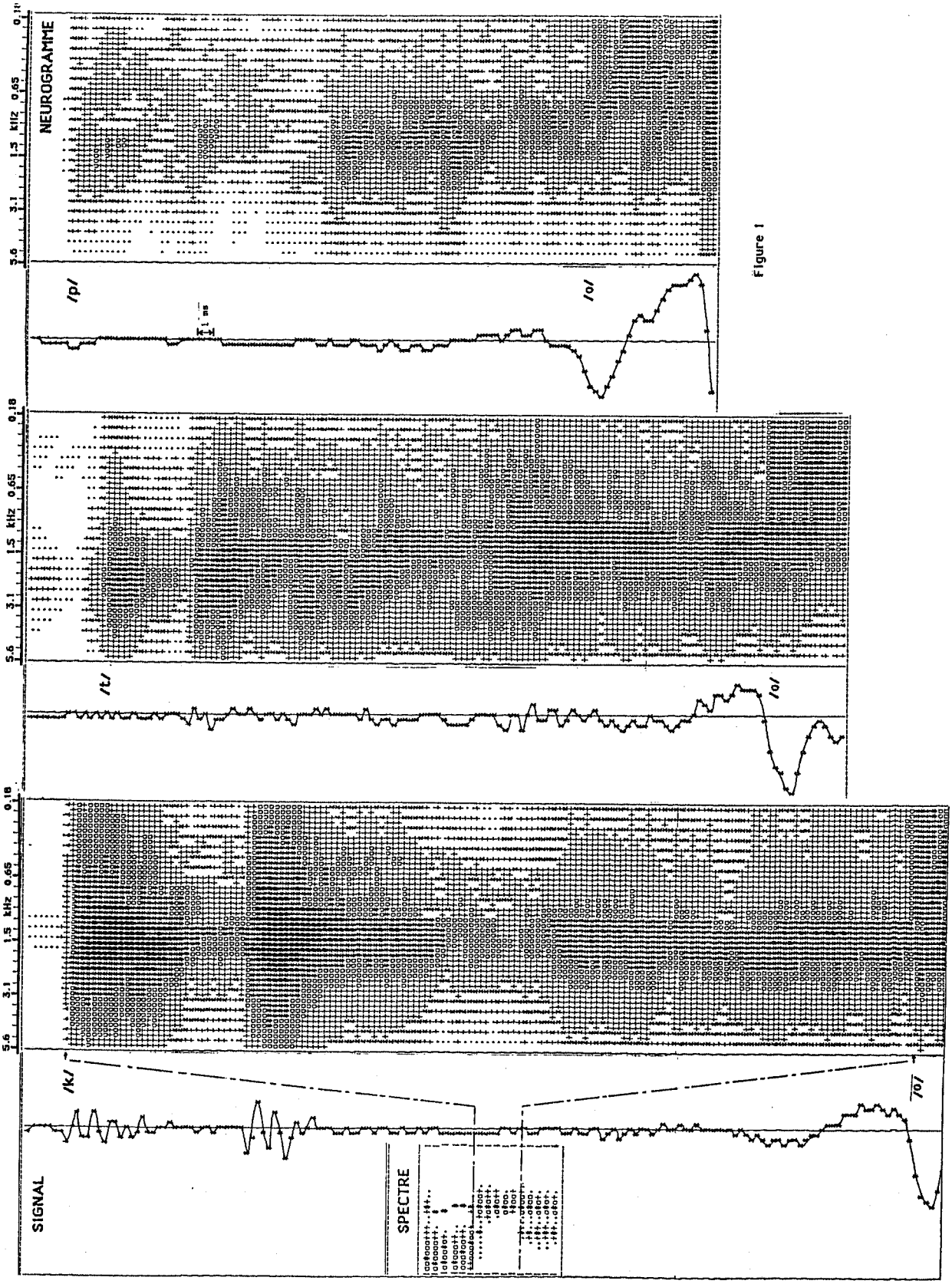


Figure 1

ACOUSTIQUE DU RELACHEMENT DES OCCLUSIVES: UNE ETUDE DE SIMULATION

Shinji Maeda

CNET, 22301, Lannion, France

ABSTRACT

Formulating an articulatory model of the CV-syllables, where C = /p, t, or k/ and V = /i, a, or u/, the speech signals, and the corresponding transfer functions for the consonant release were calculated using a vocal-tract synthesizer. A visual inspection of the transfer functions suggested that the time-varying characteristics of the release spectrum within less than 5 ms after the onset can specify, independently of the following vowel, place of articulation for the consonants. The transfer function, for which the corresponding synthesized signal is perceived as the labial, is characterized by a spectral balance that varies considerably with time. A constant gross spectral shape is characteristic to the dental, and a stable prominent peak through out the release to the velar.

