GROUPEMENT DES ACOUSTICIENS DE LANGUE FRANCAISE

Groupe de la Communication Parlée

7èmes JOURNEES D' ETUDE SUR LA PAROLE

NANCY - 19-21 MAI 1976

Avec la participation
de l' IRIA, du CNRS et de l' AFCET

V O L U M E 1

Texte des Communications

Organisées à l' UNIVERSITE de NANCY 1 par

le LABORATOIRE D' ELECTRICITE ET D' AUTOMATIQUE

le LABORATOIRE D' INFORMATIQUE

GROUPEMENT DES ACOUSTICIENS DE LANGUE FRANCAISE

Groupe de la Communication Parlée

7èmes JOURNEES D' ETUDE SUR LA PAROLE

NANCY - 19 - 21 MAI 1976

Avec la participation
de l' IRIA, du CNRS et de l' AFCET

V O L U M E 1

Texte des Communications

Organisées à l' UNIVERSITE de NANCY 1 par

le LABORATOIRE
D' ELECTRICITE ET D' AUTOMATIQUE

le LABORATOIRE
D' INFORMATIQUE

I.M.P.- C.U.C.E.S

Comité Organisateur des journées

Mmes M.C. HATON
M.T. MAS

MM J. BREMONT
J.P. HATON
M. LAMOTTE
J.M. PIERREL
C. SANCHEZ
C. VIGNERON

Les septièmes journées d'Etude sur la Parole ont été organisées sous le patronage du groupement des Acousticiens de Langue Française, avec la participation de l'IRIA, du CNRS et de l'AFCET, par le Laboratoire d'Informatique et le Laboratoire d'Electricité et d'Automatique à l'Université de Nancy I, les 19, 20 et 21 Mai 1976.

Les thèmes de travail qui avaient été retenus pour cette rencontre sont les suivants :

Thème 1 : Reconnaissance automatique de la parole continue
Thème 2 : Synthèse de la parole (simulation de la source, synthèse par règles, appareillages)
Thème 3 : Analyse du signal vocal

Deux tables rondes ont également été organisées, portant sur :

a) l'aide aux handicapés
b) la conversion graphème-phonème

Les comptes-rendus de ces journées se composent de deux volumes. Le premier volume contient les communications ordinaires portant sur les trois thèmes précités. Le second volume contient le texte des conférences plénières, les exposés de synthèse sur les trois thèmes, les discussions et les comptes-rendus des tables rondes.

Nous tenons à exprimer notre gratitude à Monsieur le Président de l'Université de Nancy I, pour l'aide qu'il nous a apportée. Nous remercions également Monsieur le Directeur du Service Interuniversitaire CUCES, qui a grandement facilité le tirage des compte-rendus.

Jean-Paul HATON — Michel LAMOTTE

Co-organisateurs

The 7th French Workshop on Speech Communication was held at the University of Nancy I, on May 19, 20 and 21, 1976. It was organized by the Laboratoire d'Informatique (J.P. HATON) and the Laboratoire d'Electricité et d'Automatique (M. LAMOTTE) under the sponsorship of the Groupement des Acousticiens de Langue Française (GALF), with participation of IRIA, CNRS and AFCET.

The following topics were studied :

1) Automatic recognition of continuous speech,
2) Speech synthesis (simulation of the vocal source, synthesis by rules, devices),
3) Analysis of the speech signal.

Two panel discussions completed the program :

a) aids to handicapped
b) grapheme -to-phoneme conversion.

The proceedings of this workshop consist of two volumes. The first volume contains the contributed papers related to the three themes. The second one contains the invited papers, the review papers and the panel discussions.

We should like to express our indebtedness to Professor BOULANGE , Rector of the University of Nancy I. We also wish to thank the Director of the Service Interuniversitaire CUCES for publication of the proceedings.

Jean-Paul HATON Michel LAMOTTE

Co-organizators

RECONNAISSANCE DE LA PAROLE CONTINUE

SYNTHESE DE LA PAROLE

ANALYSE DU SIGNAL VOCAL

ANALYSE DU SIGNAL VOCAL (Suite)

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 au 21 mai 1976

ENTREE VOCALE CONTINUE EN ENSEIGNEMENT ASSISTE PAR ORDINATEUR

Camille BELLISSANT

Laboratoire d'Informatique de l'Université Scientifique et Médicale de GRENOBLE

RESUME :

Après avoir rappelé les différents types de réponses en Enseignement Assisté par Ordinateur et les méthodes de reconnaissance les mieux adaptées à chaque cas, on présente une méthode générale s'appuyant sur l'existence de "champs fixes" dans la phrase à reconnaître. Une analyse phonétique est ensuite décrite.

SUMMARY :

Following a recall of various types of answers in Computer Assisted Instruction and the best suited recognition methods in each case, a description of a general method based on the presence of "islands of confidence" within the utterance is proposed. The paper ends with a presentation of a phonetic analysis procedure.

ENTREE VOCALE CONTINUE EN ENSEIGNEMENT ASSISTE PAR ORDINATEUR

Camille BELLISSANT

Laboratoire d'Informatique de l'Université Scientifique et Médicale de GRENOBLE

I- INTRODUCTION

L'Enseignement Assisté par Ordinateur (E.A.O.) est un domaine particulier du dialogue Homme-Machine pouvant avoir des finalités différentes : l'acquisition de connaissances, la maîtrise d'une forme de raisonnement, l'illustration d'un phénomène, ... Tout l'art du rédacteur de cours consiste à organiser le dialogue entre l'élève et la machine de telle façon que le parcours soit progressif, que les différentes notions soient exposées dans un ordre logique, que la vitesse de progression soit bien adaptée au niveau de connaissance de l'élève... Ceci implique un type de conversation entre le programme et l'utilisateur permettant des échanges rapides et à contexte limité. Il est exclu, par exemple, de poser plusieurs questions imbriquées dans une même phrase, ou de faire référence à des notions non corrélées dans une question. La réponse de l'élève dans de tels cas ne serait pas exploitable pour la suite du dialogue. Il importe donc d'obtenir des réponses claires, non ambiguës. Ces réponses peuvent, en gros, être classées en deux grandes catégories : les mots ou groupes de mots isolés et les phrases complètes, ces dernières étant bien sûr les plus difficiles à analyser.
L'utilisation de la parole synthétisée en E.A.O. remonte à 1967 pour l'apprentissage des langues étrangères. Les premiers essais d'entrée vocale en Enseignement Assisté datent de 1972 {9}. Les réponses étaient toutes de type message isolé et faisaient intervenir des techniques de reconnaissance globale {2}. Il s'agit plus dans ce cas, de discrimination entre un petit nombre de messages connus que de véritable reconnaissance. Pour accepter en entrée des phrases complètes s'apparentant à des réponses libres, seules des techniques de reconnaissance analytique peuvent produire quelques résultats.

Il semble donc logique d'adapter le type de reconnaissance, globale ou analytique, au type de la réponse - donc de la question - à chaque étape du dialogue.

Un dernier point intéressant à considérer est l'usage que l'on veut faire d'une entrée vocale. Deux cas sont à envisager : ou bien la langue parlée sert à véhiculer des informations sur un sujet donné, ou bien elle est elle-même objet de l'étude. Dans le premier cas, on peut supposer que les phrases seront correctement prononcées et construites alors que dans le second cas il s'agit justement d'améliorer la construction et la prononciation défectueuses chez un élève apprenant la langue. Les problèmes abordés sont très différents et les moyens de les résoudre également.

II - METHODE GENERALE

A chaque étape du dialogue entre l'élève et le système d'E.A.O., correspond un micro-environnement relatif au domaine qui vient d'être abordé. Le vocabulaire de référence est également évolutif. La réponse à reconnaître peut être soit un message isolé (groupe nominal, valeur numérique, réponse par "oui", "non", "je ne sais pas"), soit une phrase continue. Dans le premier cas un algorithme de reconnaissance globale est plus rapide, et ceci d'autant plus qu'à chaque étape correspond un vocabulaire propre donc de taille restreinte. Dans le second cas -parole continue- il semble que l'algorithme des "champs fixes" ou "îlots de certitude" soit bien adapté à un contexte sémantique qui demeure restreint.

Cet algorithme peut être décrit en quatre phases :

-1- Recherche dans le signal de formes déterminées à l'avance (mots-clés, verbes ou constructions particulières)

-2- Elaboration à partir du ou des champs fixes ainsi déterminés, de la liste des phrases possibles en tenant compte du contexte de l'étape du dialogue

-3- Parcours du reste du message en vue d'éliminer celles des phrases proposées en -2- dont la représentation phonémique est trop éloignée de la chaîne phonétique du signal

-4- En cas de solutions multiples, tentative de paraphrasage pour lever les indécisions.

On trouvera dans {3} un algorithme de recherche d'une forme dans un signal linéaire qui peut s'appliquer à la phase -1-.
La phase -2- peut être réalisée simplement au moyen d'une grammaire dont lesrègles de production sont fonctions de l'état d'avancement du dialogue et varient à chaque étape. La représentation phonémique des phrases engendrées est directement fournie par la grammaire par l'emploi de règles phonologiques génératives {5} sans passer par une forme orthographique.
La phase -3- nécessite une analyse phonétique décrite dans le paragraphe suivant.
La phase -4- tente un paraphrasage, lorsque cela est possible, en s'appuyant sur une description sémantique du contexte de l'étape du dialogue. Cette description sémantique est un réseau associatif des notions et attributs relatifs au domaine abordé. Ce réseau est créé par le rédacteur de cours; c'est lui qui représente l'armature logique des étapes de l'apprentissage. Son existence conditionne naturellement la possibilité de réalisation de la phase -4-.

III-ANALYSE PHONETIQUE

L'appareil que nous avons réalisé {4} produit pour chaque unité de temps dont la durée est réglable de 1 ms à 10 ms, les 9 valeurs suivantes : amplitude pic-pic, nombre de passages par zéro, nombre d'extrema relatifs (changement de signe de la dérivée) pour trois bandes de fréquence. Nous avons adopté les valeurs de fréquences préconisées dans {8} et {10}. Il apparaît que ces trois bandes sont suffisantes pour une reconnaissance globale; pour une reconnaissance analytique il serait souhaitable d'augmenter leur nombre et de mesurer également le fondamental.
La mesure des passages par zéro est une extraction de paramètres souvent décriée par les équipes s'intéressant aussi bien à la synthèse qu'à la reconnaissance de la parole. Néammoins pour la reconnaissance seule, elle donne en temps réel et avec un appareillage non coûteux une bonne paramétrisation du signal de parole (voir à ce sujet {1} et {7}). Il faut d'ailleurs remarquer que lorsqu'on abaisse l'intervalle de mesure de 10 ms à 1 ms, on s'approche de l'allure effective du signal.

L'analyse phonétique que nous proposons respecte le principe général suivant : éliminer d'abord les problèmes triviaux et appliquer des algorithmes dont la complexité est proportionnée à la finesse de discrimination souhaitée. En désignant par Ai, Zi, Di les valeurs respectives de l'amplitude pic-pic, du nombre de passages par zéro du signal, du nombre de passages par zéro de la dérivée du signal pour chaque unité de temps et pour chaque bande (i = 1,2,3), on applique successivement les étapes suivantes :

A- Isoler les segments pour lesquels Ai = Zi = 0 dans les trois bandes. (silences et occlusives)

B- Isoler les segments pour lesquels A1 < ε , A2 < ε et A3 > K3 (fricatives non voisées et possibles plosives).

C- Pour ces segments distinguer dans les plans Z2/Z3 et Z3/D3 les phonèmes /s/, /f/, /ʃ/. Voir graphique n° I.

D- Isoler les segments pour lesquels A1 > K1 , A2 < ε et A3 < ε (consonnes voisées non continues).

E- Pour ces segments confirmer ou infirmer la présence du voisement dans le plan Z1/Z2 . En effet le voisement apparaît nettement par une concentration Z2 = 0 , Z1 > 0. Voir le graphique n° II montrant la différence entre /da/ et /ta/ .

F- Pour le reste du message effectuer une segmentation en fonction de la variabilité des niveaux des Ai.

G- Pour ces segments distinguer dans le plan Z3/D3 les fricatives voisées des voyelles. Pour les voyelles, proposer une classification faisant intervenir les six dimensions Z1, Z2, Z3, D1, D2, D3, prises 2 à 2 (voir par exemple le graphique n° III où /ɛ̃/ et /ã/ sont suffisamment séparés), soit dans leur ensemble pour séparer /i/ et /ü/ .

Remarque 1 : Cette méthode suppose un prétraitement pour déterminer les centres de gravité, les densités de probabilité et les contours des ellipsoïdes définis par les 6 paramètres Zi et Di. Cet apprentissage est réalisé en faisant lire à chaque nouveau locuteur un certain nombre de phrases standard. On trouvera dans {6} et {11} certaines des idées retenues pour ce prétraitement.

Remarque 2 : A la fin de l'analyse phonétique, chaque segment est caractérisé par une liste de phonèmes pondérés qui est

ensuite comparée dans la phase -3- aux différentes représentations phonémiques engendrées par la phase -2-. Les phonèmes sont classés dans cette liste par probabilité décroissante.

IV - REFERENCES

{1}- J.M. BAKER . A new time-domain analysis of human speech and other complex waveforms. Ph. D. dissertation . Carnegie-Mellon University. 1975.

{2}- C. BELLISSANT . Enseignement Assisté et Téléphone. Congrès de l'AFCET. Grenoble . 1972.

{3}- C. BELLISSANT . Un algorithme de reconnaissance de forme dans un signal linéaire. communication brève aux 6èmes journées d'étude sur la parole . Toulouse . 1975.

{4}- C. BELLISSANT, R. BOUTTAZ . Un système d'entrée vocale sur ordinateur. Séminaire de programmation. Université de Grenoble. 1975.

{5}- F. DELL . Les règles et les sons. introduction à la phonologie générative. Hermann . 1973.

{6}- A. MARTELLI . An application of heuristic search methods to edge and contour detection. C.A.C.M. vol. 19,2 . 1976.

{7}- D.R. REDDY, L.D. ERMAN, R.B. NEELY . A model and a system for machine recognition of speech . IEEE tr. on audio and electr.1973.

{8}- D.R. REDDY, P. VICENS . A procedure for segmentation of connected speech. J. Audio Eng. Soc.,vol. 16. 1968.

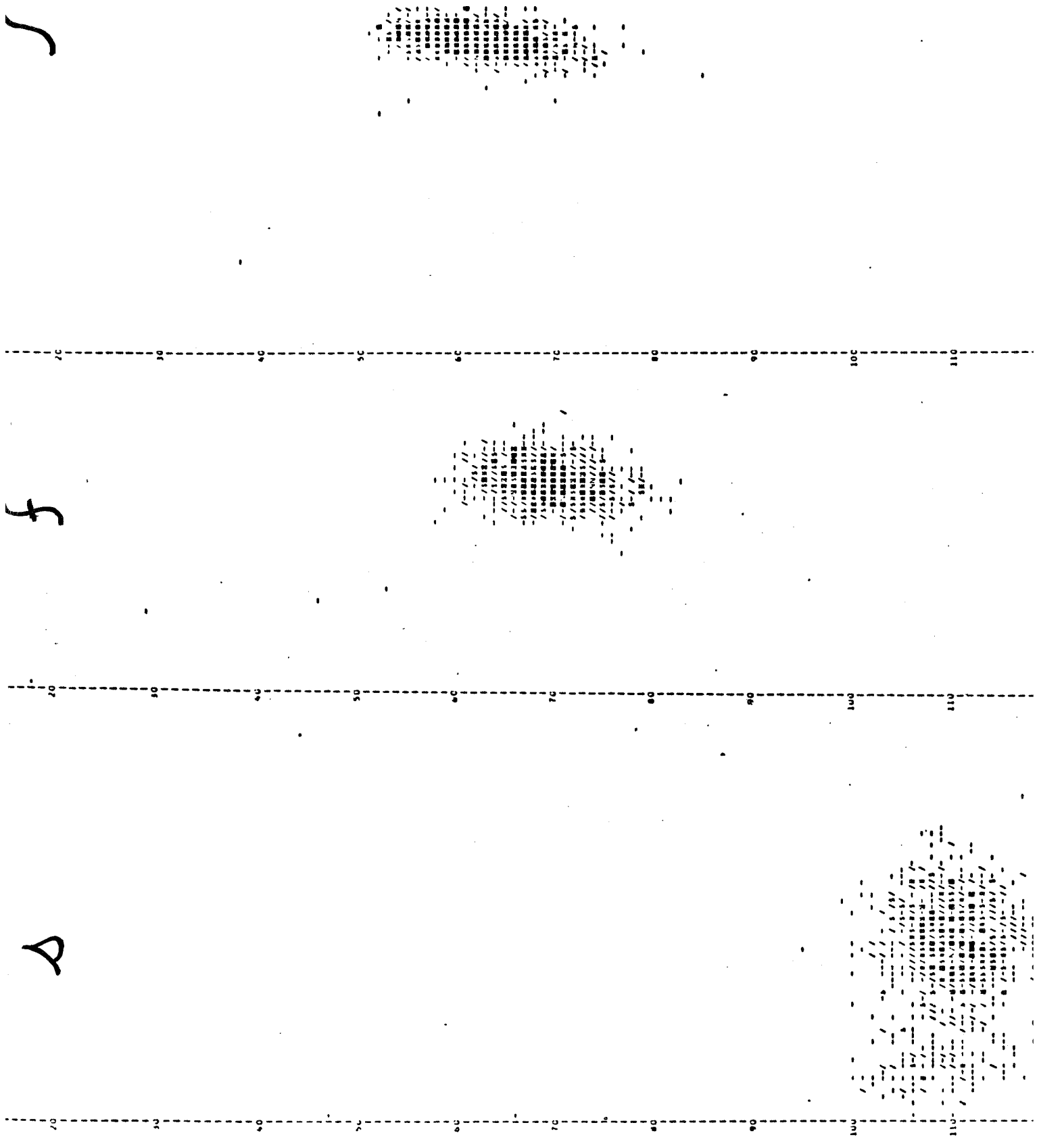
{9}- P. SUPPES, C. BELLISSANT, D.DANFORTH, J. TERHUNE . Some computer experiments on speech recognition. IMSSS. Stanford University.1972.

{10}- P. VICENS .Aspects of speech recognition by computer. Ph. D. dissertation. Stanford University. 1969.

{11}- S.S. YAU, S.C. CHANG . A direct method for clustering analysis. Pattern Recognition vol. 7. 1975.

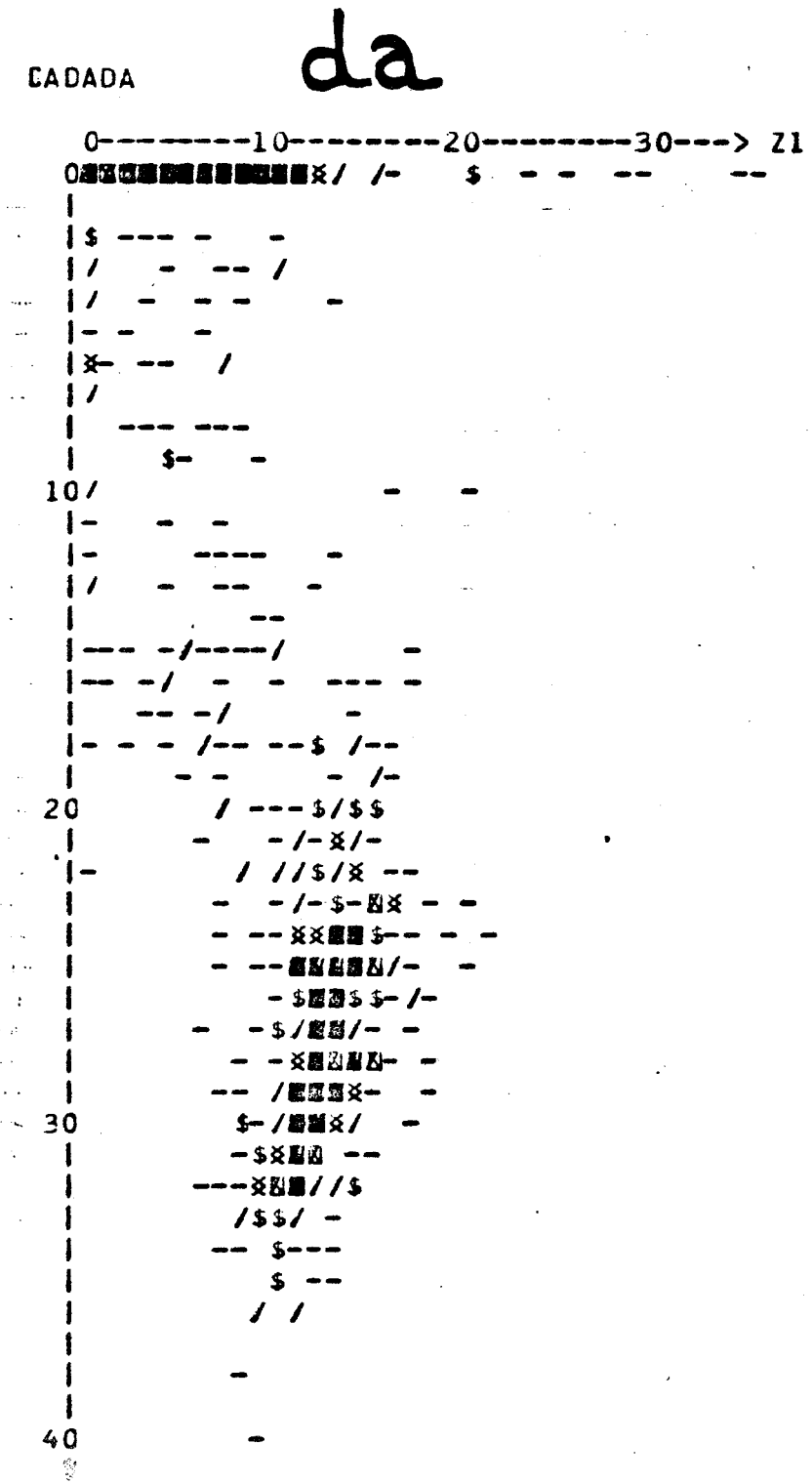
GRAPHIQUE I

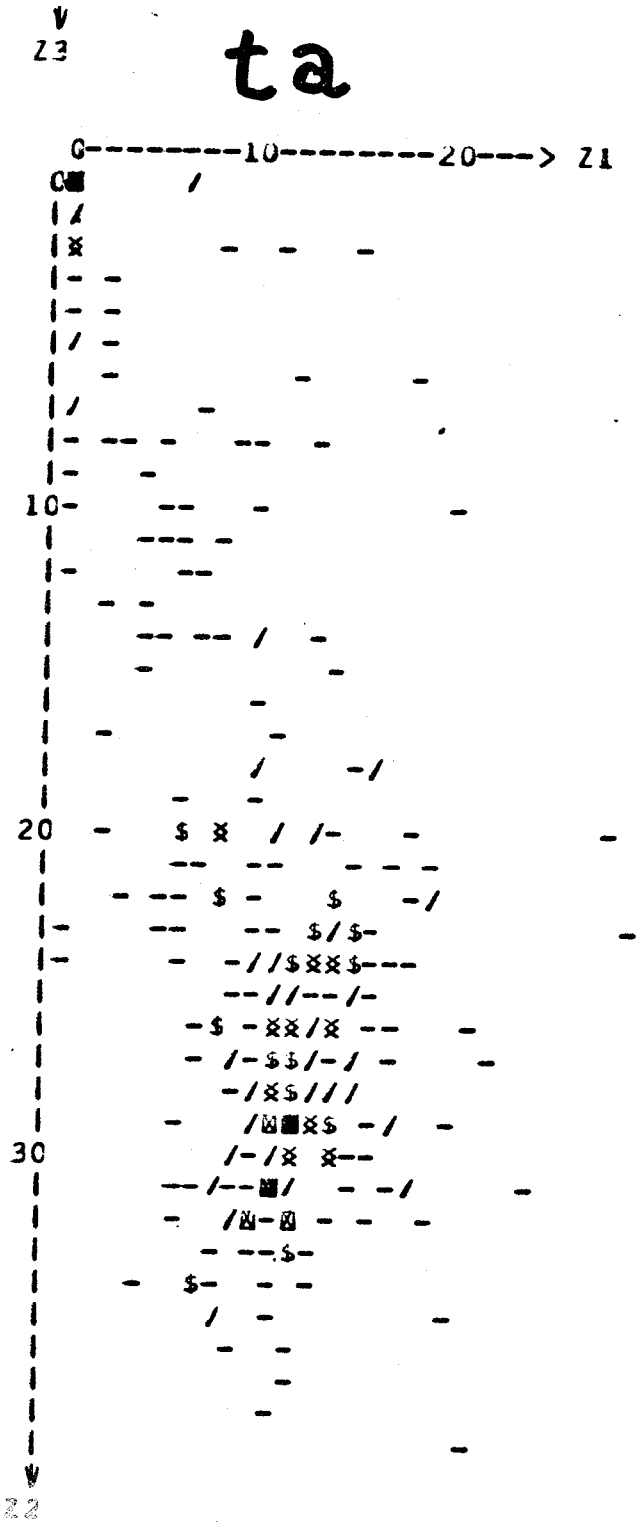
(plan Z2/Z3)



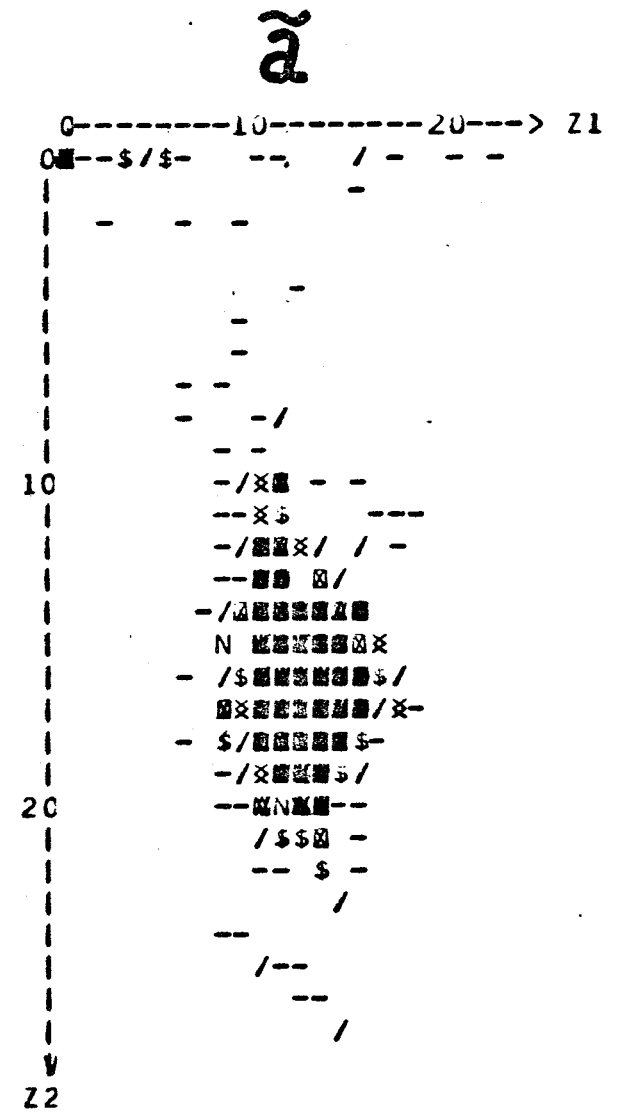
GRAPHIQUE II
(plan Z1/Z2)

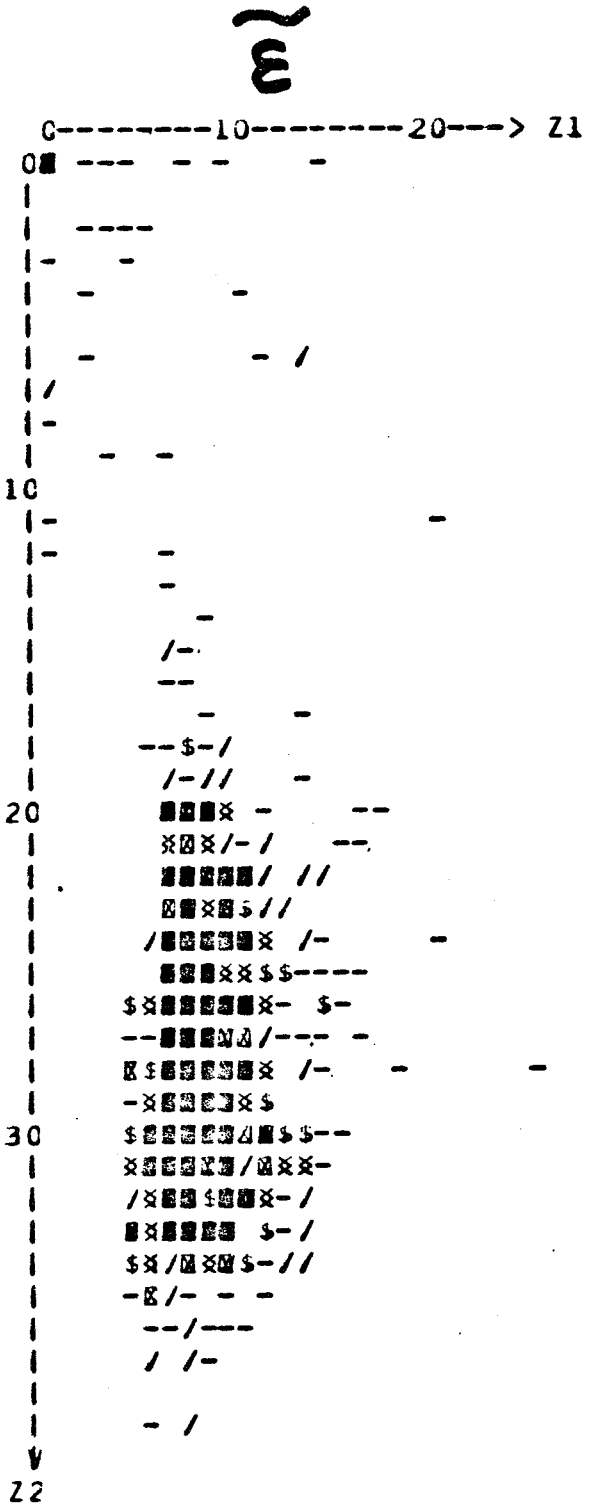
CADADA





GRAPHIQUE III
(plan Z1/Z2)





7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

APPLICATION DES METRIQUES STATISTIQUES A LA RECONNAISSANCE DES ZONES QUASI-STATIONNAIRES DU SIGNAL VOCAL

Ch.BERGER-VACHON et G.MESNARD
Laboratoire de PHYSIQUE-ELECTRONIQUE
Université de LYON I

RESUME :

Dans cet article, les auteurs envisagent la reconnaissance de la parole à l'aide d'une métrique statistique appliquée aux zones quasi-stationnaires du signal vocal.

La définition et la recherche des paramètres nécessaires à l'application de la technique statistique sont développées. C'est ainsi qu'on essaie de localiser les phonèmes à l'apprentissage, de déterminer les meilleures zones fréquentielles pour reconnaître les signaux ainsi que la position optimale des échantillons sonores à utiliser.

Les résultats obtenus par application de la métrique sont présentés sur les chiffres puis discutés dans le cadre de la reconnaissance de la parole en continu.

SUMMARY :

In this paper, the authors consider an application of statistical techniques to quasi-stationary zones embedded in vocal signals, for speech recognition.

The definition and research of necessary parameters for the application of statistical methods are developped. Localisation of phonemes at the learning stage, research of best frequency channels for recognition of vowels, most suitable time sampling are considered.

The recognition of digits is used to evaluate the quality of the method ; results are discussed according to an extension for continuous speech recognition.

APPLICATION DES METRIQUES STATISTIQUES A LA RECONNAISSANCE DES ZONES QUASI-STATIONNAIRES DU SIGNAL VOCAL

Ch.BERGER-VACHON et G.MESNARD
- LYON I -

I - INTRODUCTION

Le choix entre techniques déterministes et techniques statistiques est souvent très difficile à faire dans les problèmes de reconnaissance des formes en général et de la parole en particulier [1,2,3]. Le plus souvent, un calcul de distance est effectué ; la décision est faite en comparant les distances entre elles ou en les comparant à des seuils [4,5].

Parmi les méthodes employées, certaines sont plus directement liées à l'identification de traits phonétiques dans le langage [6,7], tandis que d'autres utilisent des prétraitements statistiques évolués avant l'application des techniques de décision [8,9,10] ; ces prétraitements statistiques peuvent être eux-mêmes précédés par des transformations du signal plus ou moins élaborées [11].

En définitive, les diverses approches en reconnaissance dérivent de la même idée de façon plus ou moins explicite : on cherche à repérer des éléments importants dans le signal et ensuite on évalue une distance avec les différents éléments d'un lexique. La décision peut être faite à l'aide d'une comparaison simple avec le dictionnaire ou en considérant une suite de symboles arrangés d'après les lois d'une grammaire. Que l'on considère la parole comme étant constituée par une suite de mots (ce qui est très artificiel) ou comme étant représentée par un ensemble de segments séparés par des silences (cas de la parole continue), les méthodes restent semblables ; les unités de base (mots ou segments) sont toujours formées par des combinaisons de phonèmes.

Dans ce qui suit, on étudie les possibilités de reconnaissance des zones quasi-stationnaires et énergétiques du signal (on se limitera aux voyelles) en utilisant des méthodes à tendance statistique ; elles sont précédées par une transformation de Fourier du signal (effectuée par un vocodeur) et par une comparaison déterministe.

En laissant de côté l'aspect syntaxico-sémantique de la parole, on se prive d'un apport important pour la reconnaissance ; mais de telles considérations pourraient masquer l'analyse de l'efficacité des méthodes que nous nous proposons d'étudier. Nous voulons répondre aux questions suivantes :

- quels sont les éléments à préciser avant de pouvoir appliquer les métriques statistiques ?
- quels sont les résultats obtenus sur un vocabulaire standard (on utilisera les chiffres) ?
- comment peut-on étendre ces résultats si on envisage la parole continue ?

II - LA DEFINITION DU SIGNAL VOCAL

Elle dépend essentiellement des méthodes utilisées. C'est ainsi qu'on s'intéresse aux voyelles et à leur caractérisation par des moyennes, des variances et des covariances.

2-1 La localisation des voyelles à l'apprentissage

La recherche des caractères des voyelles passe par une localisation des phonèmes qui n'est pas triviale "a priori".

Considérons une voyelle prononcée de façon isolée ; l'évolution de son énergie est représentée par la figure 1. On repère successivement le bruit ambiant, la transition, une zone quasi-stationnaire (plus ou moins régulière), une transition et le retour au bruit ambiant. Parallèlement, on indique une courbe donnant la stabilité. Elle est définie à partir de l'expression :

$$s_i = \sum_{j=1}^{14} \left| S_{i,j} - S_{i-1,j} \right| \qquad (1) \quad ,$$

où $S_{i,j}$ est la valeur observée sur la i^e ligne de valeurs d'un spectrogramme (i varie avec le temps), j caractérisant une bande de fréquences (j^e colonne).