INTRODUCTION

Dans un article précédent [1], nous avons décrit les sources d'excitation du conduit vocal pendant le relâchement des occlusives. Les signaux de source subissent des modifications spectrales particulières qui sont interprétées comme une consonne spécifique par les auditeurs. On sait qu'un certain nombre d'indices acoustiques qui spécifient la place d'articulation pour les consonnes initiales et intervocaliques sont distribuées le long du continuum. Cette étude concerne l'acoustique du burst de relâchement.

Stevens et Blumstein [2,3]ont montré que des propriétés acoustiques invariantes de la place d'articulation sont présentes dans la forme globale du spectre du burst. Il est vraisemblable que la résonance de la cavité frontale [4]est à la base de l'invariance des propriétés, un spectre "plat" pour les labiales, un spectre "montant" pour les dentales, et un spectre "compact" pour les vélaires. En fait, l'hypothèse sous-jacente est que la cavité frontale est acoustiquement isolée de la cavité postérieure, et que la cavité antérieure est le facteur qui domine dans la modification du spectre pendant un temps relativement long après le relâchement.

Nos propres expériences avec un modèle articulatoire simulant le relâchement, indiquent cependant que ceci n'est pas toujours vrai. Dans de nombreuses circonstances, il y a un effet dû à la cavité postérieure sur les spectres suivant immédiatement le relâchement du burst. Ces observations nous ont amené à étudier en détail le rapport entre l'articulation du relâchement et ses manifestations acoustiques. Notre approche consiste, (i) à formuler un modèle de production pour les syllabes CV, où C = /p, t, ou k/ et V = /i, a, ou u/, (ii) à réaliser des expériences de synthèse pour vérifier si le modéle est capable ou non de produire ces mêmes syllabes avec une qualité raisonnable, et (iii) à examiner visuellement l'évolution temporelle des fonctions de transfert du conduit vocal pour déterminer les caractéristiques acoustiques associées avec chacune des trois catégories de consonnes.

MODELE ARTICULATOIRE POUR LES SYLLABES CV

On suppose une forte anticipation de la voyelle pendant l'articulation de la consonne précédente. La fonction d'aire du conduit vocal est celle de la configuration de la voyelle, à l'exception de la partie proche du point d'occlusion où l'aire de section varie dans le temps, représentant un geste localisé pour la consonne. Nous considérons les trois types de configuration du relâchement suivants: (i) le type labial (Type-L), où l'étendue de la constriction est limitée à une seule section correspondant au point de la fermeture (voir Fig.1(a)). On notera que la fonction d'aire est définie par 21 tubes uniformes connectés. (ii) le type dental (Type-D), où la section de fermeture et également la cavité postérieure s'agrandissent de façon à maintenir une transition non abrupte entre la section d'occlusion et la cavité postérieure pendant le relâchement (voir Fig.1(b)), (iii) le type vélaire (type-V) où les deux cavités postérieure et antérieure s'agrandissent pendant le relâchement (voir Fig.1(c)).

On pourra noter ici que pour le type-L, à cause de la forte discontinuité de la fonction d'aire au relâchement, l'effet acoustique de la cavité postérieure doit être réduit. Dans le cas du type-D, la cavité postérieure est connectée de façon non abrupte au point d'occlusion et une influence acoustique plus importante de la cavité postérieure est prédite. En plus de l'effet de la cavité postérieure, pour le type-V, on s'attend à un effet dû à l'étroitesse de la cavité antérieure.

L'extension de la section de fermeture est spécifiée par une fonction de montée exponentielle décrite dans une étude antérieure [1].

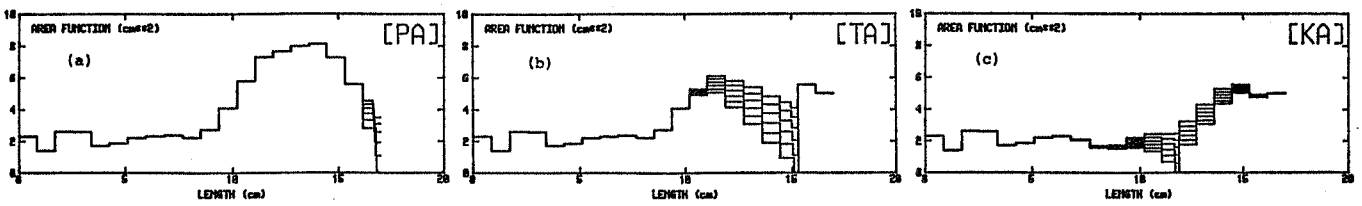


Fig.1 The transition of the area function from C to V, sampled at every 20 ms; Type-Labial at (a), Type-Dental at (b), and Type-Velar at (c).