Le vocodeur ETA du CNET à Lannion qui a donné nos enregistrements possède 14 canaux fréquentiels.

Nous avons décidé, en première approximation, de définir les lignes correspondant à une voyelle par $s_i \leqslant 14$; elles sont situées entre 2 limites L_1 et L_2 autour d'un point M (figure 1). Nous avons ensuite affiné cette prélocalisation en introduisant la variabilité de s_i, lorsque la ligne i est contenue dans l'intervalle $[L_1, L_2]$. Ceci conduit à une nouvelle localisation sur laquelle on étudie la stabilité: un algorithme de convergence conduit à deux valeurs l_1 et l_2 représentant, par définition, les limites de la voyelle considérée.

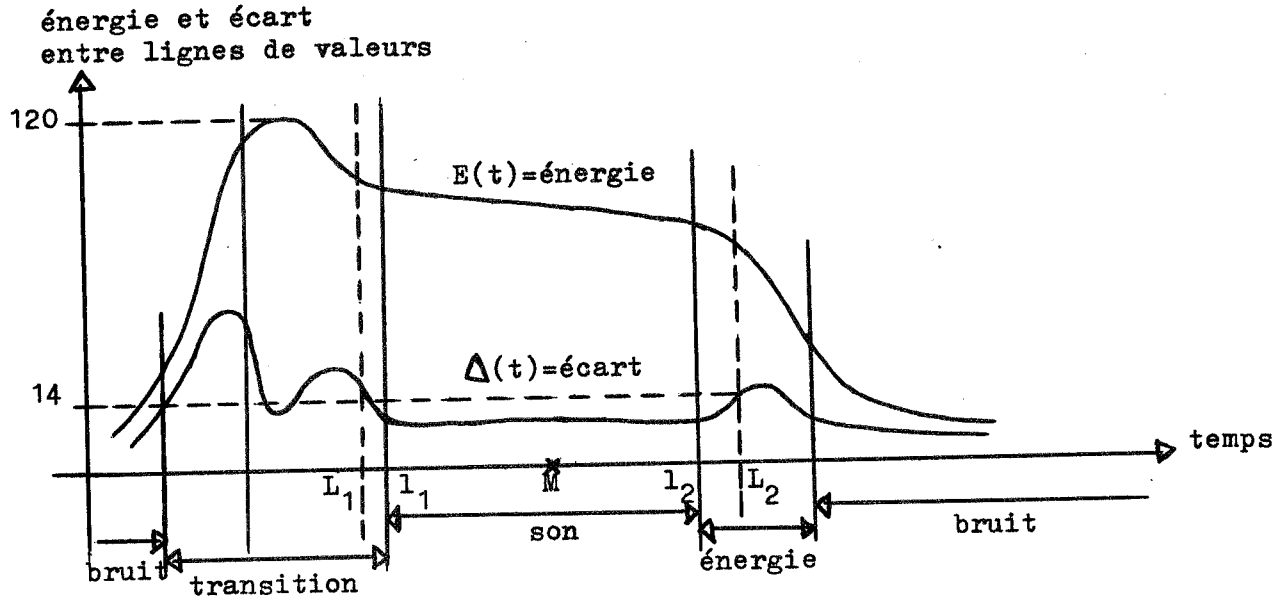


Figure 1 - Evolution de l'énergie et de la stabilité du signal en fonction du temps.

* * * *

2-2 Les bandes fréquentielles utilisées en reconnaissance

Les 14 bandes fréquentielles analysées par le vocodeur du CNET exigent une place mémoire très importante quand on souhaite effectuer une reconnaissance à l'aide de petits ordinateurs de taille industrielle. Nous nous sommes proposés de sélectionner 7 canaux sur des critères d'efficacité de séparation et de corrélation de l'information.

2-2-1 - Le calcul de la divergence

Entre les limites l_1 et l_2 (cf. § 2-1), une voyelle est caractérisée par 14 moyennes et 14 variances (cf.figure 2).

Canal	Moyenne	Variance
1 (BF)	12.00	0.93
2	12.40	1.10
3	13.00	0.86
4	11.50	1.09
5	11.50	1.12
6	9.20	0.79
7	6.50	1.05
8	5.50	1.02
9	3.60	1.04
10	7.20	1.24
11	5.30	1.89
12	3.80	2.70
13	5.50	1.40
14 (HF)	4.00	0.82

Figure 2 - Moyennes et écart-types de la voyelle /a/ calculés à partir des lignes de valeurs situées entre l_1 et l_2.

* * * *

Considérons deux voyelles, notées v_1 et v_2, et le j^e canal, C_j ; on peut tracer (cf.figure 3) les distributions de probabilité $P(X/v_1)$ et $P(X/v_2)$ de l'amplitude X observée sur C_j lorsque v_1 et v_2 sont prononcées.

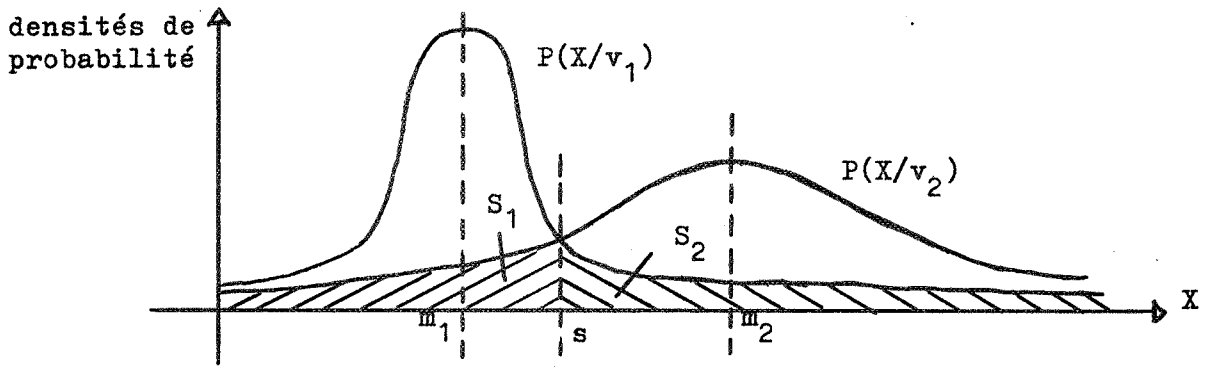


Figure 3 - Détermination de l'erreur de classification.

* * * *

Dans le cas général, les courbes $P(X/v_1)$ et $P(X/v_2)$ admettent deux intersections [12], mais pratiquement seule celle (d'abscisse s) qui est située entre m_1 et m_2 est à prendre en considération. D'après la théorie du maximum de vraisemblance [13], on décidera qu'un signal v conduisant à une valeur observée x sur C_j appartient à la classe v_1 lorsque $x < s$, et que $v \in v_2$ lorsque $x > s$; on suppose,

bien sûr, que seules les classes v_1 et v_2 sont possibles et qu'elles sont équiprobables. Ceci équivaut à décider en faveur de la classe qui est la plus probable a posteriori (lorsque le signal x a été observé).

L'erreur de classification E(s) est proportionnelle à la somme des deux surfaces S_1 et S_2 (figure 3).
On peut écrire :

$$E(s) \# \int_{-\infty}^{s} \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(x-m_2)^2}{2\sigma_2^2}} dx + \int_{s}^{\infty} \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} dx \quad (2)$$

On définit aussi la qualité :

$$Q(s) = 1 - E(s) \quad (3)$$

On voit que plus Q(s) est élevé, plus le canal C_j sera apte à séparer les voyelles v_1 et v_2.

2-2-2 - La séparation globale des voyelles

Nous allons rechercher "l'heptet" (ensemble de 7 canaux) "le plus efficace" pour séparer les 13 voyelles qui ont servi de support à cette étude. Deux notions sont retenues :

- la "puissance séparative" des canaux
- la corrélation des informations.

. la puissance séparative des canaux

Q(s) est au maximum égal à 2. Pour chacun des 14 canaux, nous avons calculé la quantité $Q_{ijk}(s)$ correspondant :

- au k^e canal
- au couple de voyelles v_i et v_j (il y a $\frac{13 \times 12}{2} = 78$ couples de voyelles possibles).

Pour chacun des canaux, l'ordinateur a compté le nombre de fois (Q'_{ijk}) où Q_{ijk} est supérieur à un seuil s'; on obtient les résultats représentés sur la figure 4.

. la corrélation entre les canaux

Nous avons calculé la corrélation entre les canaux C_i et C_j à l'aide de la formule suivante :

$$\rho_{(C_i,C_j)} = \frac{\sum_{k=1}^{13} (\overline{X_{ik}} - \overline{\overline{X_i}})(\overline{X_{jk}} - \overline{\overline{X_j}})}{\sqrt{\left[\sum_{k=1}^{13} (\overline{X_{ik}} - \overline{\overline{X_i}})^2\right]\left[\sum_{k=1}^{13} (\overline{X_{jk}} - \overline{\overline{X_j}})^2\right]}} \quad (4)$$

où :

. $\overline{X_{ik}}$ est la moyenne des valeurs X observées sur C_i lorsque la k^e voyelle v_k est prononcée et $\overline{\overline{X_i}}$ est la "moyenne des moyennes" correspondant aux 13 voyelles (moyenne par rapport à k des quantités $\overline{X_{ik}}$),

. $\overline{X_{jk}}$ et $\overline{\overline{X_j}}$ sont les valeurs correspondantes pour le canal C_j.

Il est évident que lorsque les valeurs observées sur C_i et C_j sont corrélées, les informations apportées par les deux canaux se recouvrent. Donc plus $|\rho(C_i,C_j)|$ sera faible, meilleur sera le couple (C_i,C_j). Tous les couples (C_i,C_j) ont été envisagés dans un heptet.

k \ Seuils s'.	1.98	1.95	1.90	1.85	1.80
1	1	1	3	4	5
2	6	12	12	13	20
3	6	10	15	16	23
4	9	13	19	22	29
5	6	12	19	20	25
6	13	18	21	25	28
7	14	15	23	26	30
8	16	19	23	28	35
9	1	4	5	7	14
10	2	4	9	16	19
11	0	1	2	5	6
12	0	0	1	1	4
13	0	0	0	1	1
14	0	4	15	15	17

Figure 4 - Comparaison de l'efficacité Q_{ijk} des canaux pour effectuer la séparation des voyelles avec des seuils. On indique les valeurs $Q'_{s',k}$ indiquant le nombre de fois où $Q_{ijk} > s'$.

* * * *

. la qualité d'un heptet

Soit $\mathcal{C} = (C_{\sigma_1}, C_{\sigma_2}, \ldots, C_{\sigma_7})$ un ensemble de 7 canaux où $\sigma_1, \sigma_2, \ldots, \sigma_7$ est une des C_{14}^7 combinaisons possibles des 14 premiers entiers positifs pris par groupe de 7.

Le canal C_{σ_p} est corrélé avec les six autres canaux de l'heptet. Son taux de corrélation linéaire moyen est défini par :

$$\rho_{C_{\sigma_p}} = \frac{1}{6} \sum_{\substack{q=1 \\ q \neq p}}^{7} |\rho(C_{\sigma_p}, C_{\sigma_q})| \qquad (5)$$

La qualité $V_{C_{\sigma_p}}$ de C_{σ_p} dans l'heptet est définie par :

$$V_{C_{\sigma_p}} = Q'_{\sigma_p} \left(1 - \rho_{C_{\sigma_p}}\right) \qquad (6)$$

On souhaite que les canaux effectuent des séparations ; nous avons donc retenu pour Q'_{σ_p} le nombre de séparations pouvant être effectuées avec une qualité Q_{σ_p} supérieure à 1,95. Q'_{σ_p} est donc lu dans la 3e colonne du tableau représenté sur la figure 4, ($Q'_{\sigma_p} = Q'_{1.95,\sigma_p}$).

La qualité $Q_{\mathcal{C}}$ de la combinaison $\mathcal{C}$ est donnée par :

$$Q_{\mathcal{C}} = \sum_{p=1}^{7} V_{C_p} \qquad (7)$$

Nous avons calculé $Q_{\mathcal{C}}$ pour l'ensemble des combinaisons 7 à 7 des 14 canaux. Les canaux 12 et 13 donnant Q'_{12} et Q'_{13} nuls, il n'a pas été nécessaire de les considérer ; il reste donc : $C_{12}^{7} = 792$ combinaisons à envisager.

Les valeurs $Q_{\mathcal{C}}$ sont indiquées sur la figure 5 ; la combinaison $\mathcal{C}_0 = 2 - 3 - 4 - 5 - 6 - 7 - 8$ arrive en tête. Le canal 1 qui représente le voisement (et qui est constant pour toutes les voyelles) n'est pas retenu. Quant au canal 14, il joue un rôle intéressant car il est très peu corrélé avec les autres canaux, les canaux 12 et 13 ayant été supprimés ; néanmoins, la faiblesse de Q'_{14} ($Q'_{14}=4$) ne lui permet pas d'être classé dans le meilleur heptet.

classement	heptet = $\mathcal{C}$	score = $Q_{\mathcal{C}}$
1	2 3 4 5 6 7 8	58.9
2	2 3 5 6 7 8 14	58.0
3	2 3 4 6 7 8 14	57.8
⋮	⋮	⋮
300	3 6 7 8 9 11 14	42.3
⋮	⋮	⋮
791	1 3 4 9 10 11 14	22.4
792	1 3 5 9 10 11 14	21.5

Figure 5 - Principaux éléments du classement des heptets de canaux.

* * * *

2-3 La métrique utilisée

La probabilité pour qu'un vecteur de mesures $\vec{X}$ soit issu de la prononciation d'une voyelle v_i est donnée par la formule suivante :

$$P(\vec{X}/v_i) = \frac{1}{(2\pi)^{3,5} \, |\Lambda_i|^{1/2}} \, e^{-\frac{1}{2} (\vec{X} - \vec{M_i})^T \Lambda_i^{-1} (\vec{X} - \vec{M_i})} \qquad (8)$$

$\vec{X}$ est formé par les 7 valeurs observées sur les canaux 2, 3, 4, 5, 6, 7, 8 à un instant donné I, lorsque la voyelle v_i est prononcée. $\vec{M_i}$ est le vecteur moyen et Λ_i la matrice des covariances correspondant à v_i et à l'instant I_i.

La matrice Λ_i n'est pas toujours inversible. Dans tous les cas que nous avons observés, lorsque Λ_i n'est pas inversible, c'est que v_i conduit à un résultat certain (donc non aléatoire) sur l'un des canaux de $\mathcal{C}_0$. Il suffit alors d'exclure ce canal de $\mathcal{C}_0$; on le considèrera avec les méthodes à caractère déterministe.

On en tient compte en calculant la probabilité moyenne (par voyelle et par canal) pour qu'un ensemble de vecteurs $Y = X_1, \ldots, X_{np}$ soit produit par la prononciation du segment s_p (Y correspond à un segment de parole ; il peut contenir plusieurs voyelles) :

$$P(Y/s_p) = \sqrt[n_p]{\prod_{k=1}^{n_p} \sqrt[\nu_{kp}]{P(\vec{X}_i/v_{kp})}} \qquad (9)$$

où s_p est le p^e segment de parole

n_p est le nombre de voyelles de ce segment

v_{kp} est la k^e voyelle de s_p

ν_{kp} est le nombre de canaux de $\mathcal{C}_0$ retenus pour v_{kp}.

III - APPLICATION A LA RECONNAISSANCE

La reconnaissance des segments de la parole se fait à l'aide de l'équation (9). Pour chacun des segments possibles, on calcule $P(Y/s_p)$ et on effectue la décision en faveur du segment le plus probable.

Une autre stratégie consiste à classer les segments à l'aide de la formule 9 et à reconstituer les phrases syntaxiquement possibles.

3-1 La recherche des voyelles

Précisons la détermination des vecteurs $\vec{X}_1, \ldots, \vec{X}_{n_p}$ correspondant aux voyelles du segment s_p. Tout d'abord, on définit un seuil énergétique de début. On considère par exemple le mot /ZERO/ prononcé isolément. La variation de son énergie en fonction du temps est indiquée sur la figure 6. On voit qu'il y a une montée énergétique avant la voyelle /e/, montée qui peut être repérée très facilement en considérant un seuil E_d et le "point d'ancrage" A_d qui lui correspond.

La recherche de A_d est très fiable, alors que la localisation du début du /z/ est très délicate.

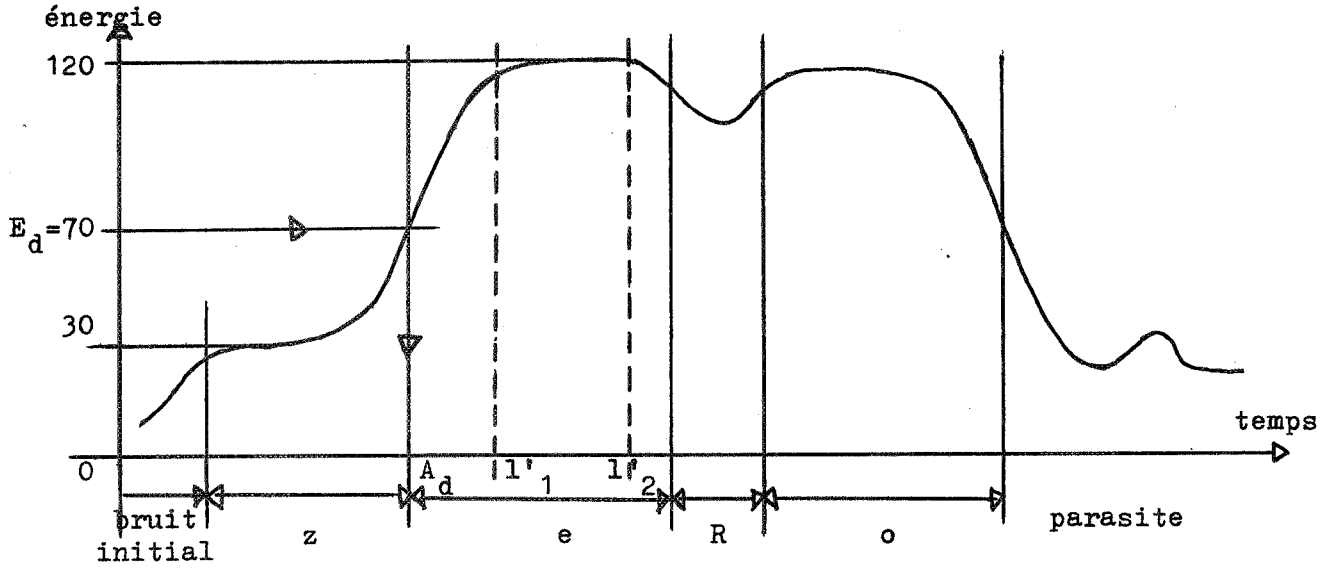


Figure 6 - Recherche du point d'ancrage correspondant à une prononciation de ZERO.

* * * *

Nous allons repérer les zones stables correspondant au /e/ et au /o/ ; on choisit, comme exemple de description, la localisation du /e/.

Dans chacun des spectrogrammes de ZERO, nous avons déterminé le début l'_1 et la fin l'_2 du son /e/ ; l'_1 et l'_2 sont repérées à partir de A_d. Cette localisation est effectuée en comparant les spectrogrammes avec les valeurs théoriques du /e/ déduites de l'étude décrite au paragraphe 2-1. On suppose que l'_1 et l'_2 sont distribuées normalement lorsqu'un ensemble de prononciations de ZERO est envisagé (cf.figure 7) :

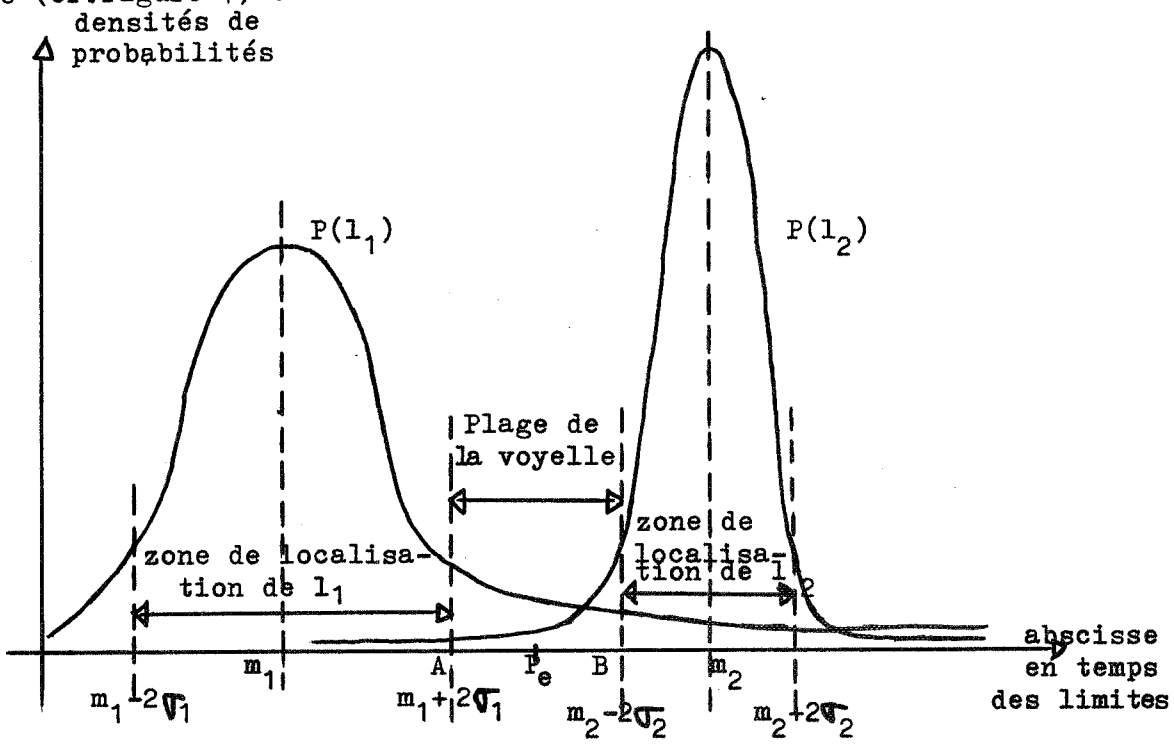


Figure 7 - Distributions des limites l'_1 et l'_2 de la voyelle /e/ de ZERO. AB est la plage de ce phonème ; dans cette zone, on a 95% de chances de trouver /e/.

* * * *

Les points A et B sont définis par :

$$\left.\begin{aligned} \int_{-\infty}^{A} P(l_1)\, dl_1 &= 0{,}95 \\ \int_{B}^{\infty} P(l_2)\, dl_2 &= 0{,}95 \end{aligned}\right\} \quad (10)$$

d'où : $A = m_1 + 2\sigma_1$ et $B = m_2 - 2\sigma_2$

Lorsque

$$A < B \quad (11)$$

il existe une zone dans laquelle nous avons 95% de probabilité de trouver la voyelle /e/ ; pour des segments s_p n'excédant pas 3 voyelles, l'inéquation 11 est pratiquement toujours vérifiée.

L'abscisse P_e représentative de la voyelle /e/ est définie par :

$$P_e = \frac{m_1 + 2\sigma_1 + m_2 - 2\sigma_2}{2} = \frac{m_1 + m_2}{2} + \sigma_1 - \sigma_2 \quad (12)$$

La moyenne M_e et les covariances Λ_e sont calculées à partir des éléments situés sur la p^e ligne de valeurs (après A_d) pour chacun des spectrogrammes de ZERO.

3-2 Application numérique

Pour préciser l'efficacité de la formule (9), on l'applique à la reconnaissance des chiffres isolés. On remarque que cette reconnaissance est proche des conditions correspondant au cas de la parole continue. En effet, si on considère une suite de chiffres prononcés sans prendre de précautions particulières, on remarque que les chiffres sont très souvent précédés par une zone de faible énergie. C'est ainsi que ZERO, DEUX, TROIS, QUATRE, CINQ, SIX. SEPT, HUIT et NEUF débutent par un phonème faiblement énergétique (/ɥ/ et /n/ ont peu d'énergie). Il suffit alors d'introduire un artifice de prononciation avant UN, pour obtenir une séparation automatique de ces éléments. La décomposition des chiffres est indiquée sur la figure 8.

Mot	Voyelle considérée	Commentaires
ZERO	/e/ et /o/	
UN	/œ/	
DEUX	/ø/	
TROIS	/a/	
QUATRE	/ɑ/	TR éliminé
CINQ	/ɛ̃/	K éliminé
SIX	/i/	ə éliminé (ə final automatique)
SEPT	/ɛ/	TE éliminé
HUIT	/i/	i plus aigu que dans SIX ; T éliminé
NEUF	/œ̃/	

Figure 8 - Décomposition des chiffres. Il est très facile d'éliminer les segments correspondant à TR, K, ə et TE.

* * * *

Chacun des éléments Y contenu entre deux zones faiblement énergétiques est successivement comparé à tous les archétypes des chiffres après avoir fait les éliminations nécessaires (cf.figure 8).

La séquence des opérations pour calculer $P(Y/s_p)$ est la suivante :

. le signal vocal est analysé par un vocodeur à canaux et on ne considère que les canaux 2, 3, 4, 5, 6, 7, 8.

. On suppose que Y est constitué par un ensemble de vecteurs $\overrightarrow{X_1}$, $\overrightarrow{X_2}$, ..., $\overrightarrow{X_k}$, ..., $\overrightarrow{X_{n_p}}$ correspondant aux n_p zones stables de s_p. Chacun des vecteurs $\overrightarrow{X_k}$ est repéré par rapport au point d'ancrage A_p de s_p

. $\overrightarrow{X_k}$ est à l'abscisse $A_p + P_{kp}$ sur le spectrogramme à identifier (cf. figure 7). Il est formé par les valeurs lues sur les ν_{kp} canaux correspondant à v_{kp} (k^e voyelle du p^e segment)

. la formule 9 est ensuite appliquée en supposant que s_p représente successivement les 10 chiffres ; la décision est effectuée en faveur de la probabilité la plus élevée. La reconnaissance est en général correcte. Il y a eu quelques erreurs qui sont indiquées sur la figure 9.

3-3 Extension à la parole continue

L'identification des zones de faible énergie dans le langage n'est pas un problème trivial. C'est ainsi que nous avons cherché à localiser les occlusions dans les chiffres et à les distinguer des remontées énergétiques qui leur succèdent. En considérant la ligne de valeurs de plus faible énergie (après avoir détecté une diminution de l'énergie en-dessous de 30) ainsi que la ligne de valeur de plus forte énergie

qui lui succède, l'occlusion a toujours été localisée (cf.figure 10).

spectrogrammes \ archétypes	0	1	2	3	4	5	6	7	8	9
ZERO	*							⊛		
UN		*								
DEUX			*							
TROIS				*						⊛
QUATRE					*			⊛		
CINQ						*				
SIX							*		⊛	
SEPT						⊛		*		
HUIT							⊛		*	
NEUF										*

Figure 9 - Matrice de reconnaissance des chiffres. Le signe * indique les décisions qui sont prises. ⊛ indique les décisions erronées. Chaque spectrogramme est comparé successivement à tous les archétypes.

* * * *

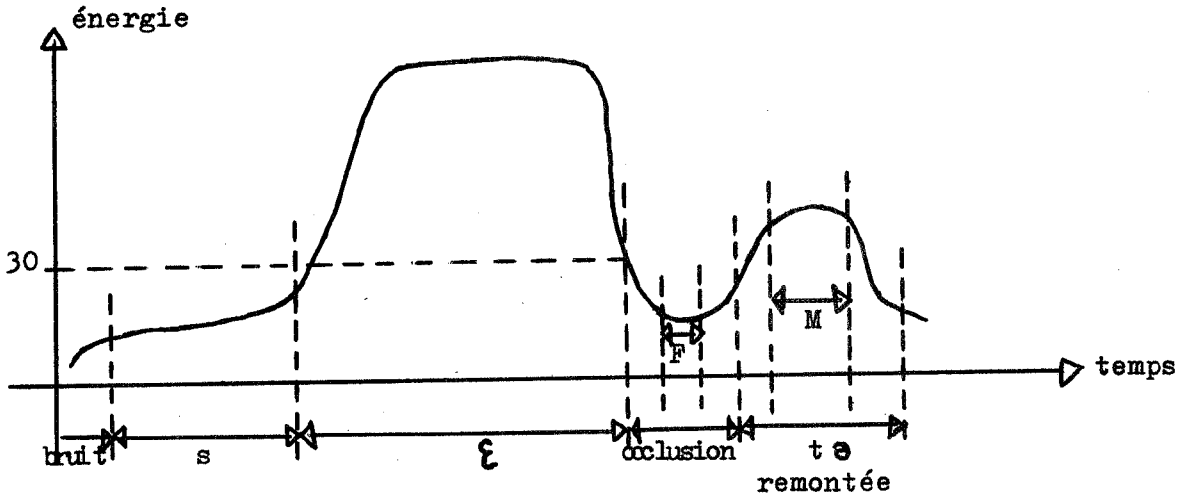


Figure 10 - Recherche de l'occlusion de SEPT
F = échantillons de faible énergie ;
M = échantillons de forte énergie.

* * * *

Si par contre, nous prenons la plus grande (s_1) des trois énergies minimales dans l'occlusion et la plus petite (s_2) des trois énergies maximales dans la remontée, (pour introduire une sécurité), il devient plus difficile de repérer l'occlusion (cf. figure 11). Néanmoins, ceci ne semble pas être un obstacle à la localisation de l'occlusion.

La comparaison systématique de tous les éléments Y (qui peuvent être

isolés entre deux zones de faible énergie) avec tous les archétypes possibles s_p nécessite l'étude des successions de phonèmes dans le langage [14] et l'introduction de techniques séquentielles pour limiter le nombre de segments s_p à envisager à chaque instant [1]. La détermination systématique du point d'ancrage A_p et la prise en considération des abscisses P_{kp} pour chaque Y et chaque s_p ne constituent pas un obstacle à l'utilisation de la méthode dans des applications exigeant le temps réel.

La considération exclusive des zones stables dans notre étude et les résultats numériques obtenus (cf.figure 9) introduisent une limitation plus sérieuse. La formule 9 doit être utilisée en complément de méthodes plus complètes conduisant à une première reconnaissance.

Enfin, les métriques que nous avons calculées permettent d'associer, pour chaque Y, un classement des segments s_p ; l'emploi de procédés syntaxico-sémantiques est capable d'extraire une phrase d'une matrice de symboles ordonnés.

chiffre	limite supérieure s_1 de l'occlusion	limite inférieure s_2 des remontées
QUATRE	29	33
CINQ	20	45
SEPT	14	27
HUIT	22	29

Figure 11 - Détermination des énergies permettant un repérage automatique des occlusions dans un mot.

* * * *

IV - CONCLUSIONS

Nous avons présenté une méthode qui a pour but, après un prétraitement, de classer les segments du langage à l'aide de considérations sur les zones quasi-stationnaires. La prise en compte d'une métrique statistique compatible avec des moyens de calculs limités exige une étude préalable approfondie du signal ; c'est ainsi que nous avons été amenés à préciser les positions fréquentielle et temporelle des éléments retenus par l'emploi de traitements statistiques. Les résultats obtenus permettent d'envisager leur traitement ultérieur par des méthodes syntaxico-sémantiques.

L'extension à la parole continue, bien que non triviale, semble tout-à-fait possible. C'est dans ce sens que s'orientent nos recherches actuellement.

BIBLIOGRAPHIE

[1] - C.BERGER-VACHON -"Conception d'une entrée vocale automatique" - Thèse d'Etat - Lyon - 1975

[2] - R.VIVES, J.Y.GRESSER - "A similary index between strings of symbols. Applications to automatic words and languages recognition". Int.Conf.on pattern recognition - 1973 - Proc. pp 308-317 - Washington.

[3] - S.CASTAN, G.PERENNOU - "Reconnaissance de formes par apprentissage". NUCLEUS - Tome 8, N°4 - 1968 - pp 329-350 - Nucléus SA Editeur, 1 rue Chalgrin (Paris 16e).

[4] - T.M.COVER, P.E.HART - "Nearest Neighbor pattern Classification". IEEE Transactions on Information Theory - Vol. IT-13, N°1 - janv. 1967 - pp 21-27.

[5] - M.E.HELLMAN - "The nearest neighbor classification rule with a reject option". IEEE Transactions on system science and cybernetics - Vol. SSC-6, N°3 - juillet 1970 - pp 179-185.

[6] - M.J.UNDERWOOD, T.R.ADDIS, D.W.BOSTON - "The evaluation of certain parameters for the automatic recognition of spoken words". Proc. Int. Conf. on machine perception of patterns and pictures - 1972 - pp 117-125 - Teddington (G.B.).

[7] - W.D.VOIERS - "Diagnostic approach to the evaluation of speech intelligibility" - JE 71 - part. B-c. (2e Journées d'Etude sur la parole - Aix en Provence - 1971).

[8] - Y.T.CHEN, K.S.FU - "On the generalized Karhunen Loeve expansion". IEEE Trans. Inform Theory - 13, N°2 - 1967 - pp 518-520.

[9] - F.P.FISHER, E.A.PATRICK - "A preprocessing algorithm for nearest neighbor decision rules". National electronics conference - 7-9 déc.1970 - Chicago (Illinois).

[10] - G.MERCIER - "Reconnaissance des formes, approximation des fonctions de décision et application à la reconnaissance des phonèmes". Thèse de spécialité (Math.appliquées) - 1969 - Rennes.

[11] - A.MURE-RAVAUD - "Reconnaissance automatique des chiffres manuscrits et des principales voyelles parlées - Emploi de la transformation de Walsh-Hadamard". Thèse de spécialité - Lyon - 1976.

[12] - C.BERGER-VACHON, G.MESNARD, J.Y.GRESSER - "Etude théorique et expérimentale des confusions données par un vocodeur - Application à la reconnaissance de la parole". Annales des Télécommunications , 30, N°5-6 - 1975 - pp 139-148.

[13] - H.INABA, K.HIRAMATSU - "Caracteristic evaluation function and decision function in pattern recognition". J.Inst.electr.Communic. Engrs Jap. [50] N°3 - 1967 - pp 118-127.

[14] - J.P.HATON - "Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole". Thèse d'Etat - 8 janv.1974 - Université de Nancy I.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

ESSAI DE RECONNAISSANCE DE LA PAROLE CONTINUE

J. BREMONT - H. LEM - M. LAMOTTE - M.-T. MAS - M.-J. VIGNERON

Laboratoire d'Electricité et d'Automatique

RESUME :

Nous présentons un programme de reconnaissance automatique d'un discours continu. Un processus de segmentation, basé sur la recherche des zones transitoires détectées par des variations énergétiques, découpe la parole en segments, ensuite reconnus par des tests phonémiques. A partir de la chaine de "caractères" ainsi obtenue et des données définissant le vocabulaire et la grammaire du langage, on reconnaît la phrase prononcée par le calcul du meilleur chemin dans le graphe flou représentant la chaine.

SUMMARY :

This paper describes a programm for automatic recognition of connected sentences. In this programm, the speech segmentation is reached through the research of the transition zones which are detected from energetical changes ; then, the segments are recognized thanks to phonemic tests. From the "character" string which is obtained and from the typical data concerning the language vocabulary and grammar, the pronounced sentence is recognized through the calculation of the best way in the fuzzy graph of the string.

ESSAI DE RECONNAISSANCE DE LA PAROLE CONTINUE

J. BREMONT - H. LEM - M. LAMOTTE - M.-T. MAS - M.-J. VIGNERON

I - INTRODUCTION.

Bien que le problème de la reconnaissance de mots isolés ne soit pas parfaitement résolu, puisque le taux de reconnaissance optimum de 100 % n'est pas encore atteint, nous avons abordé l'étude de la reconnaissance automatique d'un discours continu, dans le but de réaliser un vrai dialogue homme-machine.

II - PRINCIPE.

Le signal vocal est d'abord analysé par un banc de 32 filtres et les signaux de sortie sont échantillonnés toutes les 10 millisecondes.

Un processus de segmentation, basé sur la recherche des zones transitoires détectées par des variations énergétiques, découpe la parole en segments, ensuite reconnus par des tests phonémiques. A partir de la chaine de "caractères" ainsi obtenue et des données définissant le vocabulaire et la grammaire du langage, on reconnaît la phrase prononcée par le calcul du meilleur chemin dans le graphe flou représentant la chaine. Une telle séquence est schématisée sur la figure [1] .

Dans un souci de simplification, on choisira comme corpus un vocabulaire usuel comprenant un petit nombre de noms, pronoms personnels, articles, prépositions, verbes et adverbes et on s'imposera une grammaire simple. Il faut noter que l'entrée des données apporte de l'information, tout en créant des contraintes qui suppriment des ambiguités lors de la reconnaissance.

Pour procéder à la reconnaissance, une démarche classique [1] consiste à s'intéresser d'abord aux mots clefs de la phrase ("ilots") et à reconnaître peu à peu leur environnement, en s'aidant des règles linguistiques, grammaticales et morphologiques propres au français, de la syntaxe et de la sémantique. La méthode, adoptée ici, procède à une recherche de solutions partielles (les mots) qui sont progressivement associés en une solution globale reconstituant les phrases prononcées.

Il vient naturellement à l'esprit de choisir le mot comme cellule-pilote dans la phrase, car c'est une entité naturelle, usuelle et chargée de signification. Il est donc logique d'envisager d'abord le mot comme solution partielle et de le ranger ensuite dans une solution globale qui se construit ainsi, chemin faisant. Cette procédure est intéressante car elle présente l'avantage de réduire le temps de calcul et l'encombrement mémoire.

Pour limiter et, éventuellement corriger les erreurs, il serait souhaitable d'introduire au niveau logiciel des boucles de retour permettant, soit de modifier la solution partielle envisagée, soit de la rejeter et de relancer une nouvelle segmentation suivie d'une séquence de reconnaissance. L'influence du contexte grammatical est ainsi introduite par une telle boucle [fig(1)]. Cependant, l'introduction de ces "rétroactions" pose un problème quant à leur implantation et nuit à la stabilité du système. De plus, cela ralentit la convergence du processus et c'est un inconvénient majeur lorsqu'on s'impose, comme c'est le cas ici, de travailler en temps réel.

III - CONSTRUCTION DES SOLUTIONS PARTIELLES.

Le vocabulaire est repertorié, chaque mot étant représenté sur 16 mémoires : le premier octet de la première mémoire indique le type du mot, suivant un code convenablement choisi, le second octet repère le mot dans le vocabulaire de référence ; la deuxième mémoire renseigne sur le genre, le temps, la personne et le nombre

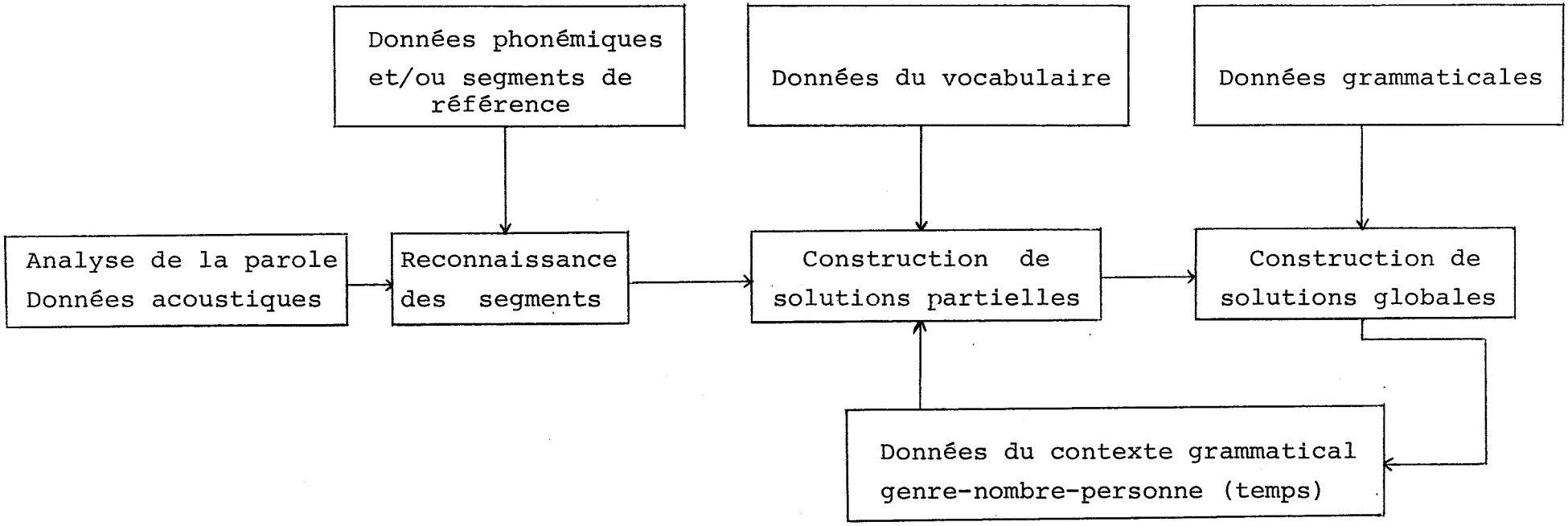

Données phonémiques et/ou segments de référence
Données du vocabulaire
Données grammaticales
Analyse de la parole Données acoustiques
Reconnaissance des segments
Construction de solutions partielles
Construction de solutions globales
Données du contexte grammatical genre-nombre-personne (temps)

Figure 1.

(c'est la "caractéristique") ; les mémoires suivantes sont consacrées aux différents caractères et permettent, grâce à un code approprié, de repérer la fin du radical du mot et aussi celle de la désinence possible, les dernières mémoires contiennent le libellé correctement orthographié.

Ce système présente un défaut évident : la rigidité du format, ce qui entraine quelquefois un "gaspillage" par suite de places vides, mais il offre l'avantage de la simplicité.

Pour construire les solutions partielles, on part de la chaine de caractères obtenue après segmentation et tests phonémiques. Le processus de reconnaissance permet de classer, par segment, trois solutions possibles affectées d'un score qui est l'indice de similarité μ tel que l'a défini Brémont [2].

Ce type d'évaluation se fait sur le plan acoustique et on peut la qualifier de "phonémique". On pourrait envisager, à ce niveau, des contraintes linguistiques (par exemple, probabilités d'existence des différents diphonèmes dans la langue française).

De plus, il serait souhaitable d'améliorer la qualité des résultats en incluant dans le programme des boucles de retour permettant de corriger, par exemple, des omissions dues à une mauvaise segmentation.

Le choix du premier caractère s'appuie sur l'indice de similarité qui lui est associé. Lorsqu'on passe d'un caractère A pris en compte à un caractère B_i possible ($i = 1,2,3$), on tente d'évaluer un degré de confiance dans la construction en cours, en définissant un score S de la façon suivante :

$S = \mu_A + \mu_B - \mu_A\mu_B$ si μ_A et μ_B sont tous deux positi

$S = \mu_A + \mu_B + \mu_A\mu_B$ si μ_A et μ_B sont tous deux négati

$S = \mu_A + \mu_B$ si μ_A et μ_B sont de signes contraires.

Notons que μ_A (ou μ_B) est compté négativement si le caractère envisagé n'appartient pas au triplet étudié (les expériences ont montré qu'il était bon de le fixer à -0,8).

Afin d'illustrer la démarche suivie, raisonnons sur un exemple. Supposons que les deux mots "LACET" et "LABOUR" figurent dans le vocabulaire des noms choisis et qu'on ait effectivement prononcé le second.

La reconnaissance par segments conduit à un tableau dans lequel figurent les segments reconnus ordonnés et affectés de leur indice μ.

Reconnaissance	$L_{0,7}$	$O_{0,6}$	$A_{0,6}$	$U_{0,5}$	$OU_{0,7}$	$R_{0,6}$	$D_{0,8}$	$U_{0,7}$
	$M_{0,6}$	$A_{0,4}$	$O_{0,5}$	$OU_{0,5}$	U	$L_{0,4}$	$P_{0,3}$	$I_{0,6}$
	$N_{0,5}$	$E_{0,2}$	$OU_{0,4}$	$O_{0,3}$	$E_{0,4}$	$M_{0,2}$	$R_{0,2}$	$E_{0,4}$
Etapes	1	2	3	4	5	6	7	8

La démarche suivie apparaît sur la figure 3.

Le problème à résoudre est de rechercher dans la séquence de segments reconnus la séquence réelle d'un mot du vocabulaire, compte-tenu du fait qu'il peut se produire des répétitions ou progression d'un caractère ou omission.

Le premier caractère reconnu étant pris comme point de départ, il faut s'assurer que l'un des caractères recherchés ensuite (par ex. A, B, OU) se trouve effectivement dans la colonne des segments reconnus, correspondant à cette étape. S'il en existe plusieurs, on choisira celui dont l'indice de similarité est le plus élevé ; ce sera le dernier caractère pris, pour l'étape suivante et on envisagera l'éventualité de sa répétition.

Etape		1	2	3	4	5	6	7	8
Caractères reconnus		LMN	OAE	A.O.OU	U.OU.O	OU.U.E	RLM	DPR	UIE
1ère S.P.	R		L	$A_{0,6}$	A	OU	OU	R	R
	A	L	A	B	B	R	R		
	O	A	B	$OU_{0,4}$	OU				
	dernier caractère pris à l'étape n-1	début	L	A	A	OU	OU	R	R
	évaluation du segment	0,7 (L)	0,4	0,6>0,4	0,4	0,5	0,6	0,2	-0,8
	évaluation de la S.P.	0,7	0,82	0,93	0,96	0,98	0,99	0,99	0,19
2ème S.P.	R		L	A	A	A			
	A	L	A	C	C	C			
	O	A	C	ET	ET	ET			
	dernier caractère pris à l'étape n-1	début	L	A	A	A			
	évaluation du segment	0,7	0,4	0,6	-0,8	-0,8			
	évaluation de la S.P.	0,7	0,82	0,93	0,13	-0,67 Rejet			

Figure 3.

L'indice de similarité correspondant est utilisé pour établir le score de la solution partielle entamée.

Dans le cas où aucun des caractères A, B, OU ne figure ici, on affecte à l'indice μ_i une valeur négative et on recommence le processus. Lorsqu'on a passé le dernier caractère du mot, le score a généralement atteint sa valeur maximale et chute de façon importante. On peut alors introduire ce mot dans la solution globale. Si, au cours de la construction partielle, l'évaluation devient très négative (inférieure à un seuil dit de "rejet"), le mot envisagé est abandonné.

IV - CONSTRUCTION DES SOLUTIONS GLOBALES.

Lorsqu'une solution partielle est atteinte (détection de la fin du radical ou de la désinence éventuelle et chûte de l'évaluation indiquant un changement de mot), on ajoute cette solution partielle à la solution globale qui la réclame. En même temps, l'évaluation de la solution partielle est cumulée à l'évaluation de la solution globale et le nombre des solutions partielles demandées par cette solution globale est diminué de 1.

Chaque solution globale (phrase) répond à des règles grammaticales strictes et simples. Ainsi, la production d'une phrase s'exprime en fonction des groupes :

G0 : début
G1 : groupe nominal sujet
G2 : groupe prépositionnel
G3 : groupe verbial
G4 : groupe pronom personnel
G5 : groupe nominal c.o.d

Chaque groupe se réécrit à l'aide de mots de type différent.

Ainsi : $Gl ::= article + \{adjectif\}_n + nom + \{adjectifs\}_m$

D'après la solution partielle qui vient d'être ajoutée à la solution globale, les règles grammaticales indiquent le type de mots qui devront être recherchés comme solution partielle suivante. Ces mots sont alors admis dans le tableau des solutions partielles avec, éventuellement, une modification de leur caractéristique : par exemple, derrière l'article "la", on attend un adjectif féminin singulier; derrière un groupe sujet au pluriel, un verbe à la 3ème personne du pluriel.

Pour le calcul de l'évaluation, on tient compte de l'évaluation de la solution partielle ajoutée et de l'accord ou non entre la caractéristique demandée et celle détectée sur le mot. A ce niveau, nous envisageons d'introduire aussi l'aspect sémantique dans le calcul de l'évaluation d'une solution globale.

BIBLIOGRAPHIE.

[1] J.-P. PIERREL, J.-P. HATON : une approche syntaxique de reconnaissance de phrases dans un contexte donné.
6èmes Journées d'Etude sur la Parole, Toulouse
(28-30 Mai 1975).

[2] J. BREMONT : contribution à la reconnaissance automatique de la parole par les sous-ensembles flous.
Thèse de Doctorat ès-Sciences
(Nancy, avril 1975).

A N N E X E

I - POUR UNE SOLUTION PARTIELLE (7 mémoires), on indique :

- l'adresse de départ dans le vocabulaire,
- le type de mot et son numéro,
- la caractéristique : genre, temps, personne, nombre (sur 2 bits pour chacun),
- le numéro de la solution globale à laquelle se rattache la solution partielle,
- le dernier caractère pris en compte dans le mot,
- le caractère de la désinence (éventuellement),
- l'évaluation courante,
- l'évaluation maximum.

II - POUR UNE SOLUTION GLOBALE (32 mémoires), on indique :

- le nombre de solutions partielles trouvées,
- le nombre de solutions partielles qui sont recherchées à l'instant considéré,
- le groupe en cours (0 à 5),
- les solutions partielles retenues (2 mémoires désignant le type, le numéro, la caractéristique).

III - TYPE DE MOTS.

1 : article
2 : préposition
3 : adverbe
4 : pronom personnel
5 : adjectif
6 : nom
7 : verbe

IV - GROUPES GRAMMATICAUX.

0 = début
1 = sujet = article + {adjectif} + nom + {adjectif}
2 = prépositionnel = préposition + article + {adjectif} + nom + + {adjectif}
3 = verbial = verbe + {adverbe}
4 = pronom personnel
5 = c.o.d = article + {adjectif} + nom + {adjectif}

V - ENCHAINEMENTS POSSIBLES DANS UNE PHRASE.

Etat actuel		Solutions partielles recherchées (état futur)	
Groupe	Type	Groupe	Type
0	1	1	5-6
	2	2	1
	4	4	7
1	2	2	1
	5	1	2-5-6-7
	6	1	2-5-7
	7	3	1-2-3
2	1	2	5-6
	4	4	7
	5	2	2-5-6-7
	6	2	2-4-5-7
	7	3	1-2-3
3	1	5	5-6
	2	2	1
	3	3	1-2-3
4	7	3	1-2-3
5	2	2	1
	5	5	2-5-6
	6	5	2-5

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

EVALUATION D'UNE METHODE ASCENDANTE D'ANALYSE LEXICALE DANS LE DISCOURS CONTINU

B. CAUSSE - D. DOURS - R. FACCA - G. PERENNOU

RESUME :

On propose une méthode d'analyse lexicale pour la parole continue qui soit seulement aidée par la syntaxe.

Elle présente les particularités suivantes :

- c'est une méthode d'analyse ascendante.
- le lexique et les chaines phonétiques à analyser sont représentées syllabiquement.
- les élisions, liaisons et enchainements sont autorisés.
- des marqueurs prosodiques et phonologiques sont pris en compte.
- des filtres syntaxiques portant sur des syntagmes très courts permettent d'éliminer certaines suites de mots inacceptables.

SUMMARY :

We propose a method of lexical analysis applied to connected speech, in which syntax is used only as a help.

It presents the following features :

- It is a bottom to top analysis method.
- The lexicon and the phonemic chains to be analysed are represented syllabically.
- Connected speech is allowed without restriction.
- Some prosodic and phonological markers are taken into account.
- Syntactic filters screening very short phrases allow for the elimination of certain innacceptable sequences of words.

EVALUATION D'UNE METHODE ASCENDANTE D'ANALYSE LEXICALE DANS LE DISCOURS CONTINU

B.CAUSSE - D.DOURS - R.FACCA - G.PERENNOU

1. INTRODUCTION

Devant les difficultées soulevées par la reconnaissance de la parole aux niveaux élémentaires (phonémique et lexical), les tentatives d'introduction de contraintes lexicales et syntaxiques, en vue d'y restreindre les choix, a soulevé un intérêt considérable. Pour s'en tenir au problème du Français, cela fut particulièrement clair au niveau des 3ème, 4ème et 5ème Journées d'Etudes sur la Parole.

Ainsi s'élaborèrent des systèmes où le langage à reconnaître était régit par une syntaxe restreinte portant sur un petit vocabulaire.

Ce qui les caractérise c'est que l'analyse est guidée par une syntaxe. De plus le choix se porte volontiers sur des méthodes descendantes partant de la phrase pour descendre au niveau des phonèmes.

En dépit de l'intérêt évident des résultats obtenus dans cette voie, quelques objections méritent d'être prises en considération. En voici deux :

(i) Ces systèmes supposent que soit définie une syntaxe de manière exhaustive : est-ce le cas pour le langage oral ?

(ii) Puisque des suites lexicales, purement grammaticales (à des degrés divers) peuvent être sémantiquement acceptables (en particulier non ambigües) et, qu'inversement, des suites grammaticales peuvent être sémantiquement inacceptables, est-il naturel de gouverner l'analyse lexicale par la syntaxe ?

L'analyse acoustique tend à se perfectionner. Nous pensons que les performances de reconnaissance sont maintenant suffisantes pour que l'analyse lexicale ne soit plus qu'aidée par des filtres syntaxiques

et sémantiques simples et performants, filtres ne représentant par exemple que des interdits très évidents, mais non une syntaxe ou un système sémantique. L'intérêt porté à une structure intermédiaire : la syllabe, croît à juste titre. Son rôle dans la reconnaissance est indéniable et certain (voir [1] par exemple).

Certaines règles de transformation phonétique n'ont de sens que par rapport à elle. Aussi est-il intéressant d'essayer de surmonter les difficultés liées au fait que la frontière d'un mot n'est pas généralement une frontière de syllabes, pour l'introduire dans l'analyse lexicale.

Enfin, les nombreuses observations faites sur la prosodie montrent que des syntagmes, relativement courts, peuvent être isolés (3è, 4è, 5è, 6è J.E.P.). Cette remarque ne fait que renforcer l'idée de l'inutilité d'une syntaxe exhaustive au niveau de l'analyse lexicale.

Ce qui suit, est une tentative pour mettre en oeuvre, dans un système exactement formulé, les remarques précédentes. Ce travail fait suite à notre communication des 6ème J.E.P. [4]. Les auteurs ont collaboré à l'ensemble de l'étude et les responsabilités ont été réparties comme suit :

Syllabation : D.DOURS
Analyse lexicale : R.FACCA - G.PERENNOU
Lexique : B.CAUSSE
Coordination : G.PERENNOU

Ajoutons que cette étude s'insère dans le projet ARIA (*).

(*) Analyse et Recherche de l'Information Acoustique.

2. SYLLABATION

La syllabation joue en rôle important dans le système de reconnaissance de la parole continue que nous envisageons. Elle permet en effet de constituer des noyaux à partir desquels s'articule la recherche des mots dans un lexique, de plus certaines règles phonologiques n'ont de sens que dans le cadre de la syllabe.

2.1. A propos de la syllabe phonétique :

Certains phonéticiens ont nié l'existence de la syllabe phonétique, elle ne serait pour eux, qu'une réalité psychologique. Nous ne les suivrons pas dans cette voie et retiendrons plutôt le point de vue de Saussure : "La limite de la syllabe se trouve là où l'on passe d'une implosion à une explosion" et plus particulièrement celui de Jakobson : "Du point de vue acoustique, on distingue dans la syllabe une partie culminante formée par une voyelle suivie d'une partie non culminante". Il est évident que cette notion de syllabation est en relation avec le degré d'aperture, ce qui se traduit au niveau acoustique par des variations d'énergie. Ceci nous conduit tout naturellement à envisager une classification des phonèmes par niveau énergétique.

2.2. Définition des classes :

Nous avons défini quatre classes de phonèmes, en fonction de leur niveau énergétique.

- La première classe, codée 0, comprend les silences et les plosives sourdes et sonores.
- La deuxième, codée 1, se compose de toutes les consonnes autres que les plosives et les liquides.
- La troisième, codée 2, comprend les semi-voyelles et les liquides. Pour les liquides une règle supplémentaire est à envisager : si deux liquides se suivent, la première est codée 2 et la seconde 1.
- La quatrième, codée 3, comprend toutes les voyelles orales et nasales.

2.3. Algorithme de découpage en syllabes :

C'est un algorithme très simple fondé sur la définition de Jakobson. Il consiste à détecter successivement une augmentation d'énergie suivie d'une baisse d'énergie et à prendre une décision pour le découpage.

En ce qui concerne le découpage, il faut bien remarquer que la syllabation obéit davantage à des habitudes qu'à des règles. C'est pourquoi nous avons cherché des lois de découpage conformes aux principes de la syllabation traditionnelle. Ces principes sont résumés dans les exemples suivants :

mot phonétique	mot codé
- ab|di|ke	030|03|030
- kak|tys	030|0310
- ab|sɛ	030|130
- par|ljɔ̃	032|1230
- par|ti|ky|lje	032|03|03|2230
- stri|jyr	01023|2320
- stryk|ty|ral	01230|03|2320
- rɛs|trik|sjɔ̃	0231|0230|1230
- røs|tryk|tu|re	0231|0230|03|230

2.3. Conclusion :

L'algorithme mis en oeuvre ne nécessite qu'une reconnaissance partielle des phonèmes. En effet seule une connaissance des classes est nécessaire. Ceci est intéressant car si les confusions entre phonèmes sont possibles à l'intérieur d'une même classe, elles sont par contre beaucoup moins probables d'une classe à l'autre et de ce fait la syllabation n'est pas perturbée par les confusions engendrées par le système de reconnaissance phonétique.

3. UN ALGORITHME D'ANALYSE LEXICALE

3.1. Principe de l'algorithme :

Le modèle génératif envisagé est le suivant :
Phrase → suite syntagmatique → suite phonétique → suite phonétique transformée (élision, liaison, enchaînement, etc...) → suite syllabique.

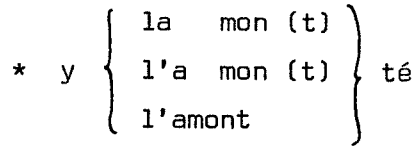
EX1 : /il a monté/l'escalier/quatre à quatre/.
→ /i la m ɔ̃ te/lɛs ka lje/ka tra katr/

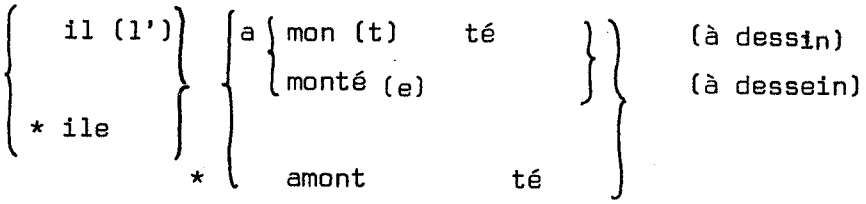
EX2 : /les petits oiseaux/chantent/ →
/le(z) pəti(z) wazo(z)/ʃ̃ãtə(t)/.
→ /le p'ti zwa zo/ʃ̃ãt/.

Le modèle d'analyse décrit ci-après constitue l'inverse du

modèle précédent. Il tient compte des liaisons, des enchaînements, des élisions, des substitutions et des commutations (erreurs éventuelles du système de reconnaissance) qui apparaissent dans une chaîne phonétique en parole continue.

Exemple : i la mɔ̃te →





* désignant une élimination par filtre syntaxique.

3.2. Description de l'algorithme :

On part d'une chaîne de syllabes phonétiques SYL fournie par l'organe de reconnaissance dont la partie syllabation a été décrite au paragraphe précédent. L'algorithme s'organise autour de quatre organes principaux :

- une mémoire LIST appelée à contenir des listes de mots "factorisant" le début de la chaîne syllabique.
- Une table DEB contenant des suites de syllabes en cours d'examen. Chacune d'entre elles pointant vers un ensemble de listes de mots chaînées entre elles dans la mémoire LIST.
- Un lexique LEX où les mots sont représentés syllabiquement.
- Un organe de décision DEC qui compare les suites de la table DEB aux mots du lexique LEX ; DEC fait aussi des transformations sur les suites de DEB ; enfin DEC tient à jour les listes de mots de LIST en fonction des résultats de ses comparaisons.

Examinons en détail chacun de ces organes :

1°/ Structure de la mémoire LIST

Elle est décrite dans la fig. 1

2°/ Structure de la table DEB

Elle est décrite dans la fig. 2

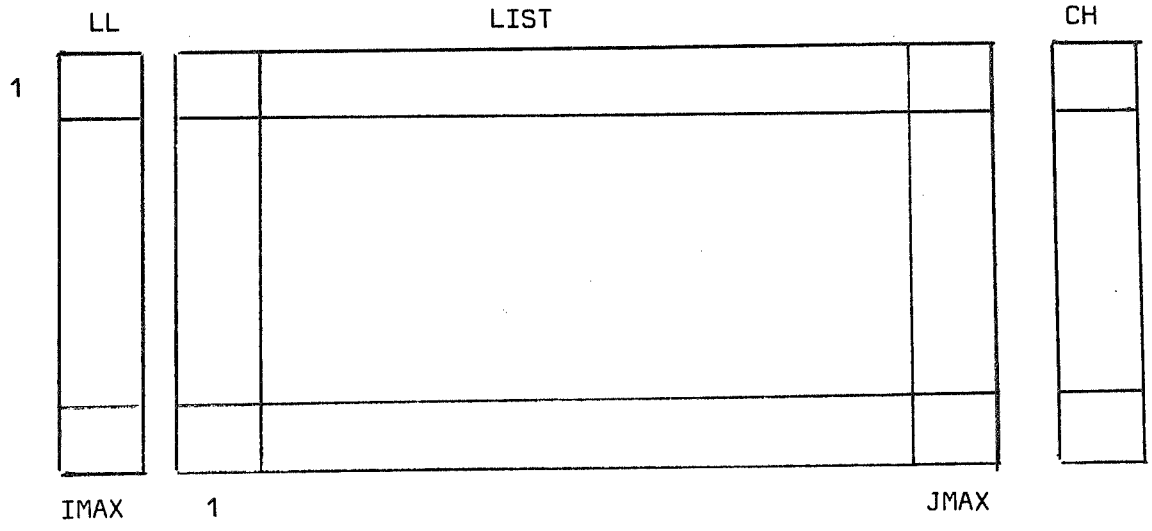
3°/ Structure du lexique LEX

Le lexique se présente comme une table de mots représentés syllabiquement et munis de divers attributs : catégorie syntaxique, genre, nombre, personne, liaisons, élisions finales, etc...

Pour les verbes réguliers seule la racine est représentée, un pointeur renvoie vers la liste des désinences possibles.

On peut accéder au lexique au moyen de la donnée d'une ou plusieurs syllabes, en réponse le lexique fournit les renseignements suivants :

La suite de syllabes
- est seulement un début de mot.
- est seulement un mot. Dans ce cas la liste des attributs est aussi fournie.
- est un début de mot et un mot dont on fournit encore les attributs
- n'est ni un début de mot ni un mot.



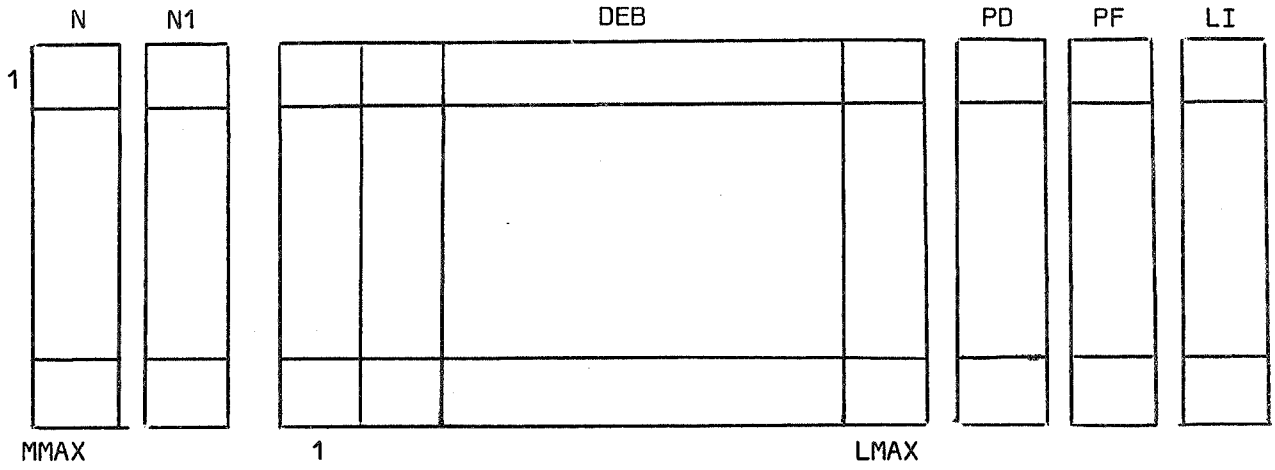
LL(i)= 0 si la ligne i est disponible
j ≠ 0 sinon (le prochain mot sera recopié à partir de la case j)

Fig. 1 suite page suivante.

LIST(i,.) = liste de mots phonétiques séparés par une syllabe blanche.

CH(i) = adresse de la liste chaînée suivante
ou 0 si fin de chaîne

fig. 1 (suite et fin)



N(i) = n° de la syllabe traitée
N1(i) = n° de la syllabe lue
DEBUT(i,.) = suite de syllabes en cours d'examen
PD(i) = pointeur première liste de la chaîne
PF(i) = pointeur dernière liste de la chaîne
LI(i) = 0 pas de liaison
1 si liaison obligatoire
2 si liaison facultative
3 si liaison interdite

Fig. 2

4°/ Structure de l'organe de décision DEC

- Initialisation :

a) LIST contient une seule liste qui est vide (la 1ère par exemple)
c'est-à-dire :
LL(1) = 1, LIST(1,J) = blanc 1≤J≤JMAX, CH(1) = 0 LL(I) = 0,
LIST(I,J) = blanc 1<I≤IMAX 1≤J≤JMAX, CH(I) = 0

b) DEB contient une seule suite qui est vide et qui pointe vers la seule liste de LIST c'est-à-dire :

N(1) = 0, N1(1) = 0, DEB(1,L) = blanc 1≤L≤LMAX, PD(1) = 1, PF(1) = 1, LI(1) = 0

N(M) = 0, N1(M) = 0, DEB(M,L) = blanc 1<M≤MMAX 1≤L≤LMAX PD(M) = 0, PF(M) = 0, LI(M) = 0

c) Soit ℓ_1 le n° de la 1ère ligne libre dans DEB

Donc $\ell_1 = 2$.

- Noyau de l'algorithme :

a) Soit (m-1) le nombre de syllabes de SYL déjà lues. S'il n'y a plus de syllabes à lire aller en e) sinon aller en b).

b) Lire la syllabe σ_m et l'écrire immédiatement à droite de toutes les suites de DEB associées à une chaîne de listes de LIST, c'est-à-dire toute suite DEB(i,.) telle que PD(i) ≠ 0. Pour chacun de ces débuts on incrémente de 1 la colonne N1.

c) Si une suite de DEB est telle que N≠N1 aller en d) sinon aller en a).

d) Soit $S_1 S_2 \text{ — } S_n S_{n+1}$ la suite de syllables à examiner.

On pose :

$d = S_1 S_2 \text{ — } S_n$ suite de syllabes avec laquelle on interroge le lexique.

$X = S_{n+1}$ syllabe lue σ_m

Nous noterons :

- $\sqcup$: la syllabe blanche (ou vide)
- $\mathcal{C}$: toute consonne
- c : toute semi-voyelle
- $\mathcal{C}_1$: toute semi-voyelle ou consonne
- Apostrophe postérieure (antérieur) l'opérateur de suppression du phonème (antérieur) d'une syllabe.

La procédure d'examen de la suite dX est représentée par les tables 1 et 2.

La table 1 est utilisée à l'initiale de la suite dX soit quand

d = syllabe vide

$X = S_1$

ou encore N = 0 et N1 = 1

La table 2 correspond au cas général :

$N \geq 1$ $N1 = N+1$

Chaque ligne correspond à une transformation possible de la suite dX.

Partant de la 1ère transformation on procède de la façon suivante :

- si les conditions d'application sont vérifiées, la suite dX est transformée en la suite $d^T X^T$. On interroge ensuite le lexique avec d^T. La table indique la ou les actions à faire en fonction de la réponse du lexique et de l'état des indicateurs FO et NFO, ainsi que le n° de la prochaine transformation.
- sinon on se reporte à la transformation mentionnée.

L'indicateur FO (fin de mot obligatoire) est positionné à VRAI

- si l'on postule une ellision devant une voyelle
- si l'on postule deux ellisions dans un mot
- en présence de combinaisons phonétiques attestées uniquement en fin de mot /4/.

L'indicateur NFO (non fin de mot obligatoire) est positionné à VRAI en présence de configurations phonétiques attestées uniquement à l'intérieur d'un mot /4/.

L'indicateur de liaison LI sera positionnée à chaque fois que d^T sera un mot admissible (cf. traitement A).