Les cavités antérieures et postérieures s'agrandissent également de façon exponentielle, mais la montée du début est atténuée et précède le début du relâchement d'une certaine quantité. Ces spécifications rendent compte de l'observation courante qu'un geste d'ouverture non abrupte du conduit vocal précède une extension abrupte de la section de fermeture.

CALCUL DU CONDUIT VOCAL

Une expérience de synthèse et l'examen visuel des fonctions de transfert (TF) a permis d'étudier les effets acoustiques et perceptifs des trois paramètres suivants: (i) la fonction d'aire de la voyelle cible, (ii) la position de l'occlusion le long du conduit vocal, et (iii) la configuration du relâchement (type-L, type-D et type-V). Les voyelles cibles sont les trois voyelles extrêmes /i, a, et u/. La position de l'occlusion varie de la section des lèvres (1ère section) jusqu'à la 9ième section, section par section. Les 3 types de relâchement différents sont combinés avec chaque voyelle et chaque position d'occlusion.

Expérience de synthèse

Le but est de vérifier si le modèle de production des syllables CV décrit ci-dessus et défini de façon intuitive, est capable ou non de produire des consonnes reconnais /p/, /t/ ou /k/ en faisant varier systématiquement les 3 types de paramètres. Si tel est le cas, il sera alors intéressant d'étudier les propriétés acoustiques distinctives associées avec chacune de ces 3 consonnes, en terme des TF variant avec le temps.

Pour la synthèse, la vitesse initiale du relâchement est fixée à 60 cm^2/s pour le type-L et -D, et à 30 cm^2/s pour le type-V. La vitesse initiale détermine la force de l'excitation cohérente ainsi que la rapidité des variations temporelles du spectre. La source du bruit de frication est représentée par une source de pression à la sortie de la section de fermeture. Son amplitude est calculée à partir de l'aire de constriction et du débit volumique [1].

La fonction d'aire variable au cours du temps et l'aire glottale ont été introduites dans un synthétiseur du conduit vocal basé sur une simulation acoustique [5]. L'aire glottale est gardée ouverte pendant le relâchement à 0.1 cm^2. Durant la voyelle, elle est variée au cours du temps de façon appropriée pour l'excitation. Un exemple de signal synthétisé et son spectrogramme sont représentés sur la Fig.2. La fonction d'aire cible pour la voyelle /a/, une occlusion située à la 5ième section à partir des lèvres, et un relâchement de type-V ont été utilisés pour la synthèse de cette syllabe.

Une écoute informelle des 81 syllabes synthétisées indique qu'il est possible de produire les consonnes /p/, /t/ et /k/ en ajustant la position de l'occlusion pour n'importe quelle voyelle, quelque soit le type de configuration du relâchement. Seul un positionnement de l'occlusion dans certaines sections produit des consonnes facilement identifiables. Une position d'occlusion intermédiaire produit des consonnes peu claires ou ambigues. Des consonnes synthétisées avec le type-L et la voyelle ouverte /a/, cependant, sont perçues de façon moins distincte que celles qui sont synthétisées avec une configuration appropriée du type de relâchement, en particulier pour les vélaires. Bien qu'un test de perception plus systématique aurait dû être conduit, nous avons conclu que la qualité des sons était suffisante pour justifier une étude des caractéristiques acoustiques associées avec chacune des 3 catégories.

Fonctions de transfert dynamiques

A cause des changement rapides de la fonction d'aire durant le relâchement, le spectre du signal présente également des variations temporelles

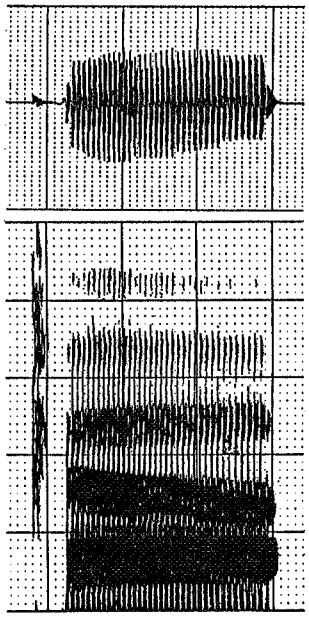


Fig.2
Synthesized signal and its spectrogram for [ka].