Les liaisons seront traitées comme une transformation de la 1ère syllabe d'une suite donc pour :

$d=S_1$ $X=S_2$ ou encore N=1, N1=2

TABLE 1

	Suite transformée		Conditions d'application de la transformation	Réponses admissibles	Actions	Aller à
	d^T	X^T				
1	⊔	S_1	Aucune	Aucune	Acquitter la ligne ℓ dans DEB	2
2	ↄə	$'S_1$	S_1 commence par une consonne sinon aller en 4	a) ↄə est seulement un début de mot et non FO b) ↄə est seulement un mot et non NFO c) ↄə est un début de mot et un mot, non FO et non NFO d) autres réponses	a) Exécuter A(ℓ,ℓ_1,⊔) puis B(1,ℓ_1,(ↄə) $'S_1$) b) Exécuter A(1,ℓ_2,ↄə) puis B(1,ℓ_2, $'S_1$) c) Exécuter a) puis b) d) aucune action	3
3	(ↄə) (ↄə)	$"S_1$	S_1 commence par deux consonnes sinon aller à FIN	a) d^T est seulement un mot et non NFO b) d^T n'est pas un mot	a) A(ℓ,ℓ_1,d^T) puis B(1,ℓ_1,$"S_1$) b) aucune action	FIN
4	i	$'S_1$	S_1 doit commencer par la semi-voyelle j qu'on supprime pour fournir $'S_1$ et $'S_1$ = a/ã/ɔ/ɔ̃/y/ɛ/e sinon aller à FIN		Exécuter A(ℓ,ℓ_1,i) puis B(1,ℓ_1,$'S_1$)	FIN
FIN					Fin du traitement	

	Suite transformée		Conditions d'application de la transformation	Réponses admissibles	Actions	Aller à
	d^T	X^T				
1	$S_1 \ldots S_n$	S_{n+1}	Aucune	a) d^T est seulement un début de mot et non FO b) d^T est seulement un mot et non NFO c) d^T est un mot et un début de mot et non FO et non NFO d) autres réponses	a) acquitter la ligne ℓ dans DEB soit $N(\ell) = N1(\ell)$ b) Exécuter $A(\ell, \ell_1, d^T)$ puis $B(0, \ell_1, X^T)$ c) Faire a) puis b) d) aucune action	2
2	$S_1 \ldots (S_n$ C)	$'S_{n+1}$	S_{n+1} commence par une semi-voyelle sinon aller en 3	a) d^T est un mot* et non NFO b) d^T n'est pas un mot * Pour la transformation 6, d^T doit être de plus un adjectif	a) Exécuter $A(\ell, \ell_1, d^T)$ puis $B(1, \ell_1, X^T)$ b) aucune action	5
3	$S_1 \ldots (S_n$ C)	$'S_{n+1}$	S_{n+1} commence par une consonne sinon aller en 5			4
4	$S_1 \ldots (S_n$ C C)	$''S_{n+1}$	S_{n+1} commence par deux consonnes sinon aller en 5			5
5	$S_1 \ldots S_n$ (C, ə)	$'S_{n+1}$	S_{n+1} commence par une consonne ou une semi-voyelle sinon aller en 7			6
6	$S_1 \ldots (S'_n \mid \tilde{\varepsilon} \mid \tilde{ɔ} \mid)$	$'S_{n+1}$	S_n se termine par ɛ ou ɔ S_{n+1} commence par N qu'on supprime pour former $'S_{n+1}$ sinon aller à 7			FIN
7	$S_1 \ldots S'_n$ (C ə)	S_{n+1}	S_n se termine par une semi-voyelle ou une consonne S'_n se termine par une voyelle sinon aller à FIN	a) d^T est seulement un début de mot et non FO b) d^T n'est pas un début de mot	a) Exécuter $A(\ell, \ell_1, \sqcup)$ puis $B(0, \ell_1, d^T X^T)$ b) aucune action	FIN
FIN					Si la ligne ℓ n'a pas été acquittée l'effacer ainsi que les listes de mots qui lui sont attachées	

Le traitement se fera de la façon suivante :

- s'il n'y a pas de liaison ou si interdite, les transformations seront faites sur $dX = S_1S_2$
- si la liaison est obligatoire on prendra $dX = 'S_1S_2$
- si la liaison est facultative on fera le traitement une fois avec $dX = S_1S_2$, une autre fois avec $d_X = 'S_1S_2$.

Examinons pour finir les différentes actions permettant de tenir à jour les listes de LIST et les suites de DEB.

- L'acquittement d'une suite de DEB se fait en égalant N à N1. Cette suite ne sera examinée qu'après l'apport d'une syllable supplémentaire.
- L'effacement d'une suite de DEB rend la ligne disponible, PD=0 N=0 N1=0 toutes les listes de LIST rattachées à cette suite sont aussi effacées.
- Le traitement $A(\ell,\ell_1,M)$ permet de dupliquer une chaîne de liste de mots de LIST rattachées à la suite de DEB figurant à la ligne ℓ, d'ajouter le mot M à chacune de ces listes, de les chaîner, de les rattacher à une ligne ℓ_1 de DEB, enfin de positionner l'indicateur de liaison $LI(\ell_1)$ suivant les attributs du mot M.

Soit n_S le nombre de syllabes du mot M. La description du traitement est la suivante :

i. soit k_1 le n° de la 1ère ligne libre dans LIST ($LL(k_1) = 0$).

On pose $PD(\ell_1) = k_1$, $k = PD(\ell)$

ii. Recopier $LIST(k,.)$ dans $LIST(k_1,.)$ sur $LL(k)-1$ positions.

Si M n'est pas le mot vide compléter à droite avec le mot M en séparant par une syllabe blanche et faire

$$LL(k_1) = LL(k) + n_S + 1$$

sinon faire $LL(k_1) = LL(k)$.

iii. Si CH(k) = 0 alors faire $CH(k_1) = 0$ et aller en iiii.
sinon rechercher k_1' la 1ère ligne libre de LIST.
Faire $CH(k_1) = k_1'$ remplacer $k_1 = k_1'$ k par CH(k) et aller en ii.

iiii. Positionner l'indicateur de liaison $LI(\ell_1)$ en fonction des attributs du mot M.

iiiii. Fin.

NB : Lorsque les filtres syntaxiques simples seront définis et mis en place, la construction d'une nouvelle liste ne sera faite que si le mot ajouté y est 'acceptable'.

- Le traitement $B(a,\ell_1,D)$ permet à la suite de A de compléter la ligne ℓ_1 de DEB, de l'acquitter éventuellement, et de résoudre le cas de suites identiques dans DEB.
Soit n_k le nombre de syllabes du début D. Le traitement est le suivant :

i. On range le début D dans $DEB(\ell_1,.)$ sur n_k positions à partir de la gauche
Si a=1 on acquitte la ligne ℓ_1
soit $N(\ell_1) = N1(\ell_1) = n_k$
Si a=0 on pose $N1(\ell_1) = n_k$
$N(\ell_1) = n_k - 1$

ii. S'il existe dans DEB une ligne ℓ' identique à la ligne ℓ_1 (aux positions de pointeurs PD et PF près) alors on chaîne les suites de listes attachées aux lignes ℓ' et ℓ_1, en faisant :

$$CH(PF(\ell')) = PD(\ell_1)$$

$$PF(\ell') = PF(\ell_1)$$

et on libère la ligne ℓ_1 en posant

$$PD(\ell_1) = PF(\ell_1) = 0$$
$$N(\ell_1) = N1(\ell_1) = LI(\ell_1) = 0$$

sinon on recherche le n° ℓ_1 de la 1ère ligne libre de DEB.

e) Effacer toutes les lignes de DEB qui n'ont pas permis de former un mot à l'issue de cette dernière étape. Les chaînes de listes rattachées à chacune de ces lignes seront aussi effacées.

4. CONCLUSION

Bien du travail reste encore à faire pour la mise en oeuvre du système et en mesurer les performances effectives, qui visiblement seront excellents dès que le lexique sera implanté en mémoire centrale.

On peut déjà en apercevoir l'intérêt :

(i) Indépendance par rapport aux applications du module principal.

(ii) Possibilité d'y faire intervenir à chaque instant, les acquis de la phonétique, les filtres syntaxiques, les lexiques, etc...

Par ailleurs, le système est susceptible de perfectionnement, nous pensons notamment aux syllabes spéciales.

REFERENCES BIBLIOGRAPHIQUES

(1) ROSSI M.
Les contraintes phonologiques dans un système de reconnaissance de la parole.
6ème Journées d'Etudes sur la Parole. Toulouse 28-30 mai 1975

(2) DELATTRE P.
Studies in French and Comparative Phonetics.
Mouton 1966. Edit. C.H. Van Schooneveld

(3) DOURS D., FACCA R., PERENNOU G.
Analyse temporelle du signal vocal, comparée à l'analyse fréquentielle classique du point de vue de la reconnaissance.
5ème Journées d'Etudes sur la Parole. Orsay 15-17 mai 1974

(4) DOURS D., FACCA R., LAURENTIE Y., MAURAND G., PERENNOU G.
A propos de marqueurs lexicosyntaxiques : quelques exemples commentés de phrases issues de l'analyseur du projet A.R.I.A.
6ème Journées d'Etudes sur la Parole. Toulouse 28-30 mais 1975

(5) DOURS D., FACCA R., MAURAND G., PERENNOU G.
L'apport de l'analyseur du projet A.R.I.A. sur quelques exemples d'analyse phonétique.
6ème Journées d'Etudes sur la Parole. Toulouse 28-30 mais 1975

(6) DOURS D., FACCA R., PERENNOU G.
Analyse d'un signal fortement structuré : le signal vocal.
Colloque national sur le traitement du signal et ses applications.
Nice 16-21 juin 1975

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

RECONNAISSANCE DE PAROLE ET SEGMENTATION

Jean A. DREYFUS-GRAF, Genève (Suisse)

Résumé :

La machine de reconnaissance de la parole proposée (phonétographe V) cherche à appliquer non seulement les lois de l'émetteur vocal, mais aussi celles du récepteur auditif, qui sont non-linéaires. Les composantes spectrales et les dynamiques des phonèmes plosifs et non-plosifs sont examinés séparément. La segmentation de la parole en phonèmes s'effectue selon des principes majoritaires, utilisant les maximums des composantes spectrales quasi-stationnaires et de leurs transitions. La matrice de reconnaissance des phonèmes est corrigée par les composantes de modulation, telles que pitch, souffle, roulement et degré de voisement. Concernant la parole naturelle, les taux de reconnaissance de phonèmes (téléphonés) sont de l'ordre de 90 % pour l'homme, représentant la limite escomptable aussi pour la machine. Pour avoir accès aux 10 % restants, on peut avoir recours à des règles autres que phonétiques, telles que phonologiques, lexicales ou grammaticales. Concernant la parole codée (phonocode SOTINA), les taux de reconnaissance de phonèmes, escomptables directement, sont de l'ordre de 99.9 %, soit 100 fois meilleurs.

Summary :

The proposed speech recognition machine (phonetograph V) seeks to apply not only the laws of the vocal transmitter but also those of the hearing receiver, which are non-linear. The spectral components, and the dynamics, of the plosive and non-plosive phonemes are examined separately. The segmentation of speech into phonemes is performed by using the maxima of the spectral components and of their transitions, according to majority principles. The recognition matrixes are corrected by modulation components, such as pitch, hizz, rolling and voicing degree. Concerning natural speech, the phoneme recognition rates over the telephone are of the order of 90 % for human listeners, representing the expected rates for machines as well. The access to the remaining 10 % may use additional rules, such as phonological, lexical or grammatical. Concerning coded speech (phonocode SOTINA), the directly expected phoneme recognition rates are of the order of 99,9 %, that is 100 times better.

RECONNAISSANCE DE PAROLE ET SEGMENTATION

Jean DREYFUS-GRAF, Genève (Suisse)

1. Segmentation en mots, syllabes et phonèmes

D'une manière générale, la reconnaissance de la parole implique sa segmentation en éléments plus ou moins complexes, tels que phrases, mots, syllabes ou phonèmes. Une langue naturelle, telle que le français, comprend des milliards de phrases différentes, mais celles-ci sont réductibles d'abord à quelques milliers de mots, puis à quelques centaines de syllabes, et enfin à quelques dizaines de phonèmes alphabétiques.

Plus la segmentation sera fine et plus la mémoire de reconnaissance de parole sera déchargée, donc économique, que cette mémoire soit humaine ou technique.

Au cours des dernières 40 années, de nombreux prototypes ou systèmes ont été expérimentés en vue de réaliser la reconnaissance automatique de la parole.

Ceux que nous avons développés visaient tous à la reconnaissance des phonèmes, la segmentation étant basée sur les variations temporelles des maximums d'énergie spectrale, nommés "formants". Les fig. 1 à 5 reproduisent quelques résultats fournis par nos prototypes expérimentaux, dont les années et les noms sont rappelés dans le tableau suivant :

année	nom du prototype	réf.	résultats
1948	sonographe, avec enregistreur à 4 composantes	/1/	fig. 1
1950	sténo-sonographe, avec enregistreur à 6 composantes	/2/	fig. 2
1952	phonétographe I-II, avec télé-imprimeur "Robotyper"	/3/	fig. 3
1961	phonétographe III , " " " "	/4/	fig. 4
1971	phonétographe IVa , avec enregistreur à 16 pistes calculateur PB250 et imprimante	/5/	fig. 5

Les recherches en reconnaissance de parole, effectuées dans d'autres laboratoires sont orientées généralement vers des segmentations en éléments plus complexes que les phonèmes, tels que des mots /6/, des di-phonèmes /7/, ou des syllabes centrées sur des voyelles/8/.

2. Approche digitale

Afin d'utiliser pleinement l'assistance par ordinateur moderne, la reconnaissance de la parole abandonne progressivement la technologie analogique pour se convertir à la technologie digitale. A la limite, l'onde acoustique elle-même est soumise d'emblée à la conversion analogue-digitale /9/. On peut encore signaler diverses méthodes utilisant par exemple la "prédiction linéaire" /10/,/11/, la"programmation dynamique" /12/, les "sous-ensembles flous" /13/, "l'analyse factorielle" /14/,"l'analyse synchronisée avec le pitch /15/".

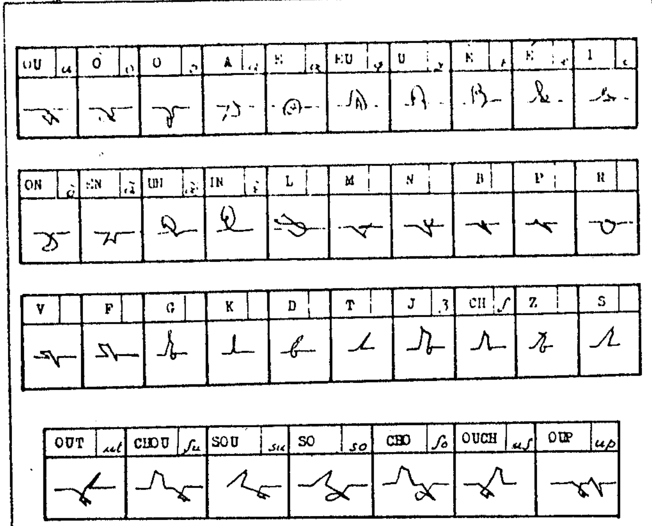


Fig.1. 1948 : Phonémogrammes de phonèmes isolés et de syllabes, par le 1er "sonographe" à 4 composantes

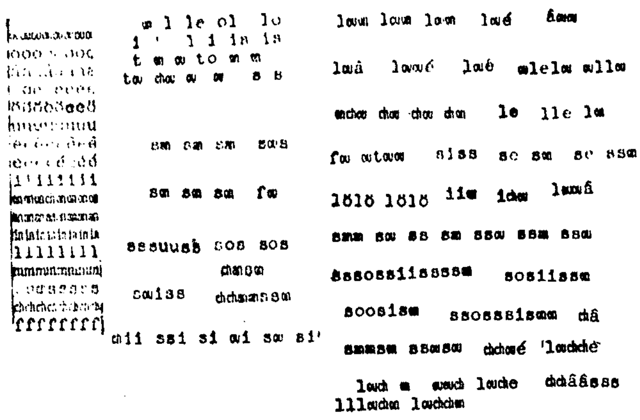


Fig.3. 1952 : Phonétogrammes de phonèmes prolongés (auto-répétés 8 fois par seconde), et de quelques mots, fournis par le "phonétographe I, II"(imprimant).

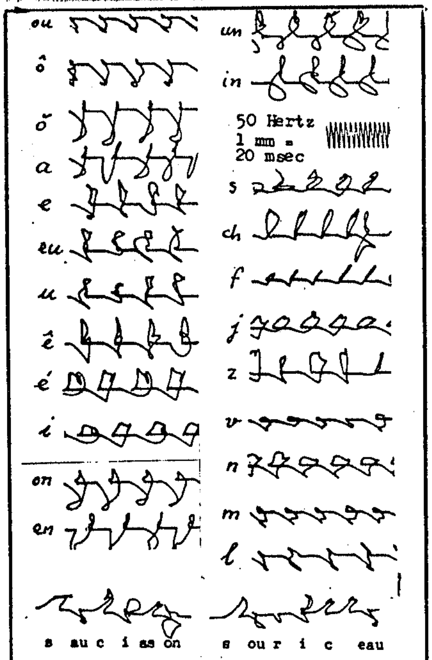


Fig.2. 1950 : Phonémogrammes de phonèmes (auto-répétés 10 fois par seconde) et des mots "saucisson","souriceau", fournis par le "sténo-sonographe" à 6 composantes.

a) alphabet phonétique français

médaamm éé mmnééssiiiöö vvou vvêêêrréé mmininintceeennan
la pââârolle deevenniiiir âââkssiiionon vou ââssiiissstế aa la
preemmmiiiêêêrr métâmooorrrfffoozz oootoommmâââtiiike du llananangââjj
pââââârlllé an llananangâjj éékrii pââar llle fonnnééétogrrââaff
proototiiipe trrououououââ aaalleemanan vvoort viirrd taaat

ssonon ffonéétiiik frrassê vououououââiiiiêêêlll éé konoonssoonnn teennnuuu
ouououououououououou ôôôôôôôôôô oooooooooo ââââââââââ aaaaaaaaaa eeeeeeeee
öööööööööö uuuuuuuuu êêêêêêêêêê éééééééééé iiiiiiiii
nââazzzââalliizzzzzzzéé onononononononon anananananananananan unununununununun inininininininin
llllllllll mmmmmmmmmm nnnnnnnnnn rrrrrrrrr konoonsssonnnn fffllluktué
jjeee vvee zee chchchchchch ffffffffff sssssssss
konoonssonnn pllllloozzziiv bee deee gee pe te k
ffinin mmmêrssii de voootrrr ââtanonssiiiouon

b) alphabet phonétique international

medaamm ee mmmeessiiiøø vvu vvɛɛɛrrəe mmɛ̃ɛ̃ɛ̃təəennɑ̃
la paaarolle dəevenniiiir aaakssiiiɔ̃ɔ̃ɔ̃ vu aassiiissste aa la
prəemmmiiiɛɛɛrr metamooorrrfffoozz oootoommmaaatiiikə dy llɑ̃ɑ̃ɑ̃gaaʒʒ
paaaarllle ɑ̃ llɑ̃ɑ̃ɑ̃gaʒʒ eekrrii paaur llle fonnneetogrraaaff
prootəotiiipə trruuuaa rreeegllllaaaʒ nnnyyymmɛerɔɔ dø
eeʃɑ̃ɑ̃ɑ̃tiiiiɔ̃ɔ̃ ɑ̃ɑ̃glɛɛ uerd gɛɛtss ɛɛɛkʃʃʃnn iiitiiinng iiiz
prruuuff of puuuding aaalləəmɑ̃ɑ̃ vvoort viirrd taaat

ssɔ̃ɔ̃ ffoneetiiik frrɑ̃sse vuuuaaaiiiɛɛɛlll ee kɔ̃ɔ̃ɔ̃ssɔɔnnn təennnyyy
uuuuuuuuuu oooooooooo ɔɔɔɔɔɔɔɔɔɔ aaaaaaaaaa aaaaaaaaaa əəəəəəəəəə
øøøøøøøøøø yyyyyyyyyy ɛɛɛɛɛɛɛɛɛɛ eeeeeeeeee iiiiiiiiii
naaazzzaaalliizzzzzee ɔ̃ɔ̃ɔ̃ɔ̃ɔ̃ɔ̃ɔ̃ɔ̃ɔ̃ɔ̃ ɑ̃ɑ̃ɑ̃ɑ̃ɑ̃ɑ̃ɑ̃ɑ̃ɑ̃ɑ̃ œ̃œ̃œ̃œ̃œ̃œ̃œ̃œ̃œ̃œ̃ ɛ̃ɛ̃ɛ̃ɛ̃ɛ̃ɛ̃ɛ̃ɛ̃ɛ̃ɛ̃
llllllllll mmmmmmmmmm nnnnnnnnnn rrrrrrrrrr kɔ̃ɔ̃ɔ̃sssonnnn ffflllyktye
ʒʒəəə vvəə zəə ʃʃʃʃʃʃ ffffffffff ssssssssss
kɔ̃ɔ̃ssonnn pllllloozzziiv bəə dəəə gəə pə tə k
ffɛ̃ɛ̃ mmmɛrssii də vooɔtrrr aatɑ̃ɑ̃ssiiiɔ̃ɔ̃ɔ̃

Fig.4. 1961 : Phonétogrammes d'un discours et de phonèmes prolongés (auto-répétés 6x par seconde), imprimés par le "robotyper" du "Phonétographe III"
a) en alphabet phonétique français, b) en alphabet phonétique international

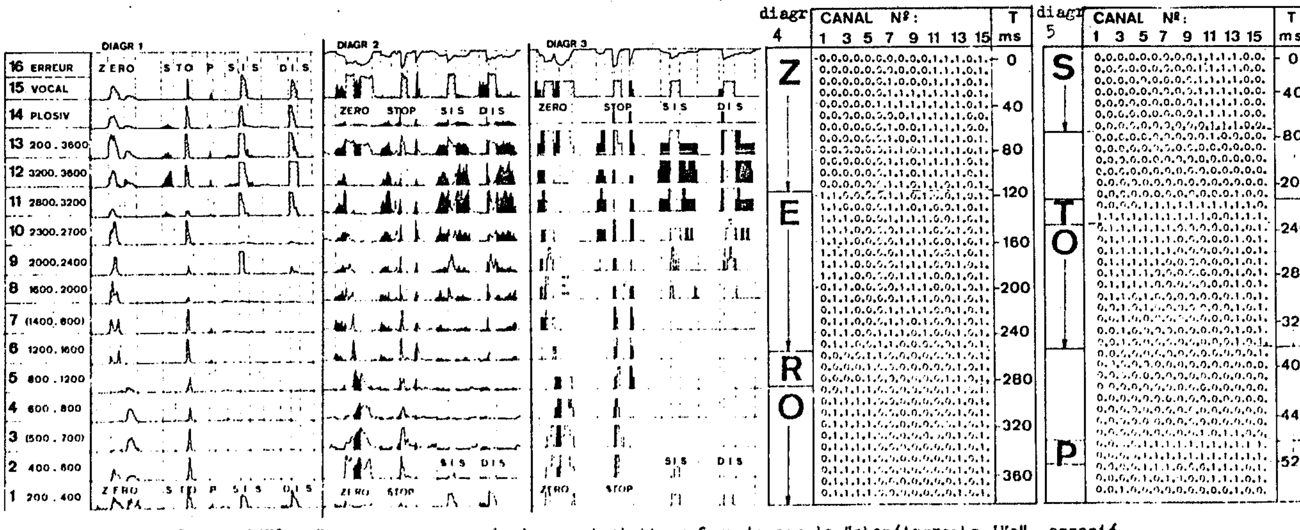


Fig.5. 1971 : Spectrogrammes analogiques et digitaux fournis par le "phonétographe IVa", associé à un enregistreur "Sanborn" à 16 pistes et à un calculateur "PB 250".
DIAGR 1 : spectres linéaires (sans compresseur) des mots "ZERO","STOP","SIS","DIX"
2 : " logarithmiques (avec 4 compresseurs sélectifs, à simples boucles)
3 : " logarithmiques et quantifiés à 3 niveaux 0,1,2
4,5 : " digitalisés à 2 niveaux 0,1, avec reconnaissance des mots "ZERO","STOP"

Toutefois, on doit constater que l'approche digitale retrouve, sous des formes perfectionnées, les mêmes difficultés que l'approche analogique. Ces difficultés se manifestent d'autant plus que les 3 conditions suivantes doivent être davantage respectée :

a) Reconnaissance en temps réel
b) Limitation des fréquences analysées à la bande téléphonique (300-3400 Hz) et perturbation par les bruits.
c) Adaptation à plusieurs locuteurs, masculins et féminins.

3. Développement d'un nouveau phonétographe

Les phonétographes III et IV ne pouvaient pas suivre la vitesse d'élocution normale, ni s'adapter à plusieurs locuteurs, car leurs analyseurs spectraux étaient simplifiés, et qu'ils ne disposaient d'aucun extracteur de pitch.

La fig. 6 illustre le principe d'une machine de reconnaissance, qu'on peut nommer "phonétographe V", et qui propose des systèmes plus élaborés et mieux adaptables. Le schéma est décrit en technologie hybride, mi-analogique et mi-digitale, la partie analogique pouvant être digitalisée, par la suite, selon les besoins.

4. Quelques remarques préliminaires

a) Fonctions de transfert : émetteur vocal et récepteur auditif

La plupart des approches digitales récentes, telles que mentionnées ci-dessus, considèrent les fonctions de transfert de l'émetteur vocal (sources d'excitation et conduit vocal), mais non celles du recepteur auditif, qui sont essentiellement non-linéaires.

Ainsi, la réponse du récepteur auditif à des sons prolongés suit des lois doublement logarithmiques /16/. D'autre part, sa réponse à des impulsions brèves dépend d'un grand nombre de facteurs, tels que raideur de l'attaque (pente de la variation d'énergie) et vitesse de répétition des impulsions. Par exemple, l'oreille distingue 2 impulsions brèves, distantes de 20 millisecondes seulement, mais la répétition continue de ces mêmes impulsions, 50 fois par seconde, est perçue comme un son grave unique.

b) Prédiction linéaire synchrone et récepteur auditif

L'analyse de la parole par "prédiction linéaire synchrone" ne semble pas conforme à celle qui pourrait être effectuée par le récepteur auditif : celui-ci n'est guère capable, en effet, de synchroniser l'analyse spectrale des signaux modulés par le conduit vocal avec le rythme des impulsions glottales, c'est-à-dire d'effectuer 100 à 500 analyses complètes par seconde, selon qu'il s'agit d'une voix grave ou aiguë.

c) Existence physique des phonèmes

La possibilité d'une segmentation phonétique est contestée par de nombreux auteurs qui nient l'existence même des phonèmes, et surtout celle des consonnes plosives p,t,k. Pourtant le fait quel'oreille

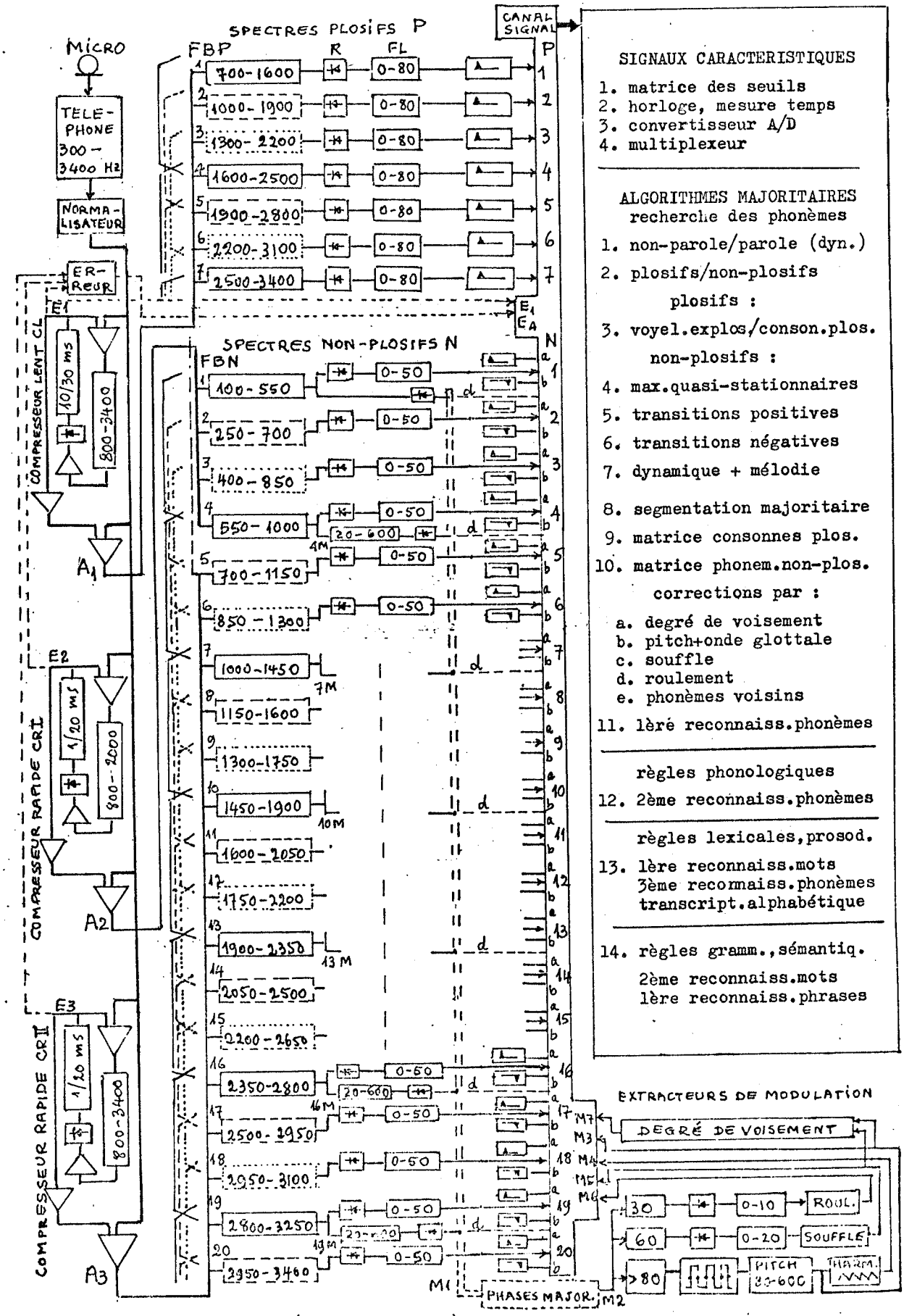


Fig.6. Schéma de principe (phonétographe V)

distingue les mots "chaque", "chatte" et "chape" prouve que les plosives finales, même dévoisées, sont reconnaissables. On peut en conclure que les pentes ascendantes de leurs énergies spectrales sont déjà significatives, les autres critères, tels que durées, "locus" ou influences par l'environnement étant d'importance secondaire.

d) Restrictions au critère de la dynamique

De nombreux auteurs considèrent que le critère de la dynamique est primordial pour effectuer la séparation entre voyelles et consonnes, l'énergie des voyelles étant plus grande que celle des consonnes. Pourtant pour bien comprendre les consonnes "m", "n", "l" dans les mots "mime", "mine", "mille", il faut leur donner davantage d'énergie qu'à la voyelle "i".

e) Restriction au critère du voisement

D'autre part, de nombreux auteurs considèrent le critère du voisement comme fondamental pour la reconnaissance. Pourtant des mots tels que "choc", "chic", "chaque" peuvent être prononcés d'une manière voisée ou chuchotée sans que leur intelligibilité en soit notablement affectée. Ainsi la distinction fondamentale des phonèmes en général, et des voyelles sonores "o", "i", "a" en particulier, ne réside pas dans leur voisement plus ou moins accentué, mais bien dans leurs composantes spectrales.

5. Système non-linéaire pour l'extraction des signaux caractéristiques

Le système d'extraction des signaux caractéristiques de la fig. 6 essaie de se conformer aux lois de l'émetteur vocal en même temps qu'à celles du récepteur auditif, qui sont non-linéaires, et il tient compte des remarques précédentes.

Il opère essentiellement les fonctions suivantes :

a) Séparation entre l'analyse spectrale des phonèmes plosifs P et celle des non-plosifs N :

Pendant les 10 premières millisecondes après chaque silence d'au moins 60 millisecondes, dans la bande 800 à 3400 Hz, le compresseur lent CL sélectionne les fronts raides des phonèmes plosifs P, et élimine tous les phonèmes non-plosifs N. Les "phonèmes plosifs P" englobent les consonnes plosives p,t,k,b,d,g aussi bien que les voyelles explosées !a !o, etc., la séparation entre ces 2 groupes s'effectuant par la suite. Les constantes de temps du compresseur lent CL sont T_{L1} = 10 ms à la montée et T_{L2} = 30 ms à la descente /17/. La sortie auto-régulée A1 du compresseur CL alimente les 7 filtres de bande FBP 1 à 7, dont la fonction se limite à l'analyse spectrale des fronts plosifs.

b) Séparation entre la dynamique significative DS et non-significative NS

Les 2 compresseurs rapides CR1 et CR2 suppriment la dynamique des phonèmes non-plosifs au-dessus d'un certain seuil. Leurs constantes de temps sont T_{R1}=1 ms à la montée et T_{R2}=20 ms à la descente. Les sorties auto-régulées A2 et A3 alimentent les 20 filtres de bande FBN 1 à 20, qui opèrent l'analyse spectrale des parties quasi-stationnaires des phonèmes, indépendamment de la dynamique (ou variation d'amplitude) globale.

Celle-ci est récupérée ensuite par les signaux d'erreur SE1+SE2+ SE3= SE4= DS, permettant la séparation parole/non-parole et fournissant un critère auxiliaire pour la séparation voyelle/consonnes.

c) Les 2 sous-ensembles de l'analyse spectrale

Pour les parties quasi-stationnaires des phonèmes non-plosifs N, chacun des 20 filtres de bande FBN 1 à 20 présente une largeur de 450 Hz, admettant une constante de temps T = 1 milliseconde, mais ils se chevauchent par groupes de 3, ayant ainsi un pouvoir séparateur de 150 Hz (=450:3). Ils sont suivis de 20 redresseurs et de 20 filtres passe-bas 0-50 Hz (fenêtre 20 ms) délivrant 20 signaux d'analyse quasi-stationnaires dans les canaux N 1 à 20. Les variations temporelles, positives et négatives, de ces signaux seront mesurées après digitalisation et sont symbolisées par les 2x20 = 40 canaux N1a-N20a, N1b-N20b.