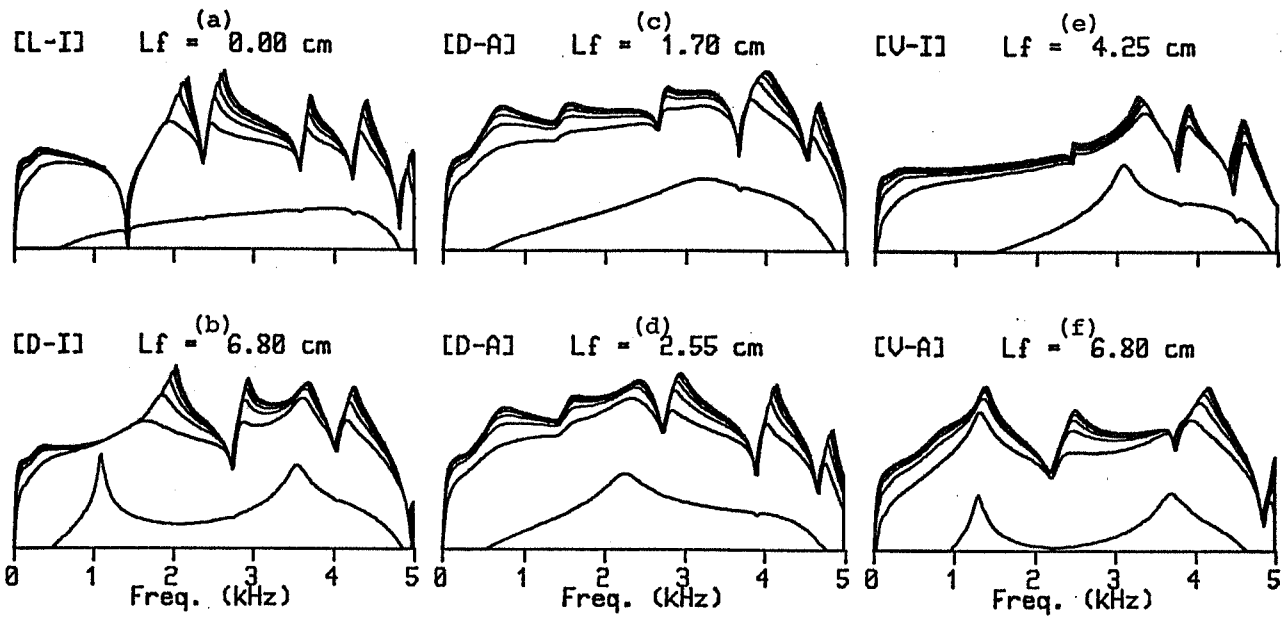


Fig.3 Calculated time-varying transfer fuction (in dB) from the onset to 5 ms after the release, sampled at every 1 ms.

rapides. Mais, une analyse de Fourier classique des bursts naturels est incapable de rendre compte de ces variations rapides du spectre, puisque ce type d'analyse contient une intégration temporelle inhérente. En revanches, ces variations rapides du spectre, en terme de TF, peuvent être calculées dans notre approche à partir de la fonction d'aire. Malgré les changements rapides de la fonction d'aire, la vitesse d'expansion de l'occlusion doit rester inférieure à 100 cm^2 /s dans des circonstances réalistes. On peut faire l'hypothèse d'une quasi-stationarité du conduit vocal. Les TF variant avec le temps peuvent en conséquence être représentés par une série de TF statiques, qui sont dérivés de fonctions d'aire statiques successives.

Dans le calcul, une fonction d'aire échantillonnée et l'aire glottale (fixée à 0.1 cm^2) sont introduites dans le même synthétiseur du conduit vocal, mais dans ce cas, le conduit vocal est excité par une impulsion de pression localisée à la sortie de la section de l'occlusion. La transformée de Fourier de la sortie calculée peut être considérée en conséquence comme la TF de pression à pression.

RESULTAT ET DISCUSSION: MANIFESTATION ACOUSTIQUE DU RELACHEMENT

Comme les variations spectrales les plus importantes se situent au début, on a examiné visuellement les TF durant les 5 premières ms. Il est intéressant de noter que, dans l'hypothèse de quasi-stationnarité du conduit vocal, les TF calculées sont indépendantes de la dynamique du relâchement, et plus spécialement de la valeur de la vitesse initiale du relâchement. Ceci implique que quoique les TF montrés ci-dessous sont calculées à des points spécifiques pour une vitesse initiale spécifique, des TF identiques doivent occurir en quelques points durant le relâchement, indépendamment de la valeur de la vitesse initiale.

L'examen des 81 séries de TF suggère que les formes de TF du début en relation avec le suivi des TF modifiées, dus à une augmentation de l'effet acoustique de la cavité postérieure, sont distinctives de chaque catégorie de consonnes, indépendamment de l'identité des voyelles cibles. La forme de la TF initale est définie de façon vague comme une forme spectrale avec peu ou sans influence de la cavité postérieure. L'influence réduite est manifestée comme les creux spectraux dans les TF.