Pour les fronts des phonèmes plosifs P, l'analyse spectrale est moins fine et plus rapide. Les 7 filtres de bande FBP 1 à 7 présentent des largeurs de 900 Hz, avec des constantes de temps T_I= 0,5 ms. Ils se chevauchent par groupes de 3, admettant un pouvoir séparateur de 300 Hz (=900:3). Ils sont suivis de 7 redresseurs R et de 7 filtres passe-bas 0-80 Hz (fenêtre 12 ms) et délivrent 7 signaux d'analyse de phonèmes plosifs dans les canaux P 1 à 7.

d) Extraction de pitch, signal glottal, souffle, roulement et degré de voisement

Les spectres des phonèmes quasi-stationnaires peuvent être modulés par 3 sources d'excitation de l'émetteur vocal : 1) par l'onde glottale, dont la fréquence fondamentale, ou pitch, peut varier entre 80 et 600 Hz, 2) par le souffle dont le spectre aléatoire descend au-dessous de 80 Hz, 3) par des roulements de /r/ ou /x/, qui se trouvent autour de 30 Hz.

Afin d'extraire ces 3 catégories de modulation, les filtres de bande "non-plosifs", tels que FBN 4, 7, 10, 13, 16, 19, sont suivis de filtres secondaires 30-600 Hz, tels que 4M, 7M, 10M, 13M, 16M, 19M, et de redresseurs, tandis que le premier filtre FNB 1 (100-550 Hz) est directement suivi par un redresseur. On obtient ainsi 7 composantes redressées dans les canaux N 1c, 4c, 7c, 10c, 13c, 16c, 19c. La résultante de leurs phases dominantes apparaît dans les canaux M1, puis M2. Elle est analysée par extracteurs : de pitch (> 80Hz), de souffle (60 Hz), et de roulement (30 Hz). Les harmoniques (HARM) d'une onde triangulaire peuvent être ajoutées au pitch pour reconstituer l'onde glottale. Finalement, on peut comparer l'énergie du souffle avec celle du pitch pour obtenir le "degré de voisement". Celui-ci est très variable, entre la voix forte et la parole chuchotée. Il peut modifier les spectres quasi-stationnaires, qui doivent être corrigés en conséquence. Les 5 signaux de modulation ainsi extraits aboutissent aux canaux M3 à M7.

e) Digitalisation des signaux caractéristiques

Récapitulons les signaux caractéristiques à traiter :

nombre	signaux	destination
7	P1-P7	analyse spectrale des fronts de phonèmes plosifs
20	N1-N20	analyse spectrale des phonèmes non-plosifs
2	E3-E4	dynamique (signaux d'erreur)
5	M3-M7	pitch, onde glottale, souffle, roulement, voisement
34		

Ces signaux caractéristiques sont digitalisés par un convertisseur analogue-digital (A/D), puis multiplexés. Leur traitement logique s'effectue dès lors par ordinateurs ou micro-precesseur.

6. Principes majoritaires

La fig. 6 ne donne qu'un bref aperçu des fonctions logiques, qui restent à développer. La segmentation et la reconnaissance de la parole s'effectuent selon des principes majoritaires, qui s'apparentent aux méthodes utilisant les probabilités ou les sous-ensembles flous.

Du fait que les filtres de bande des analyseurs spectraux se chevauchent par groupes de 3, les fluctuations individuelles s'effacent devant les phases majoritaires. De plus, les dominantes spectrales résultent de leurs maximums transitoires associés à leurs maximums quasi-stationnaires, d'une manière majoritaire.

7. Nécessité et insuffisance de la recherche des phonèmes

La recherche des phonèmes est la condition nécessaire, mais insuffisante à la reconnaissance optimum de la parole naturelle. En effet, la machine comprendra toujours la parole moins bien que l'homme, dont les taux d'erreur sont de l'ordre de 10 % concernant les "phonatomes", téléphonés. Pour réduire les taux d'erreur actuels des machines, qui sont très supérieurs à 10 %, il faut avoir encore recours à des sources de connaissances autres que phonétiques, telles que phonologiques ou lexicales, dans la mesure où elles sont disponibles. Mais nous pensons qu'il faut commencer par rechercher les phonèmes, qui sont au nombre de quelques dizaines, avant d'avoir recours aux syllabes, aux mots ou aux phrases, dont les nombres respectifs sont des centaines, des milliers et des milliards.

8. Segmentation de la parole codée (phonocodes)

Le nombre des phonèmes, qui est de quelques dizaines pour une langue naturelle, telle que le français, peut être à moins d'une dizaine, dans le cas de la parole codée ou "phonocode" /17/. Le phonocode SOTINA ne comprend que 6 phonèmes,dont 3 voyelles,O,I,A,et 3 consonnes,S,T,N.Les tests d'intelligibilité humaine ont montré que les taux d'erreur sont 100 fois plus faibles pour cette parole codée que pour le français. D'autre part, les "phono-décodeurs" se contentent d'une dizaine de signaux caractéristiques, délivrés par leurs analyseurs, au lieu de la trentaine qui serait exigée par le phonotographe V. Ainsi la parole codée permet d'étudier sous une forme très simplifiée les lois de reconnaissance de la parole naturelle.

R E F E R E N C E S

/1/ J.A.Dreyfus-Graf, "Le sonographe: éléments et principes". Annales Suisses des Sciences Appl. 14 (1948), 353...362

/2/ J.A.Dreyfus-Graf, "Le sténo-sonographe phonétique". Bull. Techn. PTT, Berne, No 3/1950, 89...95 (reproduit par l'Onde Electrique 30 (1950) et par la Revue Internationale de Sténographie.

/3/ J.A.Dreyfus-Graf, "Le typo-sonographe phonétique ou phonétographe". (phonétographe I, avec "Robotyper"). Conférence du 10 oct.1952, EAM, Genève. Bull.Techn. PTT, Berne, No 12/1952, 363...379

/4/ J.A.Dreyfus-Graf, "Phonétographe: Présent et futur". Bull. Techn. PTT, Berne, No 5/ 1961, 160...172

/5/ J.A.Dreyfus-Graf, "La reconnaissance automatique de la parole (phonétographe IVa)". L'Echo des Recherches, CNET Lannion et Issy-les Moulineaux, avril 1971.

/6/ T.B.Martin, "Application of limited vocabulary recognition systems". IEEE Symposium on Speech Recognition. Pittsburg 1974.

/7/ J.S.Liénard, "Analyse, synthèse et reconnaissance automatique de la parole". Thèse, Université Parix VI 1972.

/8/ C.J.Weinstein et al., "A System for Acoustic-Phonetic Analysis of Continuous Speech". IEEE, ASSP, Febr. 1975.

/9/ R.W.Schafer and L.R.Rabiner, "Digital Representation of Speech Signals". IEEE, Proceedings, April 1975.

/10/ J.Makhoul, "Linear Prediction: A Tutorial Review". IEEE, Proceedings, April 1975.

/11/ I.I.El-Mallawany, "Etude de vocoder à prédiction linéaire...". Thèse, Université de Grenoble, 1975.

/12/ M.Querré, "Reconnaissance de mots isolés à l'aide d'une méthode de programmation dynamique". C.N.E.T., Lannion. Compte-rendu des recherches. CEI, 1975.

/13/ J.Brémont, "Contribution à la reconnaissance automatique de la parole par les sous-ensembles flous". Thèse, Nancy, 1975.

/14/ J.P.Benzecri, "Leçons sur l'analyse factorielle et la reconnaissance des formes", ISUP, Paris, 1973.

/15/ W.J.Hess, "A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speech". IEEE,ASSP. Febr.1976.

/16/ J.A.Dreyfus-Graf, "Cybernétique auditive" (lois doublement logarithmiques de l'audition). Conférence du 9 mai 1968, GALF, Turin. Revue d'Acoustique, Paris No 14 (1971).
J.A.Dreyfus-Graf, "L'oreille comme boîte grise; sonie et tonie" ($sone_6$,$mel_{4,5}$). Conférence du 9 avril 1970, GALF, Paris.

/17/ J.A.Dreyfus-Graf, "Recognition of Coded Speech (Phonocodes)". IEEE, ICASSP, Philadelphia April 1976.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

UN VOCODEUR A CANAUX ADAPTES, SON APPRENTISSAGE : APPLICATION A LA PAROLE CONTINUE

Y. GRENIER et C. GUEGUEN
Laboratoire de Théorie des Systèmes de l'ENST

RESUME :

SUMMARY :

UN VOCODEUR A CANAUX ADAPTES, SON APPRENTISSAGE : APPLICATION A LA PAROLE CONTINUE

Y. GRENIER et C. GUEGUEN
Laboratoire de Théorie des Systèmes de l'ENST.
46, Rue Barrault

75634 - PARIS Cedex 13

1. PRINCIPE

Un vocodeur à canaux de type classique se compose d'un banc de filtres en parallèles, ces filtres étant le plus souvent du type passe-bande et analogiques.
Le vocodeur à canaux adaptés est constitué d'un banc de filtre dont chacun est accordé pour "résonner sur un phonème donné". Ces filtres sont digitaux.

Pour cela on détermine sur chaque phonème un modèle autorégressif optimal, que l'on considère ensuite comme un filtre inverse.

A l'ordre k, le modèle autorégressif a une expression de la forme $\{a_0 \ldots\ldots a_k\}$.

avec $a_0 = 1$

$$e_n = \sum a_i \, S_{n-i}$$

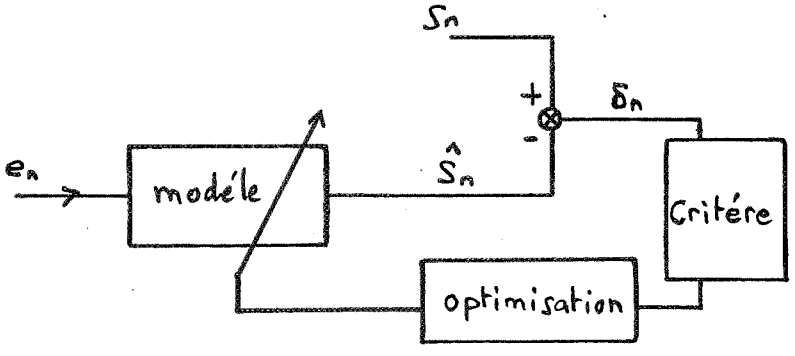


Fig. 1

les a_i optimisent au sens d'un critère donné l'erreur de prédiction $\delta_n = S_n - \hat{S}_n$; e_n est l'entrée "estimée" (bruit blanc)

Si les a_i sont connus et les échantillons précédents mesurés, la relation peut-être interprétée comme une prédiction linéaire estimant sous la forme $\hat{S}_n$ l'échantillon à l'instant n :

$$\hat{S}_n = -\sum a_i S_{n-i} \qquad \text{(supposons } e_n = 0\text{)}$$

et $\delta_n = S_n - \hat{S}_n$

Chaque canal du vocodeur constitue un filtre numérique inverse adapté ; l'apprentissage permet de déterminer les a_i du filtre "adapté" au phonème considéré). Le modèle phonémique étant construit, les coefficients a_i sont utilisés pour calculer l'erreur de prédiction sur un signal incident quelconque. Si le phonème prononcé correspond approximativement avec l'un des filtres, l'erreur de prédiction correspondante demeure faible.

Le banc de filtres adaptés fournit en sortie de chaque canal cette erreur sous forme d'une énergie, permettant de classer chaque canal suivant le degré décroissant de confiance à accorder à la reconnaissance du phonème inconnu par ce canal. L'énergie en sortie d'un filtre E_s s'écrit :

$$E_s = \sum_{n=k}^{N} (S_n - \hat{S}_n)^2 \qquad N = \text{nombre d'échantillons de la fenêtre.}$$

On montre que compte tenu de la relation :

$$\hat{S}_n = -\sum_{1}^{k} a_i S_{n-i}$$

l'énergie peut s'écrire :

$$E_s = A^t V \; A$$

avec $A = \begin{pmatrix} a_0 \\ a_1 \\ \cdot \\ a_k \end{pmatrix}$ et $V = \begin{pmatrix} v_0 & v_1 & \cdots & v_k \\ v_1 & \ddots & \ddots & v_1 \\ \vdots & \ddots & \ddots & \vdots \\ v_k & \cdots & v_1 & v_0 \end{pmatrix}$

où $v_i = \sum S_n S_{n-i}$ représente l'autocorrélation du signal incident.

$$E_s = \alpha^t v \qquad \text{avec} \quad v = \begin{pmatrix} v_0 \\ \cdot \\ \cdot \\ \cdot \\ v_k \end{pmatrix}$$

et $\alpha = \begin{pmatrix} a_0^2 + a_1^2 + \ldots + a_n^2 \\ \ldots\ldots\ldots\ldots\ldots \\ 2a_0 a_n \end{pmatrix}$

Cette dernière relation permet de considérer chaque filtre comme un discriminateur linéaire dans l'espace des vecteurs $\vec{v}$.
La réalisation du vocodeur à canaux adaptés est ainsi grandement simplifiée :

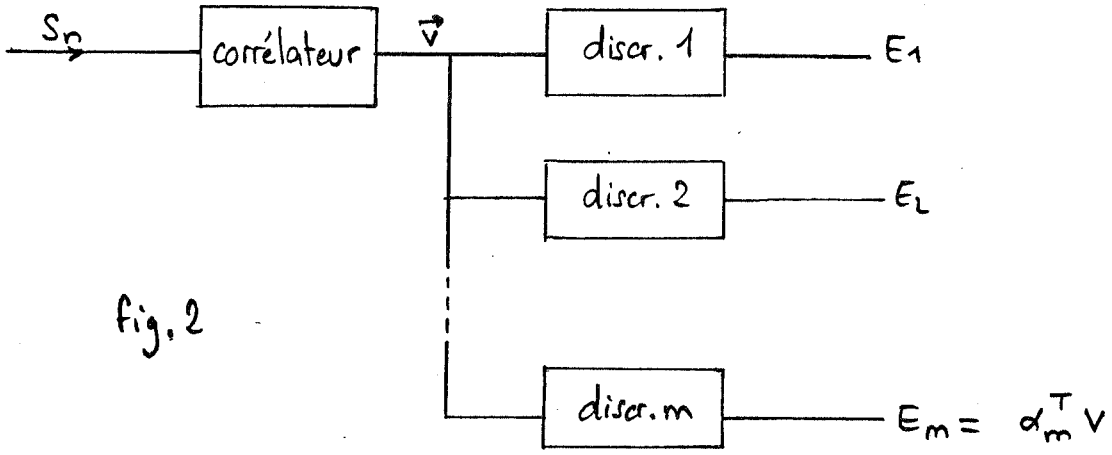


fig. 2

Chaque discriminateur réalise le produit scalaire du vecteur d'autocorrélation du signal et du vecteur α_j issus de l'apprentissage pour le J - ième phonème.

Le nombre des canaux du vocodeur est au moins égal au nombre des phonèmes à séparer, un même phonème pouvant être représenté par plusieurs modèles différents si par exemple le nuage des occurences de ce phonème à une forme trop compliquée pour être modélisable de façon efficace par une relation unique.

2. APPRENTISSAGE

La sortie E_s du filtre est :

$$E_s^2 = \alpha T V = A\ {}^t VA$$

Cette expression est également celle du critère des moindres carrés appliqué à l'erreur de prédiction S_n.La détermination du vecteur a_i minimisant E_s pour un signal dont v est donné est un problème classique en prédiction linéaire.
(algorithme de LEVINSON).

Le problème posé par l'apprentissage du vocodeur à canaux adaptés peut se ramener au précédent. Il s'agit en effet de déterminer à partir d'un ensemble d'occurence $\vec{v}_1\ \vec{v}_2\ \ldots\ \vec{v}_m$ de vecteurs d'autocorrélation calculés pour diverses productions d'un même phonème, quel est le vecteur $\{a_i\}$ ou $\vec{\alpha}$ minimisant un critère.

$$\mathcal{J} = \mathcal{J}(a_i, \vec{v}_1, \vec{v}_2, \ldots\ldots \vec{v}_m)$$

On peut donner à ce critère la forme d'une énergie en posant

$$\mathcal{J} = \sum Ei$$

ce qui s'écrit :

$$\mathcal{J} = \sum \alpha^T v_i = \alpha^T \sum v_i$$

Il suffit désormais pour déterminer la valeur de α minimisant $\mathcal{J}$ d'appliquer l'algorithme au vecteur $\vec{v} = \sum \vec{v}_1$.

La méthode opératoire est la suivante : après avoir enregistré un certain nombre de mots prononcés par un ou plusieurs locuteurs, on isole sur la bande numérisée, au moyens de marqueurs, les zones "stables" des phonèmes recherchés. Le critère définissant la "stabilité" est subjectif et visuel : sont reconnus comme stables les zones du signal où :

- pour les sons non-voisés, la forme du signal se reproduit d'une période à l'autre, sans perturbation autre que la variation de l'amplitude globale.

- pour les sons non-voisés, l'amplitude du signal demeure sensiblement constante.

Sur ces segments, l'autocorrélation du signal est calculée sous une fenêtre de Hammin de largeur 20 ms, déplacée par pas de 10 ms. Sur le vecteur moyen de chaque nuage de vecteurs correspondant à un phonème donné, et jugé homogène, on applique l'algorithme de LEVINSON.

3. APPLICATION A LA PAROLE CONTINUE.

L'ambition du vocodeur à canaux adaptés étant de s'intégrer en tant qu'étage de reconnaissance phonémique à un système complet de reconnaissance de la parole continue, certain points sont à signaler. Le vocodeur délivré à la cadence fixe de 10 ms une valeur énergétique pour chaque "canal adapté" c'est à dire approximativement pour chaque phonème. Il est donc facile de classer les phonèmes à chacun de ces instants par ressemblance décroissante, vis à vis du signal d'entrée. La sortie du vocodeur se présente donc alors comme une suite de vecteurs de reconnaissance phonémique, chacun d'entre eux comportant en tête les phonèmes les plus proches du signal d'entrée durant l'instant considéré. Un test énergétique simple permet de limiter le nombre de ces candidats de façon variable selon les instants de signal.
(cf fig 3 un exemple de sorties du vocodeur)
La sortie du vocodeur à canaux adaptés fournit donc une reconnaissance phonémique à cadence fine de 10 ms, mais sans segmentation.

Afin de transformer ces donnés en un véritable"treillis" de phonème, il faut donc les traiter pour en extraire à la fois l'information sur les limites et sur la nature des phonèmes prononcés. Ceci peut se faire en associant à chaque canal du vocodeur, donc à chaque phonème, un automate dont la tâche est de suivre instant par instant la confiance accordée au phonème correspondant.
La tâche de chaque automate est de fournir une décision de segmentation reconnaissance quand l'évolution de la confiance du phonème qu'il contrôle indique que celui-ci peut-être retenu comme candidat. La série d'entrée typique à laquelle doit répondre un automate associé à une voyelle peut-être shématisée selon la fig 4 : mémorisation de la transition de départ, calcul d'une confiance globale si le phonème se maintient durant assez longtemps avec une confiance assez grande, et décision de frontière finale sur une chute de la confiance accordée au phonème (transition de fin).

A chaque phonème peut-être affecté un automate spécialisé, sans que le temps de décodage augmente. Mais un petit nombre d'automates différents semble suffisant pour assurer des décisions correctes canal par canal. Chaque automate de décodage fournit donc, au fur et à mesure de l'entrée, des décisions sur les frontières de l'apparition du phonème qui lui correspond. L'ensemble de ces informations est regroupé, confronté aux indications de segmentation fournies par l'évolution de l'énergie du signal, et la sortie de l'étage de reconnaissance phonémique ainsi constitué est un classique "treillis phonémique", à partir duquel sont entreprises les recherches lexicales en vue du décoldage en mots de la phrase prononcée.

Lorsque l'apprentissage a été réalisé pour un seul locuteur, le vocodeur est assez sensible à un changement de locuteur, surtout en ce qui concerne les voyelles. Il est donc nécessaire de réaliser l'adaptation de ce vocodeur au locuteur. Plusieurs méthodes sont en cours d'expérimentation.

1ère METHODE.

Déterminer pour chaque locuteur une transformation multilinéaire sur le vecteur $\vec{v}$ d'autocorrélation ramenant de façon globale les axes d'inertie du nuage des $\vec{v}$ en coincidance avec ceux caractérisant le locuteur standard.

$$v' = Av \qquad E_i = \alpha_i' Tv'$$

2ème METHODE.

Déterminer les transformations analogues, non plus da façon globale, mais pour chaque phonème séparément. Cette adaptation serait assez longue et coûteuse, aussi pourrait-elle céder la place à un compromis entre les deux premières méthodes.

3ème METHODE.

Le nuage des occurance est partitionné automatiquement en k classes par une procédure du type "nuées dynamiques" et la transformation ajuste les axes d'inertie de ces k classes avec ceux des k classes du locuteur standard.

4ème METHODE.

L'apprentissage a été effectué au préalable pour n locuteurs représentatifs au mieux des diverses voix. L'adaptation à un locuteur nouveau se fait par classement de ce locuteur dans l'un des n types de voix (types considérés uniquement sur le plan des diverses formes du nuages des occurences de $\vec{v}_i$ et indépendemment de tout autre critère acoustique). On introduit ensuite dans les vocodeurs les paramètres α_i correspondant au locuteur standard de ce type de voix.

4. AVANTAGES DU VOCODEUR A CANAUX ADAPTES.

Les performances du vocodeur à canaux sont qualifiables objectivement de "très bonne". Un critère objectif de qualité ne peut-être mis en évidence, car la "valeur" d'une chaine de candidats en sortie du vocodeur dépend certes de la présence avec un bon taux de confiance du phonème réel en tant que candidat. Mais la valeur de la chaîne dépend aussi de sa cohérence avec les chaînes voisines, ou au contraire de sa non-cohérence s'il s'agit d'éliminer des candidats indésirables.

C'est en fait le système d'analyse placé en aval du vocodeur qui permet par ses résultats de juger de la qualité du vocodeur, et de son adéquation au système d'analyse.
Nous nous en tiendrons donc à l'examen visuel des chaînes de candidats pour évaluer subjectivement la confiance à accorder à ce vocodeur.
La réalisation du vocodeur est assez simple, puisqu'il se décompose en un corrélateur, suivi d'un banc de discriminateurs linéaires (produit scalaire) et de divers automates (classement des candidats-phonèmes, segmentation ...). Une réalisation hardware du corrélateur et des discriminateurs assure un fonctionnement en temps réel, le temps de calcul des automates étant réduit.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

INTERACTIONS ENTRE LES NIVEAUX LEXICAL, SYNTAXIQUE ET SEMANTIQUE EN RECONNAISSANCE DE LA PAROLE CONTINUE

J.P. HATON ET J.M. PIERREL
Laboratoire d'Informatique
Université de NANCY I
C.O. 140 54037 NANCY CEDEX

RESUME

Un système de reconnaissance/compréhension du discours continu apparaît comme un ensemble hiérarchisé complexe de processeurs très raffinés, chacun s'attachant à un aspect particulier du problème. Le but du procédé est d'extraire le maximum d'informations aux différents niveaux de traitement de façon à mener à bien la compréhension d'une phrase.

Une intéraction efficace entre les niveaux (acoustique, morphologique, syntaxique et sémantique) est une condition nécessaire au bon fonctionnement.

Nous présentons ici quelques idées directrices dans ce domaine, en nous fondant sur des exemples tirés du système étudié à Nancy.

SUMMARY

An automatic recognition/understanding system for continuous speech is a highly complex, hierarchical set of processing levels. Each level has been designed for processing a particular aspect of the problem.

Efficient cooperation and interaction between these levels are necessary in order to achieve the recognition task.

We present here some basic ideas in this field, illustrated by examples form the speech understanding system which is being implemented in Nancy.

INTERACTIONS ENTRE LES NIVEAUX LEXICAL, SYNTAXIQUE ET SEMANTIQUE EN RECONNAISSANCE DE LA PAROLE CONTINUE

J.P. HATON ET J.M. PIERREL
Laboratoire d'Informatique
Université de NANCY I
C. O. 140 54037 NANCY CEDEX

INTRODUCTION

Il y a quelques années, la plupart des recherches en reconnaissance automatique de la parole concernait l'identification de mots isolés extraits d'un vocabulaire donné. Actuellement beaucoup de chercheurs se tournent vers la reconnaissance et la compréhension du discours continu et, lors de la mise en place de tels systèmes, tous ont été amenés à considérer divers niveaux de reconnaissance : niveau acoustique, lexical syntaxique et sémantique [1], [2], [3], [4].

Il semble que de tels systèmes doivent comporter le plus d'interactions possible entre les divers niveaux le composant afin de permettre par exemple à l'un de ces niveaux de corriger les erreurs résultant d'un autre niveau. Si tout le monde est d'accord pour reconnaître la nécessité et l'utilité de telles interactions, il faut pourtant noter que jusqu'alors nous ne savons guère comment les réaliser. Beaucoup d'auteurs notent simplement leur intérêt en indiquant que de telles réalisations seront étudiées ultérieurement. Ceci provient du fait que l'étude de ces divers points est peut être l'une des plus difficile à mener dans le cadre d'un système de reconnaissance de la parole. Sans vouloir prétendre résoudre ce vaste problème, nous présentons dans cet article les diverses interactions mises en oeuvre dans le système que nous avons réalisé à Nancy [5].

Après une rapide description des différents niveaux composant notre système, nous aborderons successivement la prise en compte des informations acoustiques et les liaisons entre les niveaux lexico-

graphique et syntaxique d'une part, syntaxique et sémantique d'autre part. Nous terminerons enfin en présentant quelques résultats obtenus actuellement.

II DESCRIPTION SUCCINTE DES DIFFERENTS NIVEAUX

Dans le système que nous avons mis en place à Nancy on peut distinguer quatre niveaux différents de reconnaissance : acoustique, syntaxique, lexical et sémantique. Nous nous limitons ici à une description très sommaire, pour avoir plus de détails, on pourra se reporter à [5].

a) le niveau acoustique

Il correspond à un premier traitement du signal vocal : les données prises en compte sont celles fournies à la sortie du micro et les résultats obtenus par ce niveau peuvent être essentiellement de deux types :

- soit une représentation paramétrée de la phrase prononcée (sous forme de spectres ou de coefficients de prédiction)
- soit une représentation phonémique (sous forme de chaîne de phonèmes ou de syllabes).

Actuellement nous n'utilisons que la représentation phonémique sous forme de pseudo-chaîne de phonèmes obtenue après segmentation de la phrase, identification des segments et lissage de la chaîne résultat [6].

b) le niveau syntaxique

Il a pour donnée la grammaire du langage à reconnaître (grammaire de langage à contexte libre) [7] et fourni comme résultat des hypothèses sur les prochaines unités morphologiques pouvant apparaître dans la phrase à reconnaître. Ce niveau correspond pour nous au noyau du système de reconnaissance, il est réalisé grâce à un algorithme d'analyse syntaxique descendante travaillant de la gauche vers la droite qui est considéré comme le guide de l'ensemble du processus de reconnaissance.

c) le niveau lexical ou morphologique

Ce niveau a pour donnée la liste des hypothèses émises par l'analyseur syntaxique, et fournit comme résultat un point de reprise pour

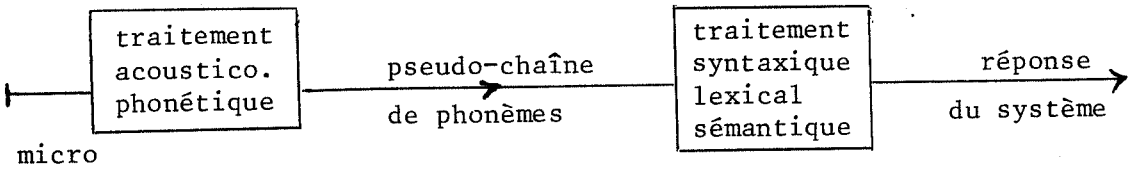
l'analyseur syntaxique correspondant à la meilleure hypothèse validée grâce à la reconnaissance phonémique. Son rôle est donc de tester successivement l'ensemble des unités morphologiques sélectionnées par le niveau syntaxique et de les classer suivant leur score de reconnaissance fourni par un module de reconnaissance analytique de mots.

d) le niveau sémantique

Il correspond à la partie la moins structurée du système et est réalisé par une procédure de dialogue qui a la charge de valider en dernier recours les résultats fournis par la reconnaissance, de lever les ambiguités ou corriger certaines erreurs. Actuellement nous ne savons pas comment générer automatiquement une telle phrase de dialogue pour des applications différentes, nous avons donc considéré que ce niveau était propre à chaque application.

III PRISE EN COMPTE DES INFORMATIONS ACOUSTIQUES ET PHONETIQUES

Lors de l'élaboration et de la mise en oeuvre du système de reconnaissance, nous avons été amené, dans un premier temps, à nettement distinguer le niveau acoustique des niveaux syntaxique, lexical et sémantique. Ceci correspond à une première étape pour laquelle on peut schématiser ainsi le système :



Néanmoins, pour tenir compte des limites et des erreurs possibles dues au niveau acoustico-phonétique et permettre quelques interactions entre ce dernier et le reste du système, il nous a fallu généraliser quelque peu la prise en compte des informations acoustiques et phonétiques tant en ce qui concerne le résultat du traitement acoustico-phonétique que les chaînes de références servant au niveau lexical.

a) résultat du traitement acoustico-phonétique

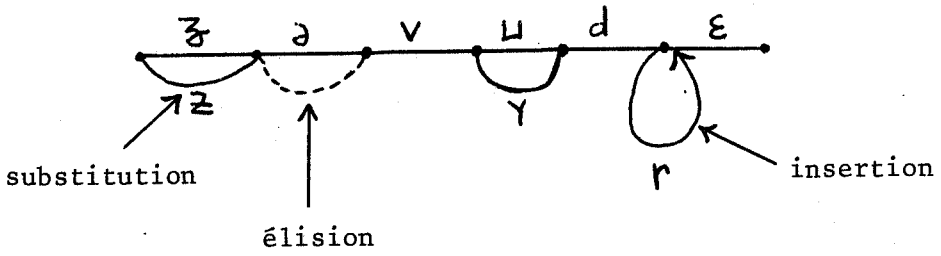
En toute rigueur, le résultat du traitement acoustico-phonétique consiste en une chaîne de phonèmes représentant l'énoncé oral de départ Cette transcription phonétique est en fait fortement entachée d'erreurs, les plus fréquentes étant des confusions entre phonèmes de même type par exemple / p / , / t / , / k / ou / b / , / d / , / g / ou encore / z / , / ʒ / , / v /. Pour tenir compte de ces erreurs de substitution et pouvoir par la suite remettre en cause le choix du phonème fait par le niveau acoustique, nous ne considérons pas seulement un phonème mais les k phonèmes les plus problables fournis lors de l'identification. Le nombre k peut être soit un nombre déterminé suivant le poids acoustique des différents phonèmes soit un nombre fixe, paramètre du système. Quelle que soit la solution adoptée, ceci augmente la complexité du module de reconnaissance de mots mais aussi ses chances de succès, car la probabilité d'avoir le bon phonème est alors bien plus importante. Dans notre système le nombre de phonèmes pris en compte a été limité à trois, ce qui semble un bon compromis entre la validité de l'information obtenue et la possibilité pour l'algorithme de recherche lexicale de travailler en temps réel.

b) définition des chaînes phonémiques de référence

Lors de la reconnaissance, on compare la pseudo-chaîne phonémique obtenue comme résultat du niveau acoustique aux chaînes de références stockées dans le lexique. Il est donc intéressant de tenir compte au niveau du lexique des variantes phonétiques, dues aux diverses prononciations, et des erreurs quasi-systèmatiques du niveau acoustique. Pour ce faire, on peut par apprentissage (éventuellement automatisé) privilégier la représentation lexicale qui a le plus de chance d'apparaître comme résultat du traitement acoustico-phonétique et indiquer les autres variantes intéressantes au niveau phonémique sous forme soit de substitution, soit d'élision, soit d'insertion.
Prenons comme exemple l'unité morphologique "je voudrais".
La représentation lexicale stricte devrait être la chaîne phonémique

" ʒ ə v u d r ɛ "

Si l'on considère que le système acoustique a une tendance quasi systèmatique d'élider les / r / qui suivent une autre consonne il serait bon de bien reconnaître la chaîne " ʒ ə v ʉ d ɛ ".
Sur ce même exemple on peut aussi tenir compte des variantes phonémiques provenant de diverses prononciations : élision du /ə/ muet, substitution du /ʒ/ par / z / ou du /ʉ/ par / y /.
Pour rendre compte de toutes ces variantes au niveau lexical, nous avons opté de définir une unité lexicale non pas sous forme d'une chaîne phonémique à une dimension mais sous forme d'une sorte de graphe permettant d'indiquer pour chaque position phonémique les insertions, substitutions ou élisions possibles. Le graphe de référence correspondant dans le lexique à l'unité morphologique "je voudrais" est alors :



Remarque :
Lors de la reconnaissance le chemin correspondant dans ce schéma à la ligne droite sera privilégié par rapport aux autres chemins, il doit correspondre à la prononciation la plus fréquente.

IV LIAISON LEXIQUE - SYNTAXE

Mis à part le niveau acoustique, l'ensemble du système repose sur le niveau syntaxique [5] ; l'analyseur syntaxique a un rôle très privilégié en ce sens qu'il émet les hypothèses qui sont ensuite reprises par le niveau lexical. Néanmoins nous avons donné une grande importance au niveau lexical qui a pour but de réduire l'indéterminisme rencontré au niveau syntaxique tant pour la grammaire que pour la définition des terminaux : l'analyseur syntaxique guide la reconnaissance mais c'est le niveau lexical qui effectue les choix en fonction

de la chaîne phonémique restant à explorer. Pour ce faire, il tente la reconnaissance des hypothèses émises au niveau syntaxique et en sélectionne une en fonction du score de reconnaissance affecté aux divers terminaux testés. Une méthode heuristique de calcul nous permet de déterminer pour chaque hypothèse testée un score de réussite S variant de 0 à 1

Pour un terminal k : $S_k = 1 - \frac{TD_I}{I}$

où TD_I est le taux de dissemblance phonémique cumulé pour les I phonèmes traités lors de la reconnaissance du terminal k.

Si $S_k <$ SEUIL 1 (SEUIL 1 fixé par expérience) on rejette l'hypothèse correspondant à ce terminal.

TD_I est donné par une formule de récurrence très simple :

$$TD_0 = 0$$

$$TD_I = TD_{I-1} + TDP_I$$

TDP_I étant un taux de dissemblance calculé pour le phonème I en fonction de la chaîne à reconnaître et de la chaîne de référence compte tenu des hypothèses d'insertion, de substitution et d'élision possibles.
De plus, on calcule pour le contexte gauche de la phrase déjà traitée un score de réussite SC cumul pondéré des scores S_k affectés aux k terminaux pris en compte. Si SC devient inférieur à un certain seuil on décide alors d'abandonner la reconnaissance pour la construction syntaxique en cours.

Outre la possibilité de"censurer" le niveau syntaxique, nous avons permis au niveau lexical de prendre des initiatives particulières dans le cas où toutes les hypothèses émises par l'analyseur syntaxique se trouvent infirmées par la reconnaissance. Un tel cas peut se produire spécialement lorsqu'il y a dans la phrase insertion d'un mot parasite, substitution d'un mot par un synonime non prévu dans la grammaire ou élision d'un mot. Si les mots considérés ne sont pas sémantiquement caractéristiques, le niveau lexical émet de nouvelles hypothèses qu'il tente ensuite de valider par reconnaissance : la

chaîne phonémique correspondant au mot recherché peut alors être précédée d'une sous-suite phonémique plus ou moins longue, dans le cas d'une substitution ou d'une insertion possible. Il s'agit donc de sélectionner un mot dans la chaîne restant à traiter et non de la reconnaître en tête. Pour déterminer les nouvelles hypothèses, le niveau lexical adjoint aux anciennes hypothèses H 1 des hypothèses supplémentaires H 2 obtenues après analyse syntaxique d'un pas effectuée en considérant comme bien reconnu les terminaux de H 1 non caractéristiques au niveau sémantique.

On obtient alors comme algorithme de détermination de ces nouvelles hypothèses :

H 1 : donné , ensemble des anciennes hypothèses

C : donné , ensemble des mots caractéristiques au niveau sémantique

H 3 = H 1 $\cap$ C

H 2 = $\emptyset$

Pour chaque $e_i \in$ H 3 considéré comme reconnu faire :

> analyse syntaxique d'un pas et détermination d'hypothèses supplémentaires Hi
>
> H 2 = H 2 $\cup$ Hi

NH = H 1 $\cup$ H 2 NH : nouvelles hypothèses émises par le niveau lexical

Un tel processus permet au niveau lexical de prendre certaines libertés par rapport à la grammaire du langage à traiter ; la grammaire ne donnant alors que la structure du noyau du langage accepté par le système. Il existe donc dans notre système des liaisons très étroites entre les deux niveaux, syntaxique et lexical, néanmoins il est possible de donnet encore plus d'importance au lexique par exemple en y incluant des catégories syntaxiques ou encore en fondant l'ensemble du système sur le niveau lexical. Il s'agit ici d'un choix à faire au niveau de la structure du système, pour notre part nous avons choisi de privilégier le niveau syntaxique sans pour autant trop amoindrir le niveau lexical.

V ANALYSE SYNTAXIQUE OU/ET SEMANTIQUE

Le rôle du niveau sémantique est de valider en dernier recours les résultats fournis par le niveau syntaxique en vérifaint que la phrase reconnue possède bien un sens dans le contexte de l'application en cours. Actuellement dans notre système le traitement sémantique a été réalisé grâce à une procédure de dialogue propre à chaque application : elle utilise les résultats de la reconnaissance obtenus par les niveaux syntaxique et lexical et s'efforce de lever les ambiguités et de traiter les erreurs. Dans ce cas l'analyse sémantique et l'analyse syntaxique se suivent ; néanmoins, lors du dialogue le niveau sémantique a une action sur l'analyseur syntaxique qui s'attend à une certaine structure de phrase en réponse à un certain type de question.

De plus comme nous l'avons vu au paragraphe IV certaines informations sémantiques sont utilisées par le niveau lexical lors de l'émission de nouvelles hypothèses, mais actuellement cela correspond aux seules interactions entre le traitement sémantique et les autre niveaux. Une telle réalisation, loin d'être idéale, correspond à une première approche et actuellement nos recherches s'orientent vers l'étude d'un système mixte où analyse syntaxique et analyse sémantique travailleraient en parallèle. A ce stade on se heurte à des difficultés importantes pour caractériser sémantiquement un langage. Jusqu'alors peu d'outils informatiques on été développés en ce sens et les solutions proposées sont le plus souvent spécifiques aux langages de programmation (caractérisation par attribut ou par double grammaire). Il est donc nécessaire de rechercher de nouveaux outils permettant une caractérisation sémantique de langage, de type langue naturelle, orientée vers la reconnaissance. De tels travaux devraient ouvrir ainsi la voie à une meilleure formalisation de l'ensemble du traitement syntaxico-sémantique nécessaire à la compréhension de la parole et à l'établissement d'un dialogue oral fructueux entre l'utilisateur et la machine.

VI RESULTATS ACTUELS

Jusqu'à présent nous avons implémenté le système syntaxico-sémantique sur le CII IRIS 80 de l'I.U.C.A. de Nancy. Les données en entrée sont formées de suites phonémiques tenant compte des performances possibles du système acoustique.

La figure A donne un exemple de listing obtenu à partir d'un télétype branché sur l'IRIS 80 : on y voit les chaînes de phonémes rentrées en donnée et le dialogue engagé avec la machine. Le système est ainsi capable de traiter en moins de 0,5 s une communication nécessitant la reconnaissance de 2 phrases de 3 à 12 mots [5].

De plus, les essais que nous avons effectués nous ont prouvé le bon fonctionnement de l'analyseur lexicographique spécialement pour ce qui est des hypothèses supplémentaires émises lors de la mauvaise reconnaissance des hypothèses faites par le niveau syntaxique. A titre d'exemple voici le chemin suivi pour la reconnaissance de la phrase "j'voudrais 1'poste 339" :

ʒ ɯ d ɛ l p ɔ s ə p w s â t â n ə f
v ʏ b i ʁ t o ʃ a t o ʃ ô p ô m ʏ s
ʒ o g e e k a e k f ɛ̂ k ɛ̂ d ʃ

(la phrase syntaxiquement correcte la plus proche était :
"je voudrais avoir le poste 339"
compte tenu de la grammaire utilisée [5] , [7])

1. analyse syntaxique :
   hypoyhèses émises : "allo, monsieur, madame, mademoiselle, le, pourrais-je, je voudrais, passez-moi, est-ce que je pourrais.

2. analyse lexicographique :
   hypothèse validée : "je voudrais" reconnu entre les phonèmes 1à5 , score de 0,7

3. analyse syntaxique
   hypothèses émises : avoir, parler à

```
STANDARD AUTØMATIQUE DE ****,JE VØUS ECØUTE
? L   R
?EU   E
? P   T   K
?AU   Ø   A
? S CH    F
? P   T   K
?EU   A   E
? T   P   K
?ØI   AU
? S CH    F
?AN   ØN  IN
? T   P   K
?AN   IN  ØN
? N   M
?EU   U
? F   S CH
?

VØUS AVEZ BIEN DEMANDE LE PØSTE  339?
? U ØU
?AI   I   E
? L   M   B
? E   I
?CH   S   F
? I AI
?AI
?

LE PØSTE  339 EST ØCCUPE.VØULEZ-VØUS PATIENTER?
? N   M
?AN
?

FIN DE CØMMUNICATIØN

STANDARD AUTØMATIQUE DE ****,JE VØUS ECØUTE
? B   D   M
? A EU    E
? F   S CH
? I   S   E
? A EU    E
? P   T   K
? I   E
? E EU    I
? L   R   E
?

VØUS AVEZ BIEN DEMANDE MR.  PIERREL ?
?ØU AU
? E   I
?
LE PØSTE DE MR.  PIERREL  EST ØCCUPE.VØULEZ-VØUS PATIENTER?
?ØU AU  Ø
? E   I
? B   N   M
?AI   E
? F   S CH
? E I \\ I AI
?

LIGNE EN ATTENTE
```

Phrase prononcée :
"le poste 339"

reponse :
"Oui merci"

réponse :
"Non"

Phrase prononcée
"Monsieur Pierrel"

réponse : "oui"

réponse : "oui merci"

Figure A

4. analyse lexicographique :
   hypothèse validée : aucune
   hypothèses supplémentaire : avoir, parler à, monsieur, madame, mademoiselle, le
   hypothèse validée : "le" reconnu entre les phonèmes 5 et 6 score de 0,9.

5. analyse syntaxique :
   hypothèses émises : poste, numéro, bureau

6. analyse lexicographique :
   hypothèse validée : "poste" reconnu entre les phonèmes 6 et 9 score de 0,8"

7. analyse syntaxique
   hypothèses émises : de, numéro de l'annuaire

8. analyse lexicographique
   hypothèse validée : numéro de l'annuaire : 339 reconnu entre les phonèmes 9 et 18, score de 0,97

9. analyse syntaxique
   condition de fin vérifiée :phrase reconnue équivalente à "je voudrais avoir le poste 339"
   l'analyseur lexicographique a replacé le terminal "avoir" inexistant dans la phrase orale de départ.

VII CONCLUSION

Nous présentons dans cet article les diverses interactions entre les niveaux acoustique, syntaxique, lexical et sémantique mises en oeuvre dans le système réalisé au laboratoire d'informatique de Nancy. Au cours de cette présentation nous avons indiqué les difficultés rencontrées spécialement au niveau sémantique. Notre travail s'oriente plus dans ce sens actuellement.

En fait, nous n'avons pas encore abordé tous les aspects de ce problème une autre poursuite intéressante concerne la mise en place de liens plus étroits entre le niveau acoustique et le reste du système. Nous

pensons surtout à la possibilité de remettre en cause les informations "acoustiques" au vu des résultats fournis par les autres niveaux.

Néanmoins un des résultats importants que nous avons obtenu concerne la liaison lexique-syntaxe. Elle nous permet de prendre une certaine liberté par rapport à la grammaire du langage à traiter. Nous pensons ainsi pouvoir traiter des langages de plus en plus importants et non figés par leur syntaxe en utilisant une grammaire noyau décrivant la structure syntaxique profonde d'un tel langage.

[1] W.R. LESSER, R.D. FENNEL, L.D. ERMAN, D.R. REEDY
"organisation of the HEARSAY II speech understanding system"
I.E.E.E. Symposium of speach recognition CMU 1974

[2] J.W. KLOWSTAT, L.F. MONSHEIM
"The Caspers linguistic analysis speech"
I.E.E.E. Symposium of speech recognition CMU 1974

[3] P.L. MILLER
"A locally organized parser for spoken input"
Communication of the ACM vol. 17 n° 11 1975

[4] J.F. BAKER
"The dragon system : an overwiew "
I.E.E.E. Symposium of speech recognition CMU 1974

[5] J.M. PIERREL
"Contribution à la reconnaissance automatique du discours continu, mise en oeuvre d'un système paramétrable aux niveaux morphologique syntaxique et sémantique"
Thèse de spécialité NANCY 1975

[6] J.P. HATON
"Acoustic segmentation and real time recognition of speech"
8ème Congrès International d'Acoustique, Londres, juillet 1974

[7] J.M. PIERREL, J.P. HATON
"Une approche syntaxique de reconnaissance de phrase dans un contexte donné"
6ème JEP, GALF, Toulouse, mai 1975

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

UN ANALYSEUR SYNTAXIQUE ADAPTE A LA RECONNAISSANCE DE LA PAROLE

Par
P. QUINTON
CNET - LANNION

RESUME :

On décrit l'analyseur syntaxique utilisé dans le système KEAL pour la reconnaissance de la parole continue. A partir des mots localisés dans la phrase, l'analyseur construit toutes les structures syntaxiques possibles, en accord avec une grammaire hors contexte qui a été préalablement définie à l'aide d'un métalangage compilé. Certaines erreurs telles que l'omission ou l'insertion de phonèmes par l'analyseur phonétique, ou la non détection d'un mot court par l'analyseur lexical peuvent être rattrapées au cours de l'analyse syntaxique.

Dans son état actuel, ce programme permet la reconnaissance de 50 à 60 % des phrases extraites de dialogues simples. Il ne nécessite que quelques secondes pour reconnaître une phrase.

SUMMARY :

We describe the syntactic analyzer which is used in the system KEAL for continuous speech recognition. After the words in an utterance have been detected by a lexical analyzer, our syntactic analyzer builds all the possible syntactic structures according to a context free grammar previously defined by means of a compiled metalanguage. This analyzer allows, in some cases, to correct some errors such as omission and insertion of phonemes by the phonemic analyzer, or non-detection of short words by the lexical analyzer.

This program allows presently for the recognition of up to 50-60% of samples of utterances that occur in simple dialogs. It takes only a few seconds to recognize a sentence.

TITRE : UN ANALYSEUR SYNTAXIQUE ADAPTE A LA RECONNAISSANCE DE LA PAROLE (*)

AUTEUR : P. QUINTON

INTRODUCTION.

L'utilisation au cours du processus de reconnaissance de la parole de contraintes telles que la syntaxe ou la sémantique est relativement récente. Elle répond à un double besoin : d'une part, se donner le moyen d'interpréter les phrases prononcées, ce qui est nécessaire lorsqu'on envisage des applications ; d'autre part, orienter la reconnaissance vers l'apport d'informations utiles à la compréhension, afin de limiter les calculs.

La réalisation d'un analyseur syntaxique adapté à la reconnaissance de la parole doit donc répondre à ces deux besoins. La nécessité de donner un sens à ce qui est dit impose l'utilisation d'un modèle des langues naturelles. De tels modèles sont étudiés depuis longtemps en vue de la communication homme-machine et de la traduction automatique. Le fait de travailler sur la parole ne remet pas fondamentalement en cause les résultats acquis, et il est préférable d'adapter plutôt que d'innover. Mais en ce qui concerne le second de ces besoins, il n'en va pas de même : l'incertitude des résultats obtenus par les programmes d'identification des phonèmes ou des mots oblige à utiliser des algorithmes d'analyse nouveaux, capables notamment de tolérer les erreurs.

Nous présentons ici l'analyseur syntaxique du système KEAL. Ce système, actuellement en cours de réalisation au C.N.E.T. à Lannion, permet la compréhension de courtes phrases utilisées pour dialoguer avec une machine. La remière partie de cet article est consacrée au modèle syntaxique que nous utilisons ; la seconde partie décrit l'algorithme d'analyse ; enfin, dans la troisième partie, nous donnons les résultats obtenus jusqu'ici ainsi que les conclusions que nous pouvons en tirer dès maintenant.

I - MODELE SYNTAXIQUE ET DESCRIPTION DU LANGAGE DE REPONSE

1 - Exemple préliminaire

Prenons pour exemple un dialogue destiné à la consultation orale de fichiers bibliographiques. Après avoir sélectionné un ensemble de fiches à l'aide des indications de l'utilisateur, la machine pose la question suivante :

(1) Que voulez-vous faire maintenant ?

à quoi l'utilisateur peut répondre :

	(2)	Je voudrais consulter la fiche numéro 1
ou	(2 bis)	Je voudrais consulter la première fiche
ou	(2 ter)	Je voudrais consulter la fiche suivante
ou	(2 quater)	Je voudrais changer de fichier

etc...

Toutes ces phrases forment ce que nous appellerons le langage de réponse. A chacune correspond une série d'actions que la machine doit exécuter, que nous appellerons "sens" de la phrase. La phrase (2) comme la phrase (2 bis) ont le sens "synthétiser le contenu de la fiche numéro un".

Le passage de la phrase (2) à son sens se fait en deux étapes : on construit d'abord la structure syntaxique de la phrase à l'aide d'une grammaire hors-contexte ; le sens est ensuite obtenu par transcription de cette structure. C'est à la réalisation de la première de ces étapes qu'est destiné l'analyseur syntaxique, la seconde étant du ressort de la sémantique.

2 - Grammaire hors-contexte

Une grammaire hors-contexte G = (T, N, ::= , S) est constituée d'un vocabulaire terminal T, d'un vocabulaire non terminal N, d'un axiome S appartenant à N, et d'une relation : : = entre N et l'ensemble des chaînes de $V=N \cup T$, que l'on note V*. Chaque couple (A,x) formé d'un symbole de N et d'une chaîne $x = a_1.a_2. \ \ldots \ .a_n$ de V* et satisfaisant A ::=x , est appelé une production de la grammaire. A est la partie gauche de la production et x en est la partie droite. Une production peut être représentée par une arborescence de hauteur 1 (cf. figure 1)

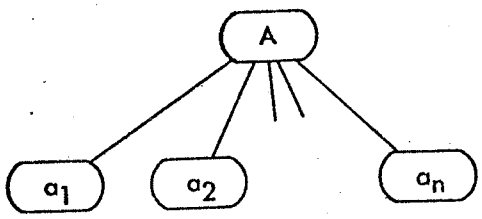


fig. 1 : Représentation de la production $A ::= a_1.a_2. \ \ldots \ .a_n$ sous forme d'une arborescence.

dont la racine est la partie gauche de la production et les feuilles sont la partie droite. En remplaçant dans cette arborescence les éléments non terminaux des feuilles par de nouvelles productions de G, on engendre des arborescences de hauteur 2 et ainsi de suite... Toute arborescence que l'on peut engendrer par ce procédé en partant de l'axiome S de la grammaire s'appelle une structure syntaxique de la grammaire. L'ensemble des chaînes de T qui forment les feuilles d'une structure syntaxique s'appelle le langage engendré par la grammaire. C'est ce modèle, quelque peu étendu, que nous utilisons pour décrire le langage de réponse d'un dialogue.

3 - Description du langage de réponse

L'annexe I donne la description d'une grammaire qui permet d'engendrer la phrase (2 bis) et lui assigne la structure syntaxique donnée par la figure 2.

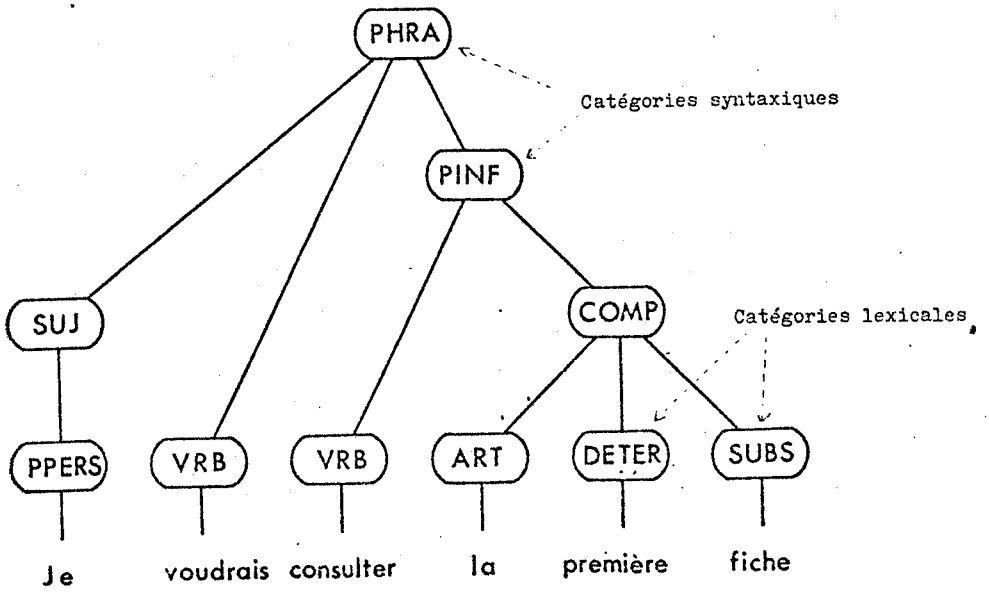


fig. 2 : Structure syntaxique de la phrase (2 bis) .

On trouve dans cette description les classes lexicales [1] qui correspondent au vocabulaire terminal T ; l'ensemble des classes syntaxiques [2], correspondant au vocabulaire non terminal N ; les variables syntaxiques [3], que l'on peut attribuer à toute classe syntaxique ou lexicale pour la subdiviser ; le lexique [4] donnant la forme orthographique de chaque mot ainsi que la valeur de ses variables intrinsèques ; enfin, en [5], les productions de la grammaire accompagnées de restrictions portant sur les variables syntaxiques des classes syntaxiques ou lexicales qui les composent. Ces restrictions permettent par exemple de vérifier l'accord en genre et en nombre d'un article et d'un substantif.

Ce genre de description, assez rudimentaire, permet de traiter la plupart des cas que l'on rencontre dans les applications envisagées actuellement pour la reconnaissance de la parole continue.

II - L'ANALYSEUR SYNTAXIQUE

1 - Les données de l'analyseur

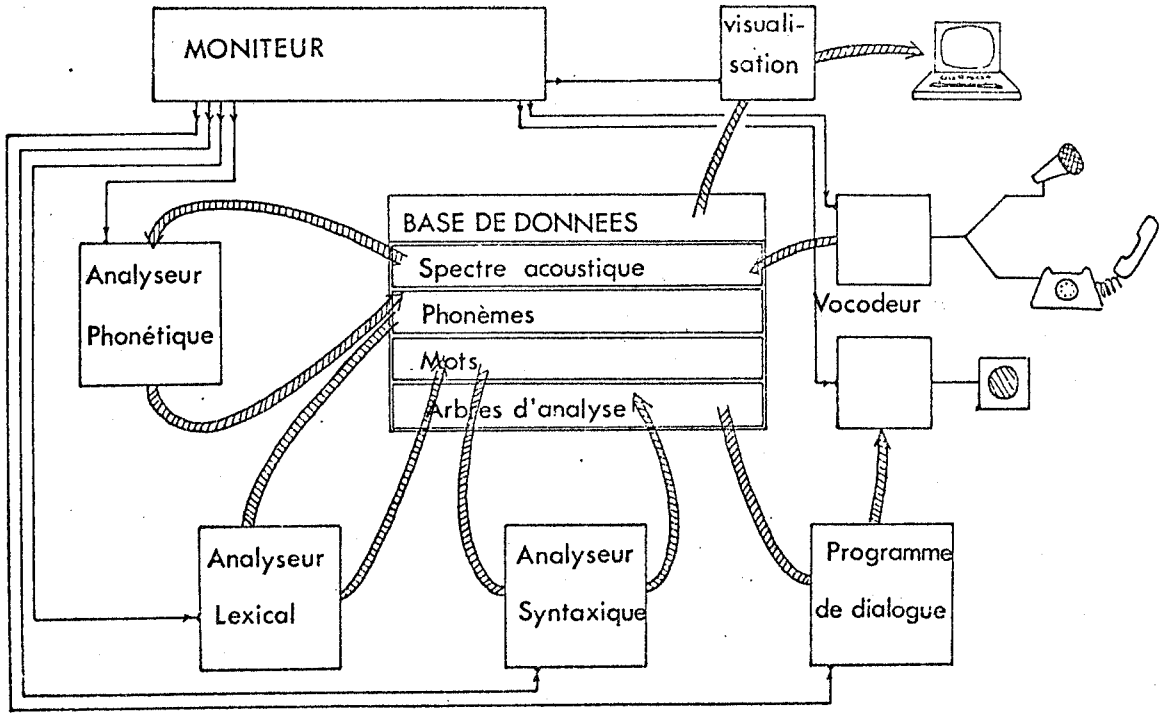


fig. 3 : Organisation générale du système KEAL.

La figure 3 donne l'organisation générale de KEAL. La phrase prononcée par un locuteur est analysée par un vocodeur à canaux qui en donne un spectre numérique. Ce spectre est ensuite traité par un analyseur phonétique[1] qui le segmente en syllabes, localise dans chaque syllabe le segment vocalique et les segments consonantiques, et identifie enfin chaque segment en proposant une liste ordonnée d'étiquettes (voir l'exemple de la figure 4).

Un analyseur lexical[2] permet ensuite de détecter les mots du lexique dans la phrase par comparaison de la transcription phonétique idéale du mot avec le résultat de l'analyse phonétique. Chaque détection est munie d'un score compris entre 0 et 1, qui indique le degré de ressemblance du mot et de la phrase. On appellera occurrence l'ensemble formé par les limites de la détection d'un mot dans la phrase et le score de la détection. L'analyseur syntaxique que l'on décrit ici a ensuite pour fonction de trouver une séquence d'occurrences contigües qui forment une phrase de la gramm ire, ainsi que la structure syntaxique de cette phrase. Cette structure est ensuite interprétée, et l'automate de dialogue décide en fonction du résultat, de la réponse à synthétiser, et de l'orientation future à donner au dialogue.

L'analyseur syntaxique a donc pour données l'ensemble des occurrences trouvées par l'analyseur lexical.

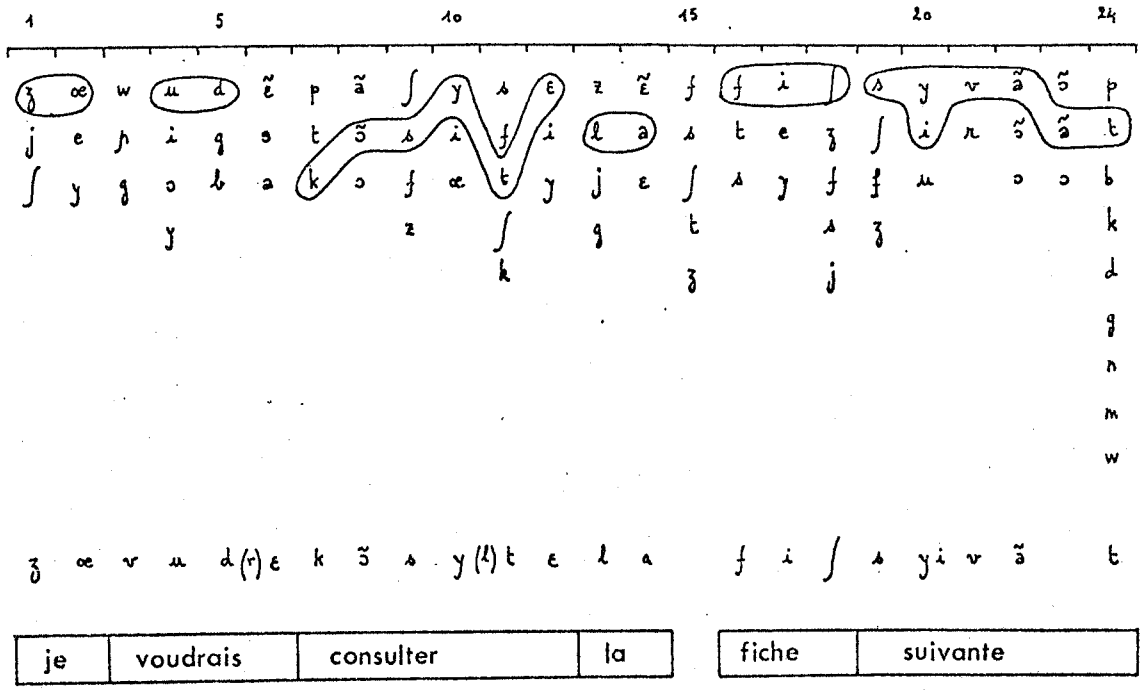


fig. 4 : Résultats de l'analyse phonétique de la phrase (2 ter). Les suites de phonèmes qui ont permis la reconnaissance des mots de la phrase sont entourées.

2 - Principes de l'analyse

Compte tenu du nombre des occurrences (environ 10 à 15 fois le nombre des mots de la phrase pour un vocabulaire de 50 mots), il est important de limiter le plus possible les calculs de l'analyseur. Ceci peut être obtenu de deux façons : d'une part, en utilisant le score des occurrences qui indique la confiance que l'on peut accorder à la détection ; ceci suggère des méthodes de type heuristique évaluant d'abord les meilleures possibilités ; d'autre part, en se servant au mieux des contraintes apportées par la grammaire, ce qui peut se faire à l'aide d'une méthode sélective utilisant les relations induites par la grammaire sur l'ensemble T (relations de précédence, de succession, etc...). La méthode que nous utilisons combine ces deux caractéristiques : l'analyse se fait de la gauche vers la droite, en construisant plusieurs structures simultanément de façon ascendante ; ces solutions sont mémorisées dans une base de données dans laquelle elles peuvent être factorisées, de façon à grouper les structures syntaxiques partielles qui ont le même comportement pour la suite de l'analyse. La stratégie d'analyse et la base de données sont indépendantes, ce qui permet de modifier l'une sans remettre l'autre en cause.

3 - Notations

Soit G = (T,N, ::= ,S) une grammaire hors-contexte. Les règles sont numérotées de 1 à d, et la p^{ième} règle est notée :

$$D_p ::= C_{p1} . C_{p2} . \ \ldots \ . C_{pp^-} .$$

Par souci d'homogénéité des notations, on introduit t règles fictives correspondant aux symboles terminaux, numérotées de d à d+t, de telle sorte que les symboles terminaux soient notés D_q avec $d+1 \leq q \leq d+t$, et $\bar{q} = 0$.

3 - Mémorisation des structures partielles

Poursuivre en parallèle plusieurs analyses implique que l'on puisse mémoriser dans une structure de données les solutions partielles obtenues à un instant donné. Nous utilisons pour ce faire des "états" tels qu'ils sont définis par Earley [3] : un état est un ensemble de structures syntaxiques dont la racine est le début d'une partie droite de règle et dont les feuilles forment une sous chaîne de la phrase à analyser ayant les mêmes frontières à droite et à gauche. Toutes les structures d'un état ont le même comportement vis à vis de l'analyse, ce qui justifie leur groupement. Dans la base de données, un état est représenté par un 5-uplet < p,j,f,l,w> (cf.figure 5).

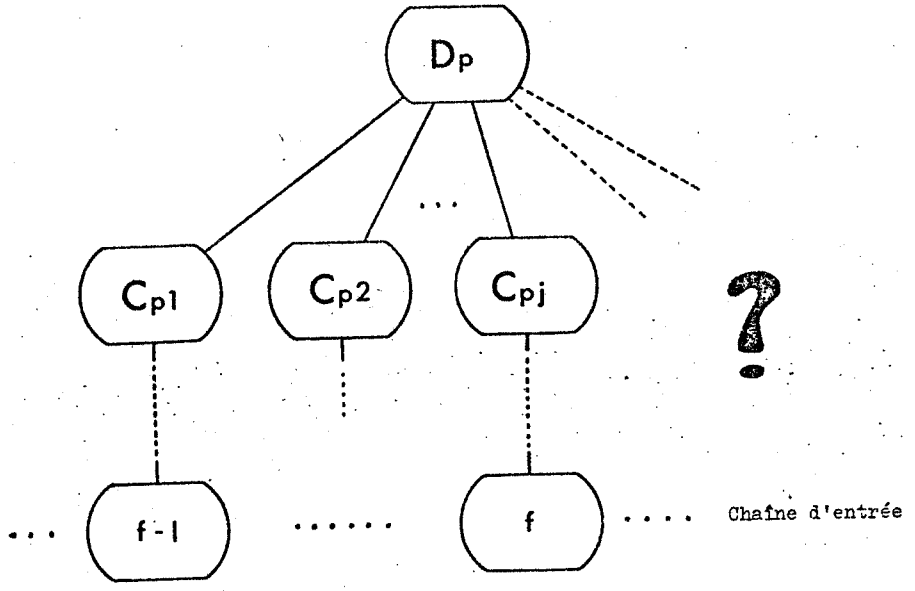


fig. 5 : Représentation d'un état $\langle p, j, f, l, w \rangle$

(on suppose que la grammaire contient la production

$D_p ::= C_{p1}.C_{p2}. \ldots .C_{pj}. \ldots C_{p\bar{p}}.$)

où :

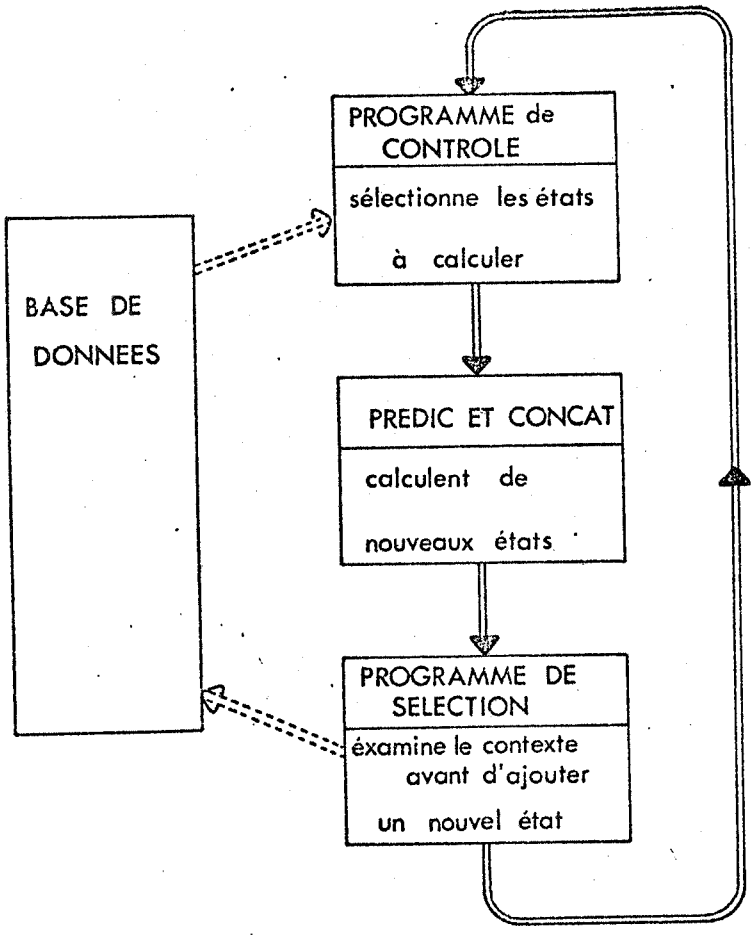
- p est le numéro de la règle,
- j est l'indice du dernier constituant reconnu C_{pj} de la partie droite de la règle,
- f est la limite droite dans la chaîne d'entrée de la sous chaine formée par les feuilles,
- l est la longueur de cette sous-chaîne,
- enfin, w est un nombre compris entre 0 et 1 qui représente le score

de la meilleure structure de l'état. A chaque structure peut en effet être attribué un score qui est calculé à partir des scores des occurrences qui forment ses feuilles, et w est le meilleur.

Si p est supérieur à d, l'état sera dit <u>état lexical</u>, et représente l'occurrence d'un mot de la classe lexicale D_p dont le score est w. Un état est dit <u>final</u> si $j=\bar{p}$, et <u>initial</u> si $j=1$.

Posé en terme d'états, le rôle de l'analyseur est de construire à partir des états lexicaux tous les états utiles pour parvenir à un état final $< p,\bar{p},n,n,w >$ tel que $S=D_p$, qui représente une ou plusieurs solutions complètes de la phrase.

4 - <u>Mécanisme de l'analyse</u>



<u>fig. 6</u> : Schéma de fonctionnement de l'analyseur.

Le schéma de la figure 6 illustre le fonctionnement de l'analyseur composé de trois modules agissant sur la base de donnée. La base de données contient les états décrits plus haut ; elle représente à un instant donné l'état de l'analyse. Elle est au départ initialisée avec les états lexicaux. L'analyseur fonctionne de façon cyclique : un programme de <u>contrôle</u> choisit d'abord les états à calculer ; deux <u>procédures de base</u> permettent ensuite de générer de nouveaux états à partir de ceux qui ont été choisis ; les états ainsi produits sont soumis à un programme de sélection qui décide en fonction du contexte s'ils doivent être ou non ajoutés à la base. Ce cycle est recommencé jusqu'à ce qu'on trouve un état représentant une solution complète, à moins que le programme de contrôle décide qu'il n'y a plus d'espoir d'en trouver.

4-1 - Les procédures de base

Le noyau de l'analyseur est constitué par deux procédures appelées "Prédic" et "Concat", dont le rôle est de faire progresser l'analyse en générant de nouveaux états.