Consonne labiale

Les TF des syllabes CV, où C est identifié comme /p/, sont caractérisés par une balance spectrale qui varie considérablement de la TF initale à la TF suivante. Fig.3(a) et (b) illustrent des exemples typiques, où sont superposés 6 coupes de TF espacées de 1 ms. Comme le niveau, en particulier le niveau des pics, augmente de façon monotone, le spectre le plus faible correspond toujours à la TF initiale (actuellement 0,1 ms après le relâchement), et le plus fort à la TF finale (5 ms après le relâchement). Dans les deux cas, la voyelle cible est /i/. Pour (a), une configuration de type-L avec une longueur de cavité antérieure (indiquée par Lf dans la figure) de 0 cm a été employée. Le cas (a) est approprié pour /p/. Pour le cas (b), le type-D avec une cavité frontale de longueur 6.80cm a été utilisé. La syllabe synthétisée correspondante a été identifiée de façon inattendue comme /pi/. Quel est l'attribut commun entre (a) et (b)? Cet attribut semble être une variabilitié considérable dans la balance du spectre. D'une part, dans le cas (a), la TF initiale présente un aspect "montant". (Notons que le large pic apparaissant juste au dessus de 4 kHz est dû en partie au filtrage

passe-bas.) D'autre part, le suivi des TF indique une augmentation rapide de l'amplitude dans les basses fréquences résultant en une balance opposée. Dans le cas (b), la TF initiale présente deux pics, l'un dans les basses fréquences, et l'autre dans les hautes fréquences, dû à une cavité frontale relativement longue. Au contraire, le suivi des TF indique une augmentation rapide de l'amplitude du pic du second formant, qui apparaît au milieu des deux pics.

On note que la propriété acoustique "descendant" proposée par Stevens et Blumstein est en accord avec cette variation de la balance spectrale. Dans leur expérience, les bursts du relâchement produits de façon naturelle étaient analysés par une transformée de Fourier à court terme, qui induit une intégration temporelle. Si le spectre variant avec le temps comme décrit précédemment était intégré, le spectre résultant aura tendance à exhiber une forme plate ou descendante.

Consonne dentale

La TF de la dentale /t/ est caractérisé par une forme spectrale globale qui est maintenue relativement constante durant le relâchement, à l'exception du niveau. Des exemples sont montrés sur la Fig.3(c) et (d). Dans les deux exemples, la voyelle cible est /a/, et le type-D est utilisé. La seule différence entre (c) et (d) est la longueur de la cavité antérieure qui est de 1.7 cm pour (c) et 2.55 cm pour (d). On observe que dans (c) un spectre montant est maintenu durant le relâchement. La version intégrée de ces TF montrera aussi une montée correspondant à la propriété invariante "montant" pour les dentales. La Fig.3(d) montre cependant un large pic au centre de l'échelle des fréquences, dû à une longueur accrue de la cavité antérieure. Comme la syllabe CV correspondant à (d) est clairement identifiée comme /ta/, la "montée" peut être considérée comme un cas particulier de la "forme spectrale globale constante". Une condition supplémentaire pour être identifiée comme dentale est l'absence de pic proéminent continu, comme noté précédemment, ce qui contraste avec les caractéristiques des vélaires décrites ci-dessous. Quand la cavité antérieure passe à 3.4 cm, résultant en une série de pics un peu plus proéminents, alors la syllabe correspondante est identifiée comme /ka/, au lieu de /ta/.

Consonne vélaire

La vélaire /k/ est caractérisée par une résonance frontale proéminente maintenue de la TF initial aux TF suivantes. La position du pic, en fréquence, est relativement fixe. Des exemples sont montrés sur la Fig.3(e) et (f). On remarque que dans (e), la résonance proéminente de la cavité antérieure un peu au dessus de 3 kHz est maintenue pendant tout le relâchement. L'identification de la consonne ne semble pas être affectée par l'apparition des pics spectraux dans les hautes fréquences. Même dans les cas où les pics des TF suivantes dominaient presque les résonances de la cavité antérieure, la consonne était identifiée comme vélaire tant que la résonance de la cavité antérieure reste à une fréquence relativement fixe. Ceci est typique des CV avec la voyelle cible /u/.

Comparons l'exemple (f) et (b). Les deux TF sont presque similaires. Mais il y a une importante différence: la position de la première résonance (inférieure) de la cavité antérieure apparaissent dans la TF initiale sur la Fig.3(f) est bien alignée avec le pic correspond au deuxième formant de la voyelle /a/ apparaissant dans les TF suivantes. Au contraire, dans le cas (b), un non-alignement peut être observé. Comme mentionné précédemment, cette syllabe est identifiée comme /pi/. Le bon alignement dans le cas (f) est dû au fait que le second formant de la voyelle /a/ est relativement bas, et que la configuration du relâchement de type-V augmente la fréquence de la première résonance de la cavité antérieure à de plus hautes fréquences.

Nos observations suggèrent que le spectre des bursts naturels peut ne pas être toujours compact. Quand le signal de relâchement est intégré dans le temps, la présence de pic(s) stable(s) peut être masquée par les pics spectraux des TF suivantes. En conséquence, il peut être difficile de distinguer, par exemple, entre les vélaires et les dentales.

CONCLUSION

Dans le cadre de cette étude, il apparaît que l'évolution du spectre au cours du relâchement d'une consonne occlusive contient une information importante sur la place d'articulation, et ce indépendamment de la voyelle qui suit. Trois types de variations spectrales pertinentes, qui peuvent constituer des propriétés distinctives des labiales, des dentales et des vélaires, ont été identifiées comme l'existence, respectivement, d'une "dynamique de la balance spectrale", d'une "forme globale constante", et d'un "pic stable proéminent".

REFERENCES

[1] Maeda,S.,"Une source d'exitation coherente dans les occlusives," 15e JEP GALF, Paris, 43-46, 1985.

[2] Stevens,K. and Blumstein,S.,"Invariant cues for place of articuration in stop consonants," J.Acoust.Soc.Am., 64(5), 1358-1368, 1978.

[3] Blumstein,S. and Stevens,K.,"Acoustic Invariance in speech production: Evidence from measurements of spectral characteristic of stop consonants," J.Acoust.Soc.Am., 66(4), 1001-1017, 1979.

[4] Kuhn,G.,"Stop consonant place perception with single-formant stimuli: Evidence for the role of the front-cavity resonance," J.Acous.Soc.Am., 65(3), 774-788, 1979.

[5] Maeda,S,"A digital simulation method of the vocal-tract system," Speech Commnication, 1, 199-229, 1982.

MODELE AUTOREGRESSIF ET SIGNAUX BRUITES - METHODE DE LA CORRELATION ETENDUE

ZERUBIA J. MENEZ J.

LASSY - UNIVERSITE DE NICE
U.A - CNRS N° 814 (Greco SARTA)

ABSTRACT

It is well known that the resolution of the autoregressive spectral estimator decreases with the signal-to-noise-ratio. An usual approach to this problem is to use a correlation subtraction method. In this paper, we studied an extended correlation matching approach presented for the first time by ATAL & JAIN [1] - The method has been tested on both synthetic and real speech. The technique developed yields satisfactory results but we have found some limits to the validation domain. Espacially, the SNR has to be limited to (30, 10 DB) because for SNR great er than 30 DB, the influence of additive noise is minor and for SNR lower than 10 DB, an important bias in frequency occurs.

INTRODUCTION

Les méthodes de modélisation AR sont très utilisées en traitement du signal vocal (codage par prédiction linéaire, synthèse de parole...) Or, dès que les données à traitersont bruitées, l'estimation des coefficients du prédicteur est biaisée et l'on observe un lissage de la DSP [2] . Ce problème est important car dans les applications , les signaux de parole sont très souvent perturbés par des bruits de différentes natures.

Dans cette communication, nous présentons brièvement une méthode développée initialement par ATAL et JAIN [1] qui apporte une nette amélioration de l'estimation des coefficients (a_i) dans le cas de données bruitées. Puis, nous exposons les résultats expérimentaux que nous avons obtenus et nous discutons du domaine de validité de la méthode proposée.

METHODE DE LA CORRELATION ETENDUE

La méthode étudiée est similaire à des méthodes déjà proposées [3], [4] : on travaille sur la matrice d'autocorrelation du signal bruité. Il s'agit d'estimer la puissance du bruit perturbateur supposé blanc gaussien-N (0, β)-et de la retrancher de la diagonale principale.

L'idée de base est de trouver les coefficients d'un filtre purement récursif d'ordre p afin qu'il y ait une parfaite adéquation jusqu'à l'ordre p+q des courbes représentant respectivement :

- la fonction d'autocorrelation du signal bruité
- la somme de la fonction d'autocorrelation de la sortie du filtre purement récursif d'ordre p et de la fonction d'autocorrelation du bruit estimé.

L'intérêt de cette méthode est d'obtenir une adéquation sur un nombre de points (p+q) supérieur à ce que l'on obtient par une méthode classique (LEVINSON à l'ordre p par exemple) résolvant les équations de YULE-WALKER.