- Prédic : à partir d'un état final $<p,\bar{p},f,l,w>$ et de la grammaire, "Prédic" construit tous les états initiaux $<p',1,f,l,w>$ tels que $D_p = C_{p'1}$. (cf figure 7).

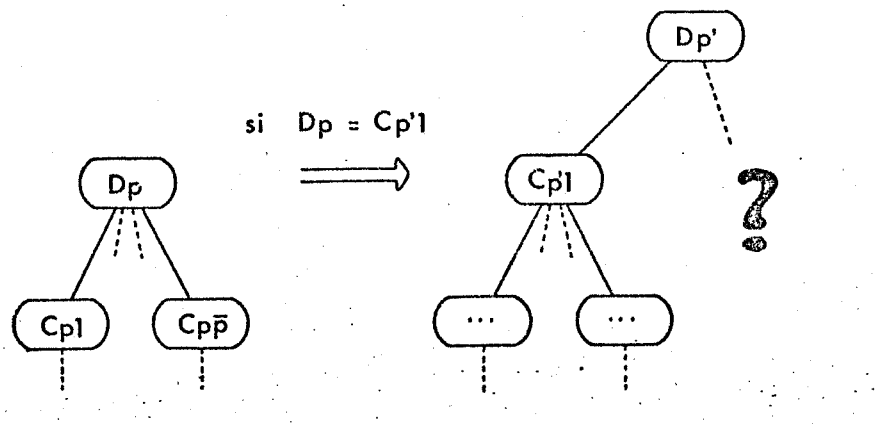


fig. 7 : Illustration du fonctionnement de "Prédic" .

- Concat : à partir d'un état non final $< p,j,f,l,w >$ et d'un état final $<p',\bar{p},f',l',w'>$ et de la grammaire , "Concat" crée l'état $< p, j+1, f', f'-f+l', w''>$ lorsque les conditions suivantes sont réalisées (cf.figure 8) :

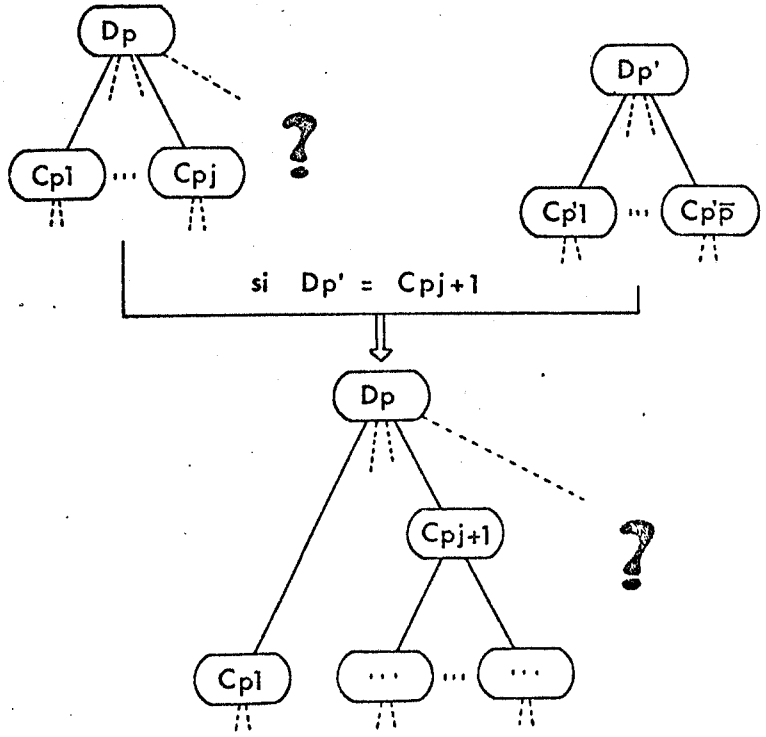


fig. 8 : Illustration du fonctionnement de "Concat".

- $f-\delta_1 \leq f'-l' \leq f-\delta_2$, δ_1 et δ_2 étant respectivement le nombre maximum de phonèmes insérés par erreur par l'analyseur phonétique. Cette condition assure que les deux états sont contigüs.

- $C_{p\ j+1} = D_{p'}$, condition pour que l'état final soit la continuation de l'état non final.

Ces deux procédures permettent ainsi la construction ascendante d'une structure syntaxique par pas élémentaires. Prédic permet la progression en hauteur de la structure, et Concat la progression en largeur.

4-2 - Sélection des états

Un état déterminé par l'une de ces procédures n'est pas immédiatement ajouté à la base de données. On utilise en effet lefait que l'analyse se déroule de gauche à droite pour restreindre le nombre des états calculés. En regardant le contexte droit de l'état, on peut déterminer s'il existe au moins un mot qui a été reconnu par l'analyseur lexical, et qui permet, d'après la grammaire, de poursuivre la solution représentée par l'état. Si ce n'est pas le cas, l'état est rejeté.

De plus, l'état permet de réévaluer le score des occurrences de son contexte : une occurrence est, en effet d'autant "meilleure" que son score est bon, mais aussi qu'elle se trouve dans un bon contexte. Cette réévaluation permet dans certains cas de corriger les effets d'une erreur de reconnaissance phonétique, qui a souvent pour conséquence de diminuer le score de détection d'un mot.

4-3 - Le programme de contrôle

Le programme de contrôle a pour fonction de choisir quels états doivent être calculés par les procédures de base, ce qui revient à décider quelles sont les solutions qui doivent être poursuivies en priorité. Nous avons utilisé successivement deux programmes de contrôle différents :

(a) calcul exhaustif

Cette technique consiste à proposer aux procédures de base tous les états qui se trouvent dans la base de donnée, dans un ordre tel que toutes les solutions soient calculées en parallèle en minimisant les calculs . La seule limitation qui intervient consiste à rejeter les états dont le score est inférieur à un seuil fixé au départ. On obtient ainsi un certain nombre de solutions dont on choisit la meilleure à la fin. L'intérêt de cette méthode provient du fait qu'on est sûr de trouver la meilleure solution.

(b) calcul des meilleurs états d'abord

Le nombre trop grand des états construits par la première méthode nous a incité à essayer une autre stratégie. A chaque fois que l'on ajoute un état à la base, on évalue la "qualité" de l'état en fonction de son contexte. Le programme de contrôle propose aux procédures de base les états qui, à un instant donné, ont la meilleure qualité. Le processus s'arrête dès qu'on trouve un état qui représente une structure complète.

La technique de contrôle (b) donne de très bons résultats : gain de temps, de place, et par conséquent possibilité d'être plus tolérant sur les erreurs. On arrive à trouver ainsi des solutions dans lesquelles un ou plusieurs mots sont manquants. La méthode (a) est par contre la seule valable pour l'analyse de textes, et nous l'avons conservée pour la mise au point de la grammaire et du dialogue.

III - RESULTATS ET DISCUSSION

Ce programme a été testé sur un corpus de 162 phrases. L'annexe II donne les résultats obtenus. Le premier test (expérience 1) portait sur la reconnaissance des nombres de 0 à 99 prononcés normalement dans un ordre aléatoire par un même locuteur. Par la suite, (expériences 2, 3 et 4), un second test a porté sur des phrases extraites de deux dialogues différents : l'un est destiné à la conception assistée par ordinateur (CAO), l'autre à l'interrogation d'un standard téléphonique automatique (STA).

Au total, le pourcentage de reconnaissance parfaite est de 65%. Ce pourcentage passe à 89%, si on lui ajoute celui des phrases reconnues en seconde position, et celui des phrases dont la reconnaissance est incomplète. Ainsi, dans presque 90% des cas, le dialogue pourra tirer profit des résultats de la syntaxe, soit directement lorsque la bonne réponse a été trouvée, soit en faisant rectifier à l'interlocuteur un mot de la réponse.

La nature des erreurs est elle-même encourageante : une analyse détaillée a montré que les erreurs proviennent le plus souvent des mêmes mots qui comportent une séquence de phonèmes difficiles à identifier : /trwa/ dans trois, ou /skri/ de description par exemple. Ces erreurs, difficiles à corriger directement au niveau de l'analyse phonétique, peuvent être évitées indirectement en modifiant le codage "phonétique" de ces mots.

La méthode employée donne donc de bons résultats, compte tenu du taux de reconnaissance phonétique (45% en première position et 80% dans une des solutions données). Le temps de calcul est faible, de 50ms à 3s selon les phrases, ce qui reste compatible avec les exigences d'une réponse en temps réel.

CONCLUSION

L'emploi de la syntaxe et de la sémantique permettent incontestablement un progrès dans la reconnaissance de la parole. On peut espérer bientôt pouvoir dialoguer avec une machine, si dure d'oreille soit elle... Certes, la syntaxe n'est pas la panacée qui permet d'éliminer les difficultés de la reconnaissance acoustique. Il semble néanmoins logique d'intégrer dès maintenant tous les éléments nécessaires à la compréhension : après tout, l'être humain n'est guère capable d'identifier les sons sans utiliser toutes ses facultés de compréhension. On ne voit donc pas pourquoi on empêcherait une machine de se servir des siennes, d'autant plus qu'elles sont, on en conviendra, nettement plus rudimentaires.

(*) Les travaux rapportés ici ont été effectués dans le cadre du contrat SESORI n° 74-80

BIBLIOGRAPHIE

1 - GRESSER J.Y., MERCIER G. : Automatic Segmentation of Speech into syllabic and phonemic units. Application to French words and utterances.
Symposium on "Auditory Analysis and Perception of Speech", Leningrad, August 1973.

2 - VIVES R. : L'analyse lexicale dans le système KEAL pour la reconnaissance de la parole continue.
7èmes journées d'Etude sur la Parole - G.A.L.F. - Nancy, mai 1976

3 - EARLEY J. : An efficient context-free parsing algorithm.
Communications of the A.C.M., Vol 13, Nr.2, 1970.

ANNEXE I : DESCRIPTION D'UN LANGAGE DE REPONSE CONTENANT LA PHRASE (2 BIS).

- [1] CLASSES LEXICALES :
VRB/verbe/,ART/article/,SUBS/substantif/,
DETER/déterminant/,PPERS/pronom personnel/.

- [2] CLASSES SYNTAXIQUES :
PHRA/phrase/,SUJ/sujet/,PINF/proposition infinitive/,COMP/complément/.

- [3] VARIABLES SYNTAXIQUES :
GNR (FEM,MASC)/genre féminin ou masculin/,
NBR (SING,PLUR)/nombre singulier ou pluriel/,
DEF (DEF,INDEF)/défini ou indéfini/,
MODE (IND,COND,INF)/mode indicatif, conditionnel ou infinitif/,
PERS (1,2,3)/personne 1ère, 2de, ou 3ième/ .

- [4] LEXIQUE :
JE=PPERS, PERS=1, NBR=SING, GNR=MASC+FEM;
VOUDRAIS = VRB, PERS=1+2, MODE=COND, NBR=SING;
CONSULTER = VRB, MODE=INF;
LA = ART, DEF = DEF, GNR=FEM, NBR=SING;
FICHE = SUBS, GNR=FEM, NBR=SING ;
PREMIERE = DETER, GNR=FEM, NBR=SING.

- [5] PRODUCTIONS :

R1 : PHRA ::=SUJ.VRB.PINF,
(NBR ET PERS) (SUJ.VRB) NON VIDE
/nombre et personne du sujet et du verbe doivent correspondre/;

R2 : SUJ ::=PPERS ;

R3 : PINF ::= VRB.COMP, MODE (VRB)=INF/le mode du verbe doit être l'infinitif/ ;

R4 : COMP ::= ART.DETER.SUBS,
(NBR et GNR) (ART.DETER.SUBS) NON VIDE
/genre et nombre de l'article, du déterminant et du substantif doivent correspondre/.

ANNEXE II : RESULTATS

Expérience	1	2	3	4	TOTAL
Nature du langage de dialogue.....	Nombres de 0 à 99	CAO* locuteur 1	STA* locuteur 1	STA* locuteur 2	TOTAL
Taille du vocabulaire.....	25	60	59		--
Nombre de règles de la grammaire.....	24	52	61		--
Nombre de phrases prononcées.....	100	24	19	19	162
Nombre de mots prononcés.....	198	112	84	84	478
Nombre de mots reconnus.....	181	97	77	77	432
Pourcentage des mots reconnus.....	91%	86%	91%	91%	90%
Nombre de phrases parfaitement reconnues..	63	16	14	13	106
Pourcentage.....	63%	66,7%	73,6%	68,4%	65,4%
Nombre de phrases reconnues en second....	24	1	1	3	29
Pourcentage.....	24%	4,1%	5,2%	15,7%	17,9%
Nombre de phrases partiellement reconnues** ..	--	5	1	2	8
Pourcentage.....	--	20,8%	5,2%	10,5%	4,4%
Total.....	87	22	16	18	143
Pourcentage.....	87%	91,7%	84,2%	94,7%	88,2%
Nombre de phrases non reconnues.....	13	2	3	1	19
Pourcentage.....	13%	8,3%	15,7%	5,2%	11,2%

* CAO : Conception assistée par ordinateur
ST : Standard téléphonique automatique

** Ce sont les phrases dont on ne connait pas tous les mots, mais dont la structure syntaxique est correcte. On ne les a pas comptées pour les nombres.

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 au 21 Mai 1976

UNE PROCEDURE DE SEGMENTATION AUTOMATIQUE DE LA PAROLE EN MOTS PROSODIQUES, EN FRANCAIS

Jacqueline Vaissière
Laboratoire de Recherche en Electronique de l'Institut de Technologie du Massachusetts (Cambridge, USA)
et Département Conception des Systèmes Electroniques du Centre National d'Etudes des Télécommunications (Lannion, France)

Résumé:

Cette communication propose une procédure de segmentation automatique de la parole continue en mots prosodiques utilisant l'information contenue dans les variations de la fréquence du fondamental (Fo). Cette procédure consiste en quatre étapes. Premièrement, les variations rapides, dues à l'identité des phonèmes sous-jacents (micromélodie) sont supprimées et un certain nombre de valeurs de Fo sont sélectionnées le long des courbes. Deuxièmement, les groupes de phonèmes entre deux pauses sont classés automatiquement en groupes finaux et non finaux, selon la direction du fondamental sur les derniers phonèmes. Troisièmement, les montées et descentes rapides du fondamental à l'intérieur des groupes sont repérées. Quatrièmement, ces mouvements sont interprétés en termes de frontières de mots prosodiques.

Summary:

We will attempt to demonstrate that by taking into account the different factors influencing Fo contours, information about at least the major boundaries inside the groups ,and information about the position of the groups inside the sentences can be obtained directly from calculations of the Fo variations along the sentences in French. The procedure consists in essentially four steps: First, the actively controlled Fo movements are discriminated from the Fo fluctuations due to the identity of the underlying phonemes. Secondly, each group is classified as a final or non final group, depending on Fo variations on its last phonemes. Third the large rises and falls inside the groups are located. Fourth, these selected Fo movements are interpreted in terms of word boundaries.

UNE PROCEDURE DE SEGMENTATION AUTOMATIQUE DE LA PAROLE EN MOTS PROSODIQUES, EN FRANCAIS.

Jacqueline Vaissière

Introduction:

L'idée d'utiliser les paramètres prosodiques dans un système de reconnaissance est relativement récente. Un programme a été réalisé dans ce but pour l'anglais américain par Lea (1), programme que le groupe de recherche sur la reconnaissance de la parole de la compagnie BBN a tenté d'intégrer avec plus ou moins de succès dans leur système général.

Le problème est de définir en premier lieu l'éventuelle contribution apportée par l'intégration des paramètres prosodiques (2) dans un système de reconnaissance. Etant donné que chaque langue possède un système prosodique qui lui est propre, cette contribution est différente selon la langue en question.

Pour le français, deux courants d'opinion semblent s'être dessinés, concrétisés par trois articles présentés lors des dernières Journées d'Etude, à Toulouse (Mai 1975).

Au delà de leurs divergences de vocabulaire et parfois d'opinion, les deux premiers articles, ceux de Di Cristo (3) et de Martin (4) se rejoignent sur un point très précis, à savoir la possibilité d'utiliser l'information contenue par les paramètres prosodiques à la fin des mots lexicaux (c'est-à-dire sur la dernière et l'avant-dernière syllabe) pour reconstituer la structure syntaxique des groupes de mots sous-jacents. La théorie de base sur laquelle repose ces programmes est l'existence en français de contrastes prosodiques portés par la fin des mots lexicaux, indiquant différents degrés de 'continuation' avec les mots suivants (cf la notion de continuation majeure et mineure de Delattre (5).)

Néanmoins nous avons noté trois points qui limitent la portée de tels programmes (dans le sens de leur efficacité dans un système général de reconnaissance). Tout d'abord, dans le cas de phrases non ambiguës (qui est de loin le cas le plus général, les phrases ambiguës étant de rares exceptions) la structure interne des groupes de mots (nous appelons groupe de mots une suite de phonèmes encadrée par deux pauses) n'est pas souvent marquée (cf l'existence de très nombreuses structures prosodiques parallèles relevées par Vaissière (6) dans son analyse de textes lus): l'utilité de tels programmes pourrait donc se limiter à ne traiter que des phrases ambiguës. De plus, il semblerait que même dans le cas de structures syntaxiques ambiguës, tous les français n'emploiraient pas exactement les mêmes procédés pour marquer la structure (cf les contradictions entre les articles de Martin d'une part, et de Di Cristo et Vaissière d'autre part). Enfin, les algorithmes proposés supposent que la segmentation en mots et même la reconnaissance de ces mots ont été réalisées lors d'une étape préalable (puisqu'ils se basent essentiellement

sur l'étude de la dernière syllabe.)

Le rôle de tels programmes devrait donc se limiter à infirmer ou à confirmer des hypothèses faites par une analyse syntaxique traditionnelle préalable: on peut imaginer par exemple qu'à chaque mot reconnu soit affecté un indice de 'continuation' dérivé de l'analyse des paramètres prosodiques, indice qui ne devra pas être en contradiction avec la structure syntaxique trouvée pour la phrase.

Le second courant est représenté par l'article de Buisson et Mercier (7). Ces auteurs ont conduit une étude préliminaire pour évaluer dans quelle mesure l'intégration des paramètres prosodiques pouvait aider à la segmentation des phrases en syntagmes (groupes de mots), et même en mots lexicaux. C'est dans ce même esprit que se situe notre communication d'aujourd'hui.

Nous avons limité notre étude à l'utilisation des pauses et de l'information contenue dans les variations de la fréquence du fondamental au cours de la phrase. Si les autres paramètres (tels que la durée et l'intensité relatives des segments successifs) ne sont pas pris en compte, c'est que l'extraction de ces paramètres est plus délicate et elle nécessite au préalable une segmentation correcte de la phrase en phonèmes et la prise en considération de la nature des phonèmes en séquence (8).

Principe général du programme de segmentation:

1. Nous prenons comme hypothèse de départ que les pauses du signal ont été correctement détectées (ce qui n'est pas un problème trivial.) Nous savons que les phrases sont généralement divisées entre elles par des pauses, et que les phrases longues (ou contenant des signes de ponctuation) sont subdivisées en groupes de mots par l'insertion de nouvelles pauses. La première division opérée est donc une division du continuum en groupes de mots, un groupe de mots pouvant correspondre par définition à une phrase grammaticale entière.

2. Les groupes de mots peuvent être divisés en deux catégories, suivant l'allure des variations du fondamental sur les dernières syllabes de ces groupes (correspondant soit à une intonation montante ou une intonation descendante): les groupes non finaux et les groupes finaux. Le premier point du programme sera donc d'utiliser les variations du fondamental avant la pause pour décider de la catégorie du groupe de mots.

3. Chaque groupe de mots peut être à son tour divisé en mots prosodiques. La frontière entre deux mots prosodiques successifs est marquée par des variations du fondamental d'un sommet vocalique (voyelle) au suivant (percues par l'auditeur comme des variations de hauteur). Ce dernier point a fait l'objet de notre article de l'an dernier et nous invitons le lecteur à s'y référer s'il veut comprendre la présente communication (il ne nous est pas possible, faute de place, de reprendre tous les points de cet article.) Le second point du programme est donc d'utiliser l'information contenue dans les variations du fondamental d'une syllabe à l'autre pour découper les groupes de mots en mots prosodiques.

La figure 1 illustre l'allure générale schématisée d'un groupe de sens non final divisé en deux ou trois mots prosodiques.

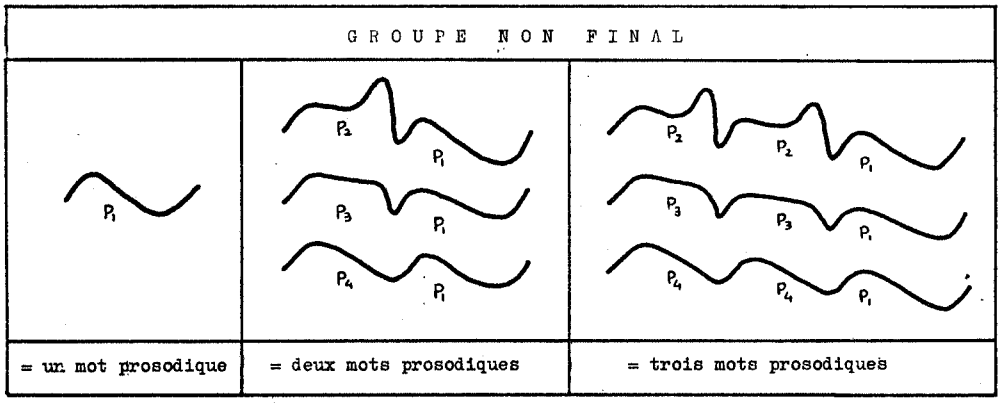


Figure 1: Allure générale d'un groupe de mots non final (terminé par un pattern P1), divisé en deux ou trois mots prosodiques. P2, P3 ou P4 indique le pattern possible des mots prosodiques situés à l'intérieur du groupe de mots. La structure prosodique schématisée pour les groupes de mots décomposés en trois mots prosodiques est une structure prosodique parallèle. Chaque mot prosodique représenté idéalise des mots qui contiennent tous le même nombre de syllabes.

La frontière entre deux mots prosodiques successifs est marquée par une sorte de V, vallée dont les deux versants seront plus ou moins pentus selon les patterns employés pour les mots (le premier versant sera naturellement plus doux, voire plat, si le locuteur a employé P4 pour le premier mot au lieu de P2).

La figure 2 donne un exemple concret de la décomposition d'une phrase en 4 groupes de mots, 3 non finaux et 1 final. Les premier et troisième groupes sont divisés en trois mots prosodiques. La figure 3 illustre un choix différent fait par un second locuteur pour découper le premier groupe de mots (division du groupe en 4 mots au lieu de 3, et répétition de P3 au lieu de P2). Un mot prosodique correspond donc à un ou plusieurs mots lexicaux (par exemple , à 'vingt dernières années' correspond un seul mot prosodique).

On peut donc conclure, en première approximation , qu'un saut important d'une syllabe à l'autre marque le début d'un mot prosodique, alors qu'une chute importante en marque la fin (attribut Ri et attribut Li, respectivement.)

Mais l'algorithme devra prendre en compte les deux phénomènes suivants:

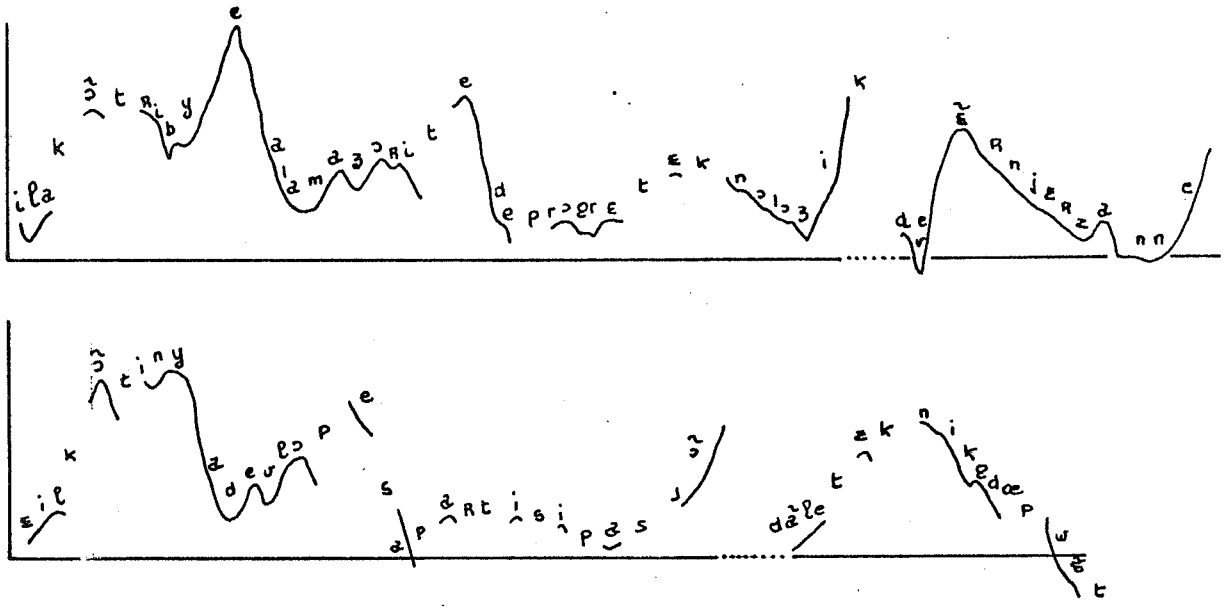


Figure 2: Courbe de fréquence du fondamental de la phrase: " Il a contribué à la majorité des progrès technologiques des vingt dernières années, et il continue à développer sa participation dans les techniques de pointe.", prononcé par un locuteur. Les traits verticaux en pointillé indiquent les frontières entre mots prosodiques (déterminées par inspection visuelle des courbes.)

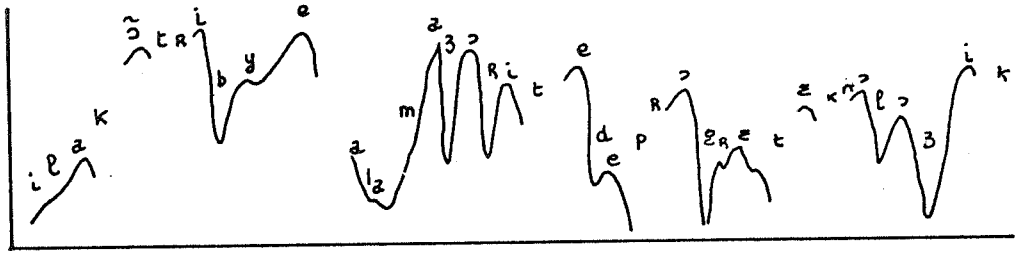


Figure 3: Courbe de fréquence du fondamental du début de la même phrase que précédemment (voir légende fig.2) prononcée par un autre locuteur.

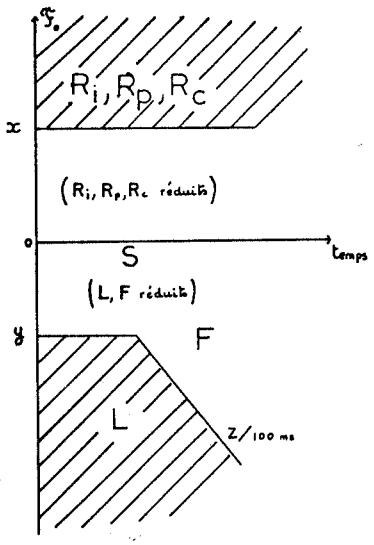


Figure 4: Amplitude des variations de Fo correspondant aux différents attributs. x, y et z représentent les seuils.

Premièrement, l'amplitude des variations du fondamental pour la réalisation des attributs Ri. Rp, Rc, F ou L peut être réduite et être alors équivalente à celle qui peut être observée pour l'attribut S (voir figure 4, ci-contre). En conséquence, afin de ne pas alourdir l'algorithme et multiplier les solutions parallèles, seules les variations de grande importance (zones hachurées) seront prises en compte. Par exemple, pour être considérée comme la réalisation de l'attribut L, une variation devra être plus grande qu'un certain seuil ($\Delta f > y$), et être réalisée rapidement ($\Delta f/t > z/100ms$). De même, pour être susceptible d'être considérée comme la réalisation de l'attribut Ri, la montée du fondamental d'une syllabe à la suivante devra excéder une certaine valeur x.

Deuxièmement, une élévation importante du fondamental peut être interprétée de trois façons différentes (voir zone 1 sur la figure 4). Suivie d'une pause, elle sera considérée comme la réalisation de Rc. A l'intérieur d'un groupe, elle pourra être considérée comme Ri ou Rp, selon qu'elle est suivie ou non d'une descente rapide (voir figure 5).

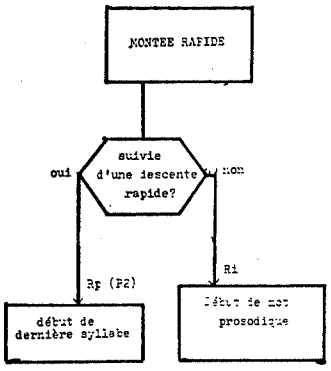


Figure 5: Interprétation d'une élévation du fondamental d'une syllabe à la suivante, en fonction du mouvement du fondamental sur les syllabes suivantes.

Le but de ce programme est donc l'interprétation en termes d'attributs de certains mouvements du fondamental, les mouvements importants et rapides, et il consiste à indiquer l'existence d'une frontière prosodique gauche ou droite d'un mot lexical en certains points très précis du signal. Une reconnaissance automatique des patterns dans leur ensemble ne sera pas tentée et ce programme n'utilise qu'une partie réduite de l'information prosodique contenue dans les variations du fondamental.

4. Le troisième point de ce programme est une élimination automatique des fluctuations du fondamental dues à la micromélodie (variations dans le sens d'une diminution du fondamental durant la production des consonnes.) (9-10). L'élimination s'effectue sur le critère suivant: les variations 'micromélodiques' caractérisent un seul phonème et elles sont plus rapides que les variations contrôlées du fondamental d'une syllabe à l'autre (auxquelles, du reste, elles peuvent se superposer). Pour certains locuteurs (voir fig.3), variations dues aux consonnes et à la réalisation des attributs peuvent être de même amplitude.

Figure 6: Elimination des variations dues à la micromélodie.

Le premier pas est donc de repérer tous les maxima et les minima locaux (voir fig.6), de calculer la déviation Δf entre un maximum et le minimum suivant, t1 la distance entre les deux points précédents, et t2 la distance entre le maximum et le maximum suivant. Si ces trois valeurs, Δf, t1 et t2 sont inférieures aux seuils x', y' et z', respectivement, la vallée formée par le minimum est supprimée , soit par la suppression du minimum (cas le plus général), soit par la suppression du second maximum (dans le cas très précis d'une diminution du fondamental entre les deux maxima locaux pouvant correspondre à l'attribut L). Le résultat d'une telle suppression est illustrée sur la figure 6b.

La méthode présentée ici pour la suppression des variations de micromélodie n'est pas la seule possible, ni même la meilleure. Une référence à la courbe d'énergie du signal (les consonnes à micomélodie s'accompagnent également d'une diminution de l'énergie globale du signal) peut être envisagée (le but ultime est de ne retenir que les variations du fondamental d'un sommet vocalique à l'autre). Cependant, un tel algorithme donne des résultats suffisamment bons dans le cas où la courbe du fondamental contient peu d'erreurs (comme tel est le cas pour le corpus qui a servi à tester le programme d'ensemble: la détection du fondamental s'est faite par ordinateur à partir du signal délivré par un accéléromètre placé sur la gorge des locuteurs.) Les mouvements du fondamental entre deux points sélectionnés (c'est-à-dire non supprimés) sont appelés mouvements locaux. Chaque mouvement local est candidat pour être reconnu comme la réalisation de attribut Ri ou de l'attribut L, s'il satisfait les conditions exposées précédemment.

Plan du programme:

La figure 7 illustre les 4 grandes étapes du programme général: premièrement, abstraction des variations dues à la micromélodie; deuxièmement, division du groupe en deux ou trois parties en séparant les mouvements locaux montants du début du groupe et les mouvements de même direction (montant ou descendant) à sa fin, et classification du groupe comme groupe non final ou final; troisièmement, repérage des mouvements locaux importants à l'intérieur du groupe (seuils x,y et z.); et quatrièmement, interprétation des mouvements importants en termes d'attributs (essentiellement L et Ri).

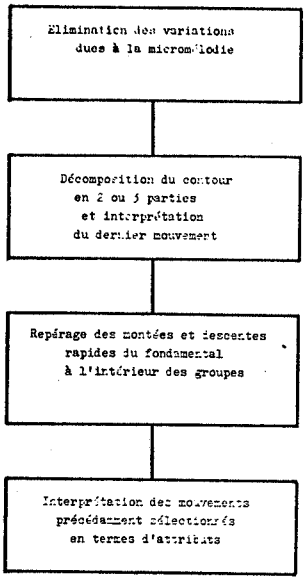


Figure 7: Plan général du programme.

Locuteurs, corpus et matériel:

Le matériel pour tester ce programme consiste en deux textes (environ 222 mots lexicaux), lus par deux locuteurs dans des conditions de laboratoire. La fréquence du fondamental a été calculée par programme (écrit par S.Maeda) à partir du signal délivré par un accéléromètre placé sur la gorge des locuteurs. Les 2 locuteurs ont été choisis pour leur façon différente de parler (le premier appartient à la catégorie P2 et le second P3).

Elimination de la micromélodie:

Les seuils employés pour suppression des variations de micromélodie pour les deux locuteurs sont les suivants (11)

	Δf	t1	t2
Locuteur 1	$<$ 25Hz	$<$ 125ms	$<$ 300ms
Locuteur 2	$<$ 70Hz	$<$ 150ms	$<$ 300ms

Division entre groupes finaux et non finaux:

Une fois la micromélodie supprimée, les mouvements locaux de Fo de même direction à la fin des groupes (avant la pause) sont regroupés et l'amplitude totale de variation est calculée. (ΔF).

Les groupes terminés par une chute finale de plus de 20Hz pour le premier locuteur, et de 50Hz pour le second sont considérés comme des groupes finaux, et les autres groupes comme des groupes non finaux. Cet algorithme permet de détecter 25 groupes finaux et 44 groupes non finaux pour le premier locuteur, et 21 groupes finaux et 53 groupes non finaux pour le second. Sur les 46 groupes finaux détectés par le programme, 43 signalent la fin d'une phrase grammaticale ou d'une proposition (une proposition peut se terminer par une intonation montante ou descendante, selon le choix du locuteur.)