FORMULATION MATHEMATIQUE DU PROBLEME

Soit x_n le signal utile perturbé par un bruit blanc $b_n N(0,\beta)$. Nous disposons exclusivement de l'observation bruitée y_n donnée par

$$y_n = x_n + b_n$$

En modélisant le signal non bruité par un modèle AR, il vient :

$$x_n = -\sum_{i=1}^{p} a_i \; x_{n-i} + \varepsilon_n \qquad (1)$$

soient :

- $\hat{r}_k$ = la fonction d'autocorrelation du Kème échantillon de la sortie du filtre purement récursif
- r_k = la fonction d'autocorrelation du Kème échantillon du signal vocal bruité.
- $\beta\varepsilon_k$ = la fonction d'autocorrelation du Kème échantillon du bruit blanc.
- E_k = l'erreur définie par $E_k = \hat{r}_k + \beta\varepsilon_k - r_k$

Le critère à minimiser est l'erreur quadratique $|E|^2$ donnée par :

$$|E|^2 = \sum_{k=1}^{p+q} |E_k|^2 \qquad (2)$$

La minimisation de l'équation (2) se traduit par la résolution d'un système non linéaire de p+q équations à p+1 inconnues. On montre [1] qu'une solution sous-optimale consiste :

. à minimiser l'erreur quadratique $\emptyset$ donnée par

$$\emptyset = \sum_{k=p+1}^{p+q} |E_k|^2$$

par rapport β, puissance du bruit perturbateur.

. Puis connaissant β, à minimiser σ_ε^2 puissance du résiduel, par rapport aux coefficients (a_i) du prédicteur AR. Cela revient à résoudre les équations de YULE-WALKER à l'ordre p en utilisant une matrice d'autocorrelation corrigée.

On pose : λ_i = ième valeur propre de la matrice d'autocorrelation

v_i = ième vecteur propre de la matrice d'autocorrelation.

$$c_j = [r_{p-1+j}, \ldots, r_j]^T$$

$$r = [r_1, \ldots, r_p]^T$$

$$\gamma_i^j = (r^T V_i)(c_j^T V_i)$$

$$\Gamma = \begin{bmatrix} \gamma_1^1 \ldots \gamma_p^1 \\ \gamma_1^q \ldots \gamma_p^q \end{bmatrix}$$

$$g = \left[\frac{1}{\beta-\lambda_1}, \ldots, \frac{1}{\beta-\lambda_p}\right]^T$$

$$\alpha = [r_{p+1}, \ldots, r_{p+q}]^T$$

Après les calculs [1], il vient

$$\emptyset = \alpha^T \alpha + 2 g^T \Gamma^T \alpha + g^T \Gamma^T \Gamma g \quad (3)$$

On notera que $\emptyset$ ne dépend que de β. Sa minimisation se fera donc à l'aide d'une méthode classique, par exemple la méthode du nombre d'or [5].

RESULTATS EXPERIMENTAUX

Nous avons travaillé sur des signaux synthétiques et sur des signaux réels. L'autocorrelation était calculée grâce à l'estimateur biaisé défini par

$$rk = \frac{1}{N-|k|} \sum_{i=1}^{N-|k|} y_i \, y_{i+|k|}$$

Dans tous les cas, les tests ont été effectués avec un SNR variant de 0 à 30 DB. Le bruit étant obtenu grâce à un générateur de bruit blanc gaussien fourni par une bibliothèque mathématique.

. Signaux synthétiques :

Nous avons étudié des signaux AR d'ordre 2 et d'ordre 8. Nous présentons ci-dessous les résultats obtenus avec un signal AR d'ordre 2 proposé par FRIEDLANDER [6] - SNR = 0DB, p=2, q=6 - Npt=256.

(voir également tableau en annexe).

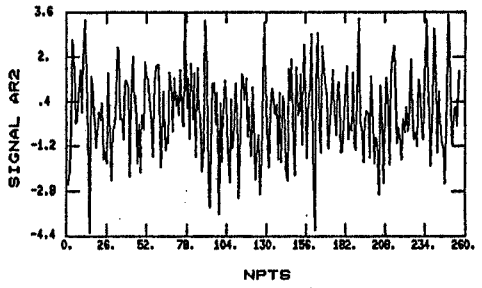


Fig. 1 : signal AR2 bruité (SNR= 0 DB)

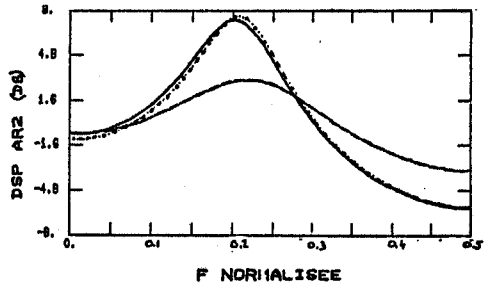


Fig. 2 : DSP du signal AR2

_____ signal non bruité ------ signal bruité

..... Signal corrigé par ATAL.

Sur les signaux synthétiques, nous avons observé que :

- la méthode de la corrélation étendue donne de bons résultats jusqu'à 0 DB pour les signaux AR d'ordre 2. Par contre, on observe l'apparition d'un léger biais en fréquence au niveau des pôles du filtre de synthèse [2] entre 15 et 10 DB. Pour les signaux AR d'ordre 8.

- Contrairement à ce que préconisaient ATAL & JAIN Nous avons trouvé que $\beta \in [0{,}2\lambda_{min}, 0{,}95\,\lambda_{min}]$ était insuffisant. Dans de nombreux cas, nous avons été obligés de prendre β de $0{,}2\lambda_{min}$ à $0{,}999\,\lambda_{min}$ pour obtenir le minimum de la fonction Φ.

. Signaux réels :

Nous avons étudié les 2 phrases suivantes: "Le Sculpteur taille la pierre" et "Un crédit est consenti par la banque" prononcées par des locutrices. Nous présentons ci-dessous les résultats obtenus pour la deuxième phrase à 15 DB avec fenêtre de HAMMING

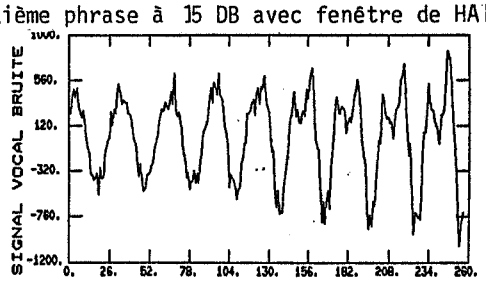
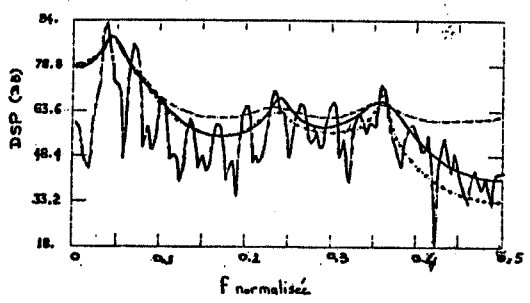


Fig. 3 : Signal vocal bruité (SNR = 15 DB, Npt=256)



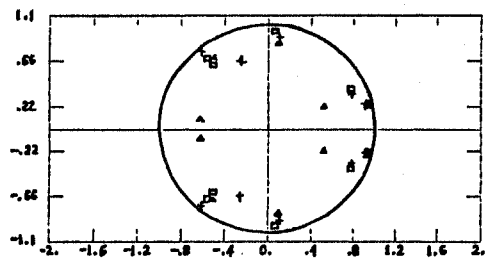
Fg. 4 : DSP signal vocal (SNR 15 DB)

— · — DSP signal non bruité (FFT)

——— DSP signal non bruité

- - - - - - DSP signal bruité

..... DSP signal bruité corrigé par ATAL. (p= 10, q=7)



Fg. 5 : Pôles signal vocal (SNR = 15 DB)

+ : signal réel non bruité

Δ : signal réel bruité

□ : signal réel bruité corrigé par ATAL. (p= 10, q=7)

Avec les signaux réels, nous avons constaté que :

- nous avions également une apparition d'un biais aux alentours de 10 DB
- La plage de variation de β était de $0,2\lambda_{min}$ à $0,99\lambda_{min}$ pour minimiser la fonction Ø
- Le choix de p et q était empirique, mais que les couples (p=8, q=5) et (p=10 , q= 7) donnaient de bons résultats dans la plupart des cas .le dernier couple confirmant la proposition d'ATAL et JAIN.

CONCLUSION

Le travail que nous avons effectué confirme les bons résultats obtenus par l'étude d'ATAL et JAIN.

Cependant, il met en évidence la nécessité de choisir un plus grand intervalle pour les variations de β ($[0,2\ \lambda_{min}\ ,\ 0,99\ \lambda_{min}]$).

Enfin, le domaine de validité de la méthode est limité par l'apparition d'un biais en fréquence non négligeable à partir de 10 DB.

BIBLIOGRAPHIE

[1] - ATAL, JAIN - "Robust LPC analysis of speech by extended correlation matching" ICASSP 85.

[2] KAY "The effects of noise on the autoregressive spectral estimator" - IEEE - ASSP 27 - Vol n° 5 - Oct 79

[3] - LE ROUX "Modèles autoregressifs et signaux bruités" GRETSI Nice 81.

[4] - VANHOVE - "Simple algorithm to improve LPC vocoder speech with noisy input" Electronics letters - vol 17 n° 4 - Feb 81.

[5] - BRENT, RICHARD - "Algorithms for minimization without derivation" . Prentice Hall Inc. Engle Wood Cliffs - 73.

[6] - FRIEDLANDER - PORAT - "On the computation of asymptotic bound for estimating autoregressive signals in white noise". Signal processing-June 85.

ANNEXE :

- Tableau récapitulatif (signal AR 2) -

	Valeurs exactes	Valeurs calculées par la méthode d'Atal
Coef. du prédicteur a_0	+ 1	+ 1
a_1	- 0,45	- 0,4483
a_2	+ 0,55	+ 0,5348
Puissance du bruit d'observation	1	1,0222
Puissance du résiduel	0,39	0,3857