Détection de la fin de mot prosodique:

Seuls les mots prosodiques se terminant par l'attrbut L non réduit (pattern P2 ou pattern P3) auront leur frontière droite détectée. Pour être reconnu comme la réalisation de l'attribut L, un mouvement local devra satisfaire les deux conditions suivantes:

	Δf	f/t
Locuteur 1	$<$ -20Hz	$>$ 15Hz/100ms
Locuteur 2	$<$ -25Hz	$>$ 17Hz/100ms

Un tel algorithme permet de détecter 94 occurences de L pour le premier locuteur (2 erreurs) et 87 pour le second (3 erreurs). Si on considère que les pauses indiquent également la fin d'un mot lexical, la position des pauses et la reconnaissance de l'attribut L permettent de détecter la frontière droite de 72°/o des mots lexicaux pour le premier locuteur et 71°/o pour le second.

Détection du début d'un mot prosodique:

Les mouvements locaux représentant un saut vers le haut du fondamental de plus de 15Hz pour le premier locuteur, et de plus de 25Hz pour le second sont reconnus comme candidats pour les attributs Rp et Ri.

Les candidats retenus précédemment, non suivis d'un mouvement local reconnu comme la réalisation de l'attribut L, sont considérés comme correspondant à l'attribut Ri (frontière gauche d'un mot prosodique). Les autres sont considérés comme Rp.

148 frontières gauches sont détectées pour le premier locuteur (5 erreurs), et 146 pour le second locuteur (9 erreurs). 64°/o des mots pour le premier locuteur et 61°/o pour le second ont donc leur frontière gauche détectée. La faiblesse de ces résultats vient essentiellement du fait que dans les mots courts (une ou deux syllables) prononcés avec le pattern P2 ou P3, la montée initiale Ri est directement connectée avec l'attribut L, et elle a été reconnue par le programme comme la réalisation de l'attribut Rp.

Correspondance entre début et fin d'un mot prosodique et d'un mot lexical:

Le début de l'attribut L se situe très exactement sur la dernière voyelle d'un mot lexical (le e muet n'est naturellement pas considéré comme une voyelle.)

L'attribut Ri correspond à la première consonne du mot lexical. Si le mot commence par une voyelle, la montée initiale peut être localisée sur la première consonne du mot, c'est-à-dire sur le début de la seconde syllabe du mot, ou sur la première voyelle du mot (choix du locuteur). Une exception non expliquée à cette règle: le mot commencant par la voyelle 'œ' tel que 'recherche' a souvent une montée sur la seconde syllabe seulement.

Conclusion:

L'intérêt d'un tel programme est limité par le fait même que les paramètres doivent être ajustés selon les locuteurs. Les seuils employés précédemment pour les deux locuteurs ont été définis à partir de l'analyse de l'un des deux textes, puis appliqués aux deux textes. Un apprentissage préalable est donc nécessaire. Et deux locuteurs ne sont pas suffisants pour juger de la validité de l'algorithme.

L'information obtenue sur les frontières peut être utilisée comme une aide à la segmentation et les mots devront être recherchés dans le dictionnaire soit à partir de leur première consonne, soit à partir de leur dernière voyelle. On peut également vérifier à l'aide de cette information si une segmentation obtenue par un autre programme est compatible ou non avec les paramètres prosodiques.

On peut donc déjà conclure que l'intégration des paramètres prosodiques dans un système général de reconnaissance comme aide à la segmentation du continuum en segments plus petits est une idée interessante, vu le faible nombre d'erreurs commises par le programme. La place exacte d'un tel programme dans un système général reste encore à définir, en fonction des autres parties de programme existantes (division en phonèmes, ou en syllabes, analyse syntaxique etc...) Le problème majeur est celui de l'adaptation des paramètres au locuteur.

Analyse syntaxique et segmentation ne sont pas les deux seules utilisations possibles des variations du fondamental. Il en existe une troisième, peut être encore plus interessante que les deux autres: l'amplitude des variations lors de la montée initiale (attribut Ri) s'est révélée lors de l'analyse être grossièrement proportionnelle à l'importance que donne le locuteur au mot dans la phrase. Lorsque la montée sur un mot est trés importante, c'est-à-dire que la montée du fondamental sur ce mot est beaucoup plus importance que la montée sur les autres mots, l'auditeur percoit sur le premier mot une sorte d'accent d'insistance (12). On pourrait donc envisager d'utiliser l'amplitude de variation du fondamental d'une syllabe à l'autre pour sélectionner la partie du discours la plus importante, c'est-à-dire celle qu'il s'agit essentiellement de reconnaitre pour comprendre le sens global de la phrase.

Références:

1. Lea W.A., An Approach to Syntactic Recognition without Phonemocs, IEEE Trans. Audio, Vol. 21, N°3, Juin 1975
2. On comprend par paramètres prosodiques: les pauses, les variations du fondamental d'un sommet vocamique (voyelle) à l'autre, les intensités relatives (corrigées) des voyelles en séquence, les durées relatives des syllabes (ainsi que la durée de chaque voyelle dans une syllabe).
3. Di Cristo A., Recherches sur la structuration prosodique de la phrase française (essai d'analyse phonosyntaxique), 6ièmes Journées d'Etude sur la Parole, Toulouse, Mai 1975, pg. 94-116.
4. Martin P., Intonation et reconnaissance automatique de la structure syntaxique, 6ièmes Journées d'Etude sur la Parole, Toulouse, Mai 1975, pg. 51-62.
5. Delattre P., Les dix intonations du Français, French Review, N°40, pg. 1-14, 1966.
6. Vaissière J., Caractérisation des variations de la fréquence du fondamental dans les phrases françaises, 6ièmes Journées d'Etude sur la Parole, Mai 1975, pg. 39-50.
7. Buisson L. et Mercier G., Utilisation de l'information prosodique en segmentation de la parole continue, 6ièmes Journées d'Etude sur la Parole, Toulouse, Mai 1975, pg. 117-123.
8. House A.S. et Fairbanks G., The Influence of Consonantal environment upon the Secondary Acoustical Characteristics of vowels, J.A.S.A. 25, pg. 105-113.
9. Lea W.A., Segmental and Suprasegmental Influences on Fundamental Frequency Contours, Southern California Occasional Papers in Linguistics N°1, Juillet 1973, pg. 17-70.

10. Boë L.J., Etude de l'interaction source laryngienne-conduit vocal dans la détermination des caractéristiques intrinsèques des consonnes du français, Bulletin de l'Institut Phonètique de Grenoble, Vol. 2, 1972, pg. 1-24.
11. Vaissière J., Automatic Segmentation of Connected Speech in French, Progress Report, Laboratoire de Recherche en Electronique, M.I.T., 1976, à paraitre.
12. Grammont M., Traité pratique de prononciation française, Paris, Librairie Delagrave, 1961 (réédition).

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

L'ANALYSE LEXICALE DANS LE SYSTEME KEAL POUR LA RECONNAISSANCE DE LA PAROLE CONTINUE

par

R. VIVES

RESUME :

L'analyseur lexical présenté travaille sur les résultats du reconnaisseur phonétique de Keal à l'aide d'un dictionnaire de mots codés phonétiquement et d'une mesure entre phonèmes définie à partir des travaux de Delattre. Une méthode d'accès hiérarchisée au dictionnaire est proposée.

SUMMARY :

In this paper we describe the design of the Keal lexical analyser operating on the phonemic recognizer outputs, using a phonemic word dictionnary and a measure between ideal phonems according to Delattre's works. A hierarchical access method is proposed for the search in the dictionnary.

TITRE : L'ANALYSE LEXICALE DANS LE SYSTEME KEAL POUR LA RECONNAISSANCE DE LA PAROLE CONTINUE (*)

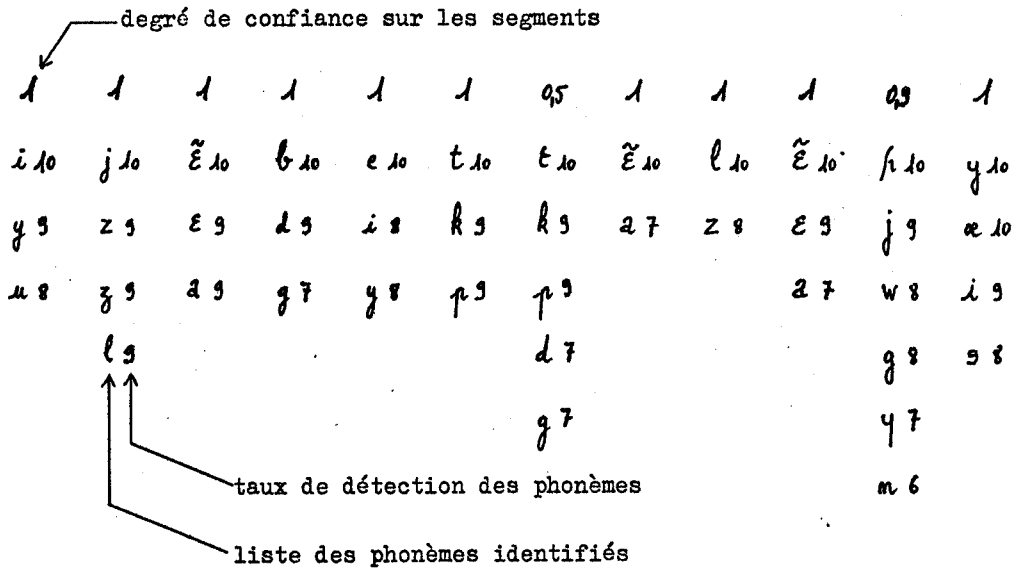
AUTEUR : R. VIVES - C.N.E.T. LANNION

I. - INTRODUCTION -

Keal est un système analytique pour la reconnaissance de la parole continue qui comporte trois niveaux d'analyse : le niveau phonémique (1), le niveau lexical, le niveau syntaxique (3).

La réponse de l'analyseur phonémique pour une phrase prononcée se présente sous la forme d'une suite d'éléments multiples pondérés.

La phrase : "il habite à Lannion" prononcée par un locuteur, a été codée par le reconnaisseur phonémique de la façon suivante :



La partie segmentation de l'analyseur phonémique attribue à chacun des douze segments trouvés, un degré de confiance, tandis que chaque réponse phonémique est munie d'un taux de détection par le reconnaisseur phonémique. Le travail du niveau lexical consiste à transformer la réponse de l'analyseur phonémique en une suite de syntagmes lexicalement possibles dont une combinaison sera validée au niveau syntaxique (voir fig. 1)

(*) Les recherches décrites dans cet article sont menées au CNET à Lannion, avec l'appui du SESORI, dans le cadre du contrat n° 74-80.

II. - PRINCIPE DE LA RECHERCHE LEXICALE DANS KEAL.

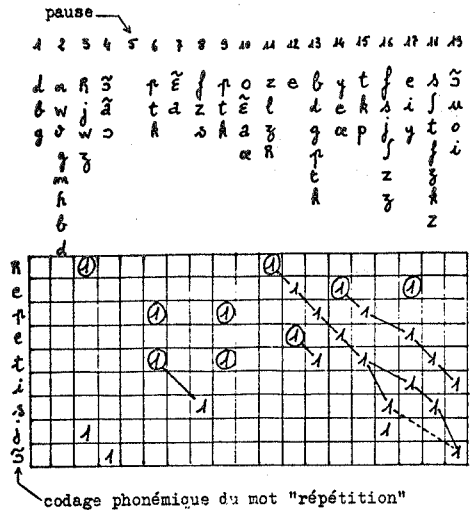
La recherche lexicale est caractérisée par trois points principaux :

- la localisation d'une forme de référence dans la forme élaborée par l'analyseur phonémique.
- la détermination d'un seuil de détection.
- la structure et le codage des éléments de référence.

II. - 1 - Dans le cas de la reconnaissance de mots isolés la localisation d'une forme de référence dans la forme élaborée par l'analyseur phonémique est triviale : la segmentation au niveau acoustique entre le bruit et la parole, donne au niveau phonémique des limites pour la forme cherchée. Dans la reconnaissance de la parole continue, on ne sait pas en général où commencent et où finissent les divers éléments de la phrase prononcée. Seules, les pauses et les silences détectés au niveau phonémique et codés par des blancs dans la suite d'éléments multiples peuvent marquer avec une certaine probabilité des débuts ou des fins de mots.

La solution que nous avons choisie est étroitement liée au calcul de notre indice de ressemblance (2). Dans ce calcul, deux listes d'éléments multiples se ressemblent selon la plus significative des sous-chaînes d'éléments qu'elles ont en commun. L'idée de comparer un élément de référence à la forme élaborée par l'analyseur phonémique revient à chercher dans la phrase prononcée toutes les occurrences de ce mot de référence. Soit, par exemple, la phrase : "non, carte répétition" dans laquelle nous voulons repérer les occurrences du mot : répétition - L'explication de ce repérage peut se faire avec les trois étapes suivantes :

- remplissage de la matrice de coincidence entre la forme élaborée par l'analyseur phonémique et le codage phonémique du mot de référence cherché.



Nous trouvons un "1" dans la case (i,j) de la matrice de coincidence s'il y a identité entre le ième phonème du mot de référence et l'une des réponses du $j^{ième}$ segment de la phrase prononcée.

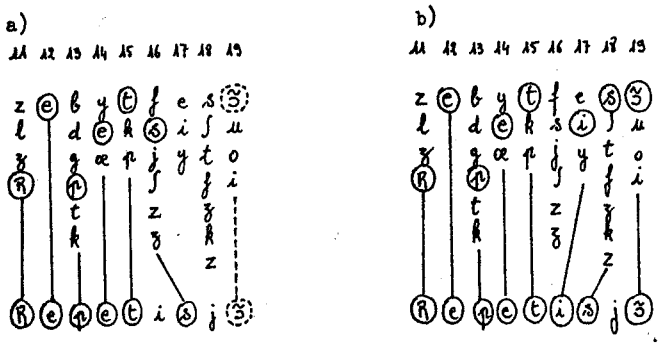
- la seconde étape est la recherche des débuts possibles du mot de référence dans la phrase. L'expérience a montré que pour un mot de référence de n phonèmes, une occurrence de ce mot dans la phrase prononcée est intéressante, si elle commence au moins par l'un des m premiers phonèmes du mot cherché, avec :

$$\text{si n est pair, } m \leq \frac{n}{2} + 1,$$
$$\text{si n est impair, } m \leq \frac{n+1}{2}$$

Dans l'exemple précédent les débuts de mot possibles ont été marqués d'un rond entourant le 1 dans la matrice de coincidence : (1).

- enfin pour chaque début de mot possible, l'algorithme de calcul de l'indice de ressemblance, ajuste, à partir des éléments les mieux reconnus, la transcription phonémique du mot de référence, sur le codage de la phrase prononcée.

Dans l'exemple précédent, nous obtenons deux cadrages intéressants à partir du 11è segment du codage de la phrase prononcée :



En a) le cadrage du phonème , marqué en pointillé, dépendra des degrés de confiance attribués aux 17è et 18è segments de la phrase. Un degré de confiance égal à 1 correspond à un segment sûr. Un degré de confiance égal à 0 est interprété comme un segment inexistant. Les performances de l'analyseur phonémique font que nous pouvons admettre la limitation suivante : deux éléments successifs appartiennent au même mot s'ils ne sont séparés que de moins de deux segments. Les pauses et les silences codés par des blancs sont munis de degrés de confiance qui permettent de pénaliser les cadrages de mots, à cheval sur ces silences : le segment silence est alors considéré comme un segment inséré dans la séquence phonémique cherchée.

Le calcul de l'indice de ressemblance a lieu pour chaque cadrage, entre le mot de référence et la partie cadrée de la phrase prononcée. Il tient compte du degré de confiance de chaque segment et du taux de détection de chaque phonème reconnu. La comparaison de cet indice au seuil de détection nous permet de retenir ou d'éliminer une occurrence du mot de référence cherché.

II. - 2 - Le choix d'un seuil de détection est guidé par deux aspects antagonistes. Dans la reconnaissance analytique de la parole continue le niveau phonémique est suivi des niveaux lexical et syntaxique: il est important de pouvoir utiliser au niveau phonémique les sources d'information des niveaux supérieurs avant de décider de la reconnaissance de tel ou tel phonème. Les réponses phonémiques multiples illustrent notre choix. Il en va de même pour la recherche lexicale qui est suivie de l'analyse syntaxique : reporter la décision de reconnaissance d'un mot au niveau syntaxique signifie qu'il faut transmettre à la syntaxe le plus d'informations possible, donc de choisir un seuil de détection de mot relativement faible, voire nul.

Une telle stratégie n'est pas applicable car les réponses de l'analyseur phonémique ne contiennent aucune indication sûre, concernant les limites des mots composant la phrase prononcée et il est alors nécessaire de considérer, par hypothèse, chaque segment de la réponse comme le début possible de l'ensemble des mots du dictionnaire. Le nombre de phrases possibles devient astronomique et l'on doit limiter le nombre de mots candidats par un seuil de détection conséquent. Dans une première série d'expériences, nous avons fixé ce seuil de détection à 0.4 de façon unique pour tous les mots. Les résultats ont montré qu'il fallait définir un seuil de détection fonction du nombre de segments de l'élément de référence cherché (nous verrons au chapître 3 qu'il faut tenir compte de la forme générale de l'élément de référence). Pour un même seuil de détection, on détecte plus facilement les mots courts (1 ou 2 phonèmes) que les mots longs. Nous avons pris un seuil égal à 0.8 pour les mots de un ou deux phonèmes et un seuil égal à 0.5 pour les autres mots.

Le seuil de détection est aussi lié à la nature de la recherche effectuée. Si l'on admet implicitement qu'une phrase prononcée contient obligatoirement l'un des mots d'une liste finie, on pourra décider que c'est l'élément qui obtient le meilleur score qui est reconnu, même si ce score n'atteint pas le seuil de détection.

II. - 3 - Le codage des éléments de référence dans le dictionnaire, a d'abord été effectué en prenant pour chaque mot le codage phonémique idéal : pour le français on utilise 16 voyelles, 3 semi-voyelles et 17 consonnes. Les applications utilisant des dictionnaires de quelques dizaines de mots ne nécessitent pas l'emploi d'une structure d'accès particulière au dictionnaire, qui peut être compulsé séquentiellement.

Pour les applications, comme l'interrogation orale d'un fichier documentaire qui peuvent utiliser plusieurs centaines de mots il est nécessaire de prévoir d'autres modes d'accès. Des tests sur l'optimisation de l'accès dans de grands dictionnaires ont été effectués dans le cadre de la reconnaissance de mots isolés (4). Nous avons mis en oeuvre un mode d'accès hiérarchisé au dictionnaire utilisant des suites d'éléments multiples comme représentants d'ensembles de mots. Un codage comme $\begin{matrix} b & a & t & o \\ \int & & & \tilde{3} \\ g & & & \end{matrix}$ est représentatif au sens de notre indice de ressemblance, des mots *bato, bat3, ʃato, ʃat3, gato,*

mais aussi gatɔ que l'on interprète alors comme un bruit de ce codage. Avec de telles classes de mots, l'analyse lexicale comporte deux étapes :

- recherche dans la phrase prononcée des représentants des classes,
- recherche des mots dont le représentant a été détecté.

Enfin dans les applications comme la reconnaissance des chiffres, qui comportent un vocabulaire restreint, nous pouvons utiliser une suite d'éléments multiples comme représentant diverses prononciations d'un même mot.

III. - CONSTRUCTION AUTOMATIQUE DE CLASSES DE MOTS POUR UN ACCES HIERARCHISE AU DICTIONNAIRE.

On appelle clé, le représentant d'une classe de mots.
Le programme de construction automatique de clés a pour données :

- un dictionnaire de mots codés phonétiquement,
- un nombre approximatif de clés désirées,
- une valeur de ressemblance minimum d'un mot avec sa clé (VRM).

En résultat, on récupère l'ensemble des classes construites. Tous les mots appartenant à une même classe, ont une ressemblance avec la clé de leur classe, au moins égale à la valeur de ressemblance minimum VRM.

Tous les calculs de ressemblance entre clés, ou entre un mot et une clé sont effectués avec notre indice de ressemblance.

III. - 1 - Description de l'algorithme.

(1) - On initialise le nombre de clés voulues par des mots tirés au hasard ; aller en (2).

(2) - s'il ne reste plus de mots à classer, aller en (5).

- sinon on calcule les ressemblances du mot suivant à classer avec les clés :

soit C_i la clé la plus ressemblante :

- si la ressemblance est inférieure à la VRM, aller en (3).
- sinon le mot est représenté par la clé C_i ; aller en (2).

(3) - si la ressemblance est supérieure à la ressemblance maximale entre 2 clés, aller en (4).

- sinon, soit C_k et C_1 les deux clés les plus ressemblantes ;

  - Si tous les mots représentés par C_k ont une ressemblance supérieure à la VRM avec C_1, alors les mots de C_k sont regroupés dans C_1 ; on initialise C_k avec le mot à classer aller en (2).

  - sinon - si tous les mots représentés par C_1 ont une ressemblance supérieure à la VRM avec C_k, alors les mots de C_1 sont regroupés dans C_k ; on initialise C_1 avec le mot à classer ; aller en (2).

- sinon la ressemblance entre C_k et C_l est prise égale à la ressemblance entre le mot à classer et C_i ; aller en (3).

(4) - s'il existe une clé ayant le même nombre de segments que le mot à classer, on agglomère * le mot à cette classe ; aller en (2).

- sinon le mot est mis dans une liste d'éléments non traitables ; aller en (2).

(5) - si la liste des éléments non traitables est vide aller en (6).

- sinon - s'il y a eu changement dans la structure des clés on remplit la liste des mots à classer avec la liste des éléments non-traitables ; aller en (2).

- sinon on initialise une clé supplémentaire avec le premier élément non-traitable ; aller en (2).

(6) fin.

III. - 2 - Nous avons traité jusqu'à présent des vocabulaires ne dépassant pas une centaine de mots. Sur le vocabulaire des 25 mots permettant la description des chiffres de 0 à 99, avec un nombre de classes approximatif égal à 4 et une VRM égale à 0.6; le programme a construit six classes :

clé n° 1 : $\tilde{œ}$ représentant 1 mot : $\tilde{œ}$

clé n° 2 : $s\ \tilde{ɛ}\ k$ / $y\ i\ s$ / $d\ u\ t$ / z représentant 7 mots : $s\ \tilde{ɛ}\ k$ / $s\ i\ s$ / $y\ i\ t$ / $d\ i\ s$ / $d\ u\ z$ / $k\ \tilde{ɛ}\ z$ / $ʒ\ \tilde{ɛ}\ t$

clé n° 3 : $e\ \tilde{ɔ}\ z$ / $s\ ɛ\ t$ / $m\ œ\ ʃ$ représentant 5 mots : $s\ ɛ\ t$ / $m\ œ\ ʃ$ / $\tilde{ɔ}\ z$ / $s\ ɛ\ z$ / $e\ \tilde{ɔ}\ z$

* exemple : l'agglomération du mot $y\ i\ t$ à la classe des mots $s\ \tilde{ɛ}\ k$ et $s\ i\ s$ représentés par la clé : $s\ \tilde{ɛ}\ k$ / $i\ s$

donne la classe des mots $s\ \tilde{ɛ}\ k$, $s\ i\ s$, $y\ i\ t$
représentés par la clé : $s\ \tilde{ɛ}\ k$ / $y\ i\ s$ / t

clé n° 4 : k a t ʁ
z e ʁ o
t ʁ w a
ɛ z

représentant 6 mots : z e ʁ o
t ʁ w a
k a t ʁ
t ʁ ɛ z
k a t o ʁ z
t ʁ ã t

clé n° 5 : d ə
v ɛ̃
e œ̃

représentant 3 mots : d ə
v ɛ̃
e œ̃

clé n° 6 : k a ʁ ã t
s ɛ̃ k

représentant 3 mots : k a ʁ ã t
s ɛ̃ k ã t
s w a s ã t

L'application à la reconnaissance a montré que la méthode n'était pas performante pour les mots courts. Les mots de un phonème comme a , ã , ɔ̃ , o , e , u , œ se retrouvent dans une même classe et sont représentés par un segment voyelle multiple qui est détecté presque partout. Les mots de deux phonèmes sont regroupés dans des classes dont la clé représente, en fait, une syllabe floue qui est recadrée sur la plupart des syllabes de la phrase prononcée. Pour les mots de plus de deux phonèmes, les clés ne sont pas systématiquement détectées dans chaque phrase : il y a donc gain de temps, mais l'amélioration reste insensible pour les applications utilisant des petits dictionnaires. L'emploi des clés à la reconnaissance repose le problème de la détermination des seuils de détection. Avec quel score z doit-on détecter une clé, pour qu'un mot normalement repéré avec un score x, soit aussi sélectionné, quand la valeur minimum de ressemblance qu'il a avec sa clé est y ?. Au pire, si x, y, z sont des valeurs comprises entre 0 et 1, il faut avoir :

$$z = x + y - 1$$

En prenant x = 0.4 et y = 0.75, il faudrait détecter les clés avec un seuil z = 0.15 pour être sûr de ne pas manquer des solutions. Une telle valeur n'est pas opérationnelle : on trouve ce qu'on cherche presque partout. Dans les expériences que nous avons menées, nous avons choisi $x = \frac{1}{2}$, y = 3/4 et z = 2/3 sans qu'il y ait de baisse significative des résultats.

IV. - INTRODUCTION D'UNE MESURE DE RESSEMBLANCE ENTRE PHONEMES.

Dans le calcul de notre indice de ressemblance, le code s a t o qui pourrait être donné par l'analyseur phonémique pour le mot ʃ a t o est trouvé aussi ressemblant à ʃ a t o qu'à g a t o alors que l'analyse phonémique a déterminé le critère fricatif sourd du premier segment. Nous avons donc voulu faire intervenir une mesure de ressemblance entre phonèmes au niveau de la détection d'un mot dans une phrase.

IV. - 1 - Définition des mesures employées.

Le but de l'expérience étant essentiellement d'apprécier l'influence d'une telle heuristique sur les résultats de la reconnaissance, nous sommes partis des travaux de Delattre (5) concernant les définitions des phonèmes moyens du français.

IV. - 1 - 1 - Mesure entre voyelles.

Dans l'espace des formants, nous avons choisi de définir une voyelle par ses deux premiers formants. Nous n'avons pas tenu compte du formant de nasalité pour $\tilde{\text{ɔ}}$, $\tilde{\text{a}}$, $\tilde{\text{œ}}$ et $\tilde{\text{ɛ}}$. Si F_1, i et F_2, i sont les valeurs moyennes des deux premiers formants de la ième voyelle, une distance $d_{i,j}$ entre la $i^{ème}$ voyelle et la $j^{ième}$ voyelle peut être définie par :

$$d_{i,j} = \left|F_{1,i} - F_{1,j}\right| + \left|F_{2,i} - F_{2,j}\right| ,$$

soit dmax la distance maximum trouvée entre deux voyelles, un indice de ressemblance ri,j entre la $i^{ème}$ et la $j^{ième}$ voyelle peut être défini par :

$$r_{i,j} = 1 - \frac{d_{i,j}}{dmax} ,$$

$r_{i,j}$ est voisin de 1 si les 2 voyelles sont très ressemblantes,
$r_{i,j}$ est voisin de 0 si les 2 voyelles ont des formants très différents.

Arrondies et à un facteur 10 près, nous avons obtenu les ressemblances suivantes :

	i	y	e	ø	ɛ	œ	a	ɑ	ɔ	o	u	ɛ̃	œ̃	ɑ̃	ɔ̃
i	10														
y	7	10													
e	8	7	10												
ø	5	8	7	10											
ɛ	5	9	7	8	10										
œ	3	7	5	8	9	10									
a	2	5	4	7	7	9	10								
ɑ	1	5	3	6	6	8	9	10							
ɔ	1	5	3	6	6	8	7	8	10						
o	1	4	5	6	4	6	5	6	8	10					
u	2	5	2	5	4	5	5	5	8	9	10				
ɛ̃	0	5	2	5	5	7	6	7	9	9	8	10			
œ̃	1	4	3	6	6	8	7	8	10	8	7	9	10		
ɑ̃	3	6	5	8	8	10	9	9	8	6	5	7	8	10	
ɔ̃	5	8	7	8	10	8	7	7	6	4	4	5	6	8	10

IV. - 1 - 2 - Mesure entre consonnes.

Nous avons utilisé la classification hiérarchisée des consonnes donnée par Delattre pour établir un indice de ressemblance entre les consonnes du français. Chaque consonne est définie par 4 traits articulatoires : le degré d'aperture, le mode d'articulation, le point d'articulation et l'indice de sonorité. Chaque trait articulatoire pouvant prendre différentes valeurs, nous avons fait une estimation simple de ces valeurs entre elles, afin de pouvoir définir une ressemblance comme la somme des valeurs des traits, que deux consonnes considérées peuvent avoir en commun.

Nous avons choisi un taux de similitude, au sens de la distance de Hamming, pour le degré d'aperture, qui peut être occlusif ou constrictif et pour le mode articulatoire qui peut être, oral ou nasal pour une occlusive, ou, fricatif ou sonnant pour une constrictive.

Entre certaines valeurs du point d'articulation nous avons choisi un taux de similitude de $\frac{1}{2}$. Il en est ainsi entre la valeur labio-dentale et les valeurs labiale, dento-alvéolaire, et latéro-dentale ; entre la valeur palato-arrondi et les valeurs palato-velaire, palatale, alvéo-palatale, palato-écarté et velo-écarté ; entre la valeur alvéo-palatale et les valeurs dento-alvéolaires, palato-velaire, palatale et palato-écarté ; entre la valeur palato-velaire et les valeurs palatale, palato-écarté et velo-arrondi ; enfin, entre les valeurs palatale et palato-écarté.

Un taux de similitude de $\frac{1}{2}$ a également été choisi entre les valeurs de l'indice de sonorité suivantes : fort et indéfini ; faible et indéfini.

Entre toutes les autres valeurs de ces deux derniers traits articulatoires, nous avons repris un taux de similitude au sens de la distance de Hamming.

Arrondies et normalisées entre 0 et 10 nous obtenons les ressemblances suivantes :

	p	b	t	d	k	g	m	n	ɲ	ŋ	ɔ	s	z	ʃ	ʒ	ʁ	l	j	ɥ	w
p	10																			
b	7	10																		
t	7	5	10																	
d	5	7	7	10																
k	7	5	7	5	10															
g	5	7	5	7	7	10														
m	5	5	2	2	2	2	10													
n	2	2	5	5	2	2	7	10												
ɲ	2	2	2	2	4	4	7	7	10											
ŋ	4	1	4	1	2	0	1	1	0	10										
ɔ	1	4	1	4	0	2	1	1	0	7	10									
s	2	0	5	2	2	0	0	2	0	3	6	10								
z	0	2	2	5	0	2	0	2	0	6	3	7	10							
ʃ	2	0	4	1	4	1	0	1	1	7	5	3	6	10						
ʒ	0	2	1	4	1	4	0	1	1	5	7	6	3	7	10					
ʁ	0	0	0	0	0	0	2	2	2	2	2	2	2	2	2	10				
l	0	0	1	1	0	0	2	4	2	4	4	4	4	2	2	7	10			
j	0	0	0	0	1	1	2	2	4	2	2	2	2	4	4	7	7	10		
ɥ	0	0	0	0	1	1	2	2	4	2	2	2	2	4	4	7	7	3	10	
w	0	0	0	0	1	1	2	2	2	2	2	2	2	2	2	7	7	7	3	10

IV. - 1 - 3 - Dans une première approche il nous a semblé suffisant de considérer la ressemblance entre consonnes et voyelles comme nulle.

IV. - 2 - <u>Modification du calcul de l'indice de ressemblance</u>.

C'est au niveau du remplissage de la matrice de coïncidence qu'il nous a paru le plus judicieux de faire intervenir la ressemblance entre phonèmes.

Le remplissage de la matrice de coïncidence est binaire : nous trouvons un "1" dans la case (i,j) s'il y a identité entre un phonème du $i^{ième}$ segment du mot de référence et l'une des réponses du $j^{ième}$ segment de la phrase prononcée.

Un mot n'est détecté que si l'un des chemins, déterminés par les "1" de la matrice, obtient un score suffisant. Il faut bien voir alors qu'il n'y a score suffisant que s'il y a parcours d'un chemin. Les valeurs de ressemblance entre phonèmes permettent de remplir la matrice par des réels entre 0 et 1, représentant la ressemblance maximum existant entre deux phonèmes de deux segments comparés. Un chemin dans la matrice peut être déterminé par les cases où l'on trouve des valeurs supérieures à un seuil donné. L'indice calculé tient compte des valeurs trouvées dans ces cases.

Nous avons fait une série d'expériences en ne prenant en compte que les chemins déterminés par des valeurs supérieures à 0,6. Les résultats de la recherche lexicale sont améliorés mais le prix, en place, au niveau de l'analyse des résultats par la syntaxe, risque d'être lourd. Sans tenir compte des ressemblances entre phonèmes, dans une vingtaine de phrases d'une application utilisant environ 60 mots, 90 % des mots prononcés sont détectés en bonne place et la syntaxe doit gérer en moyenne un nombre de mots détectés égal à trois fois le nombre de segments phonémiques de la phrase prononcée. Avec les ressemblances entre phonèmes la détection en bonne place des mots prononcés augmente de 2 % mais le nombre de détections à gérer est multiplié par 5 !

Même si l'on augmente le seuil de 0,6, il semble que cette méthode ne soit utile que dans des cas très particuliers comme la vérification d'hypothèses au niveau lexical ou la construction automatique de classes de mots.

V. - CONCLUSION.

Les résultats obtenus sont encourageants mais il semble que pour des applications faisant intervenir un grand nombre de locuteurs et des dictionnaires plus volumineux, comme l'interrogation de banques de données, il sera nécessaire d'introduire un niveau phonologique entre l'analyse phonémique et la recherche lexicale. Les phénomènes phonologiques connus comme les réductions de voyelles dans des mots grammaticaux, les liaisons, les enchaînements, sont actuellement comptés comme des erreurs de l'analyse phonémique. La syntaxe peut rattraper dans certains cas des "oublis" de la recherche lexicale, le dialogue permet de corriger des fausses interprétations mais nous ne pourrons pas dans l'avenir, ne pas tenir compte de règles, permettant de lever des ambiguités spécifiquement phonologiques.

BIBLIOGRAPHIE -

(1) Buisson L., Mercier G., Gresser J.Y., Querré M., Vivès R.

Phonetic Decoding for automatic recognition of words. Speech Communication Seminar Stockholm. Aug 1-3, 1974.

(2) Vivès R., Gresser J.Y. A similiraty index between strings of symbols - Application to Automatic Word and Language Recognition - Proceeding of the 1st International Joint Conference on Pattern Recognition - Oct. 1973 Washington DC.

(3) Quinton P. - Un analyseur syntaxique adapté à la reconnaissance de la parole - 7è J.E.P. Nancy - Mai 1976.

(4) Gagnoulet C., Buisson L., Optimisation de la recherche lexicale - 6è J.E.P. - Toulouse - Mai 1975.

(5) Delattre P., Comparing the Phonetic Features of English, German, Spanish and French - Julius Groos Verlag - 1965.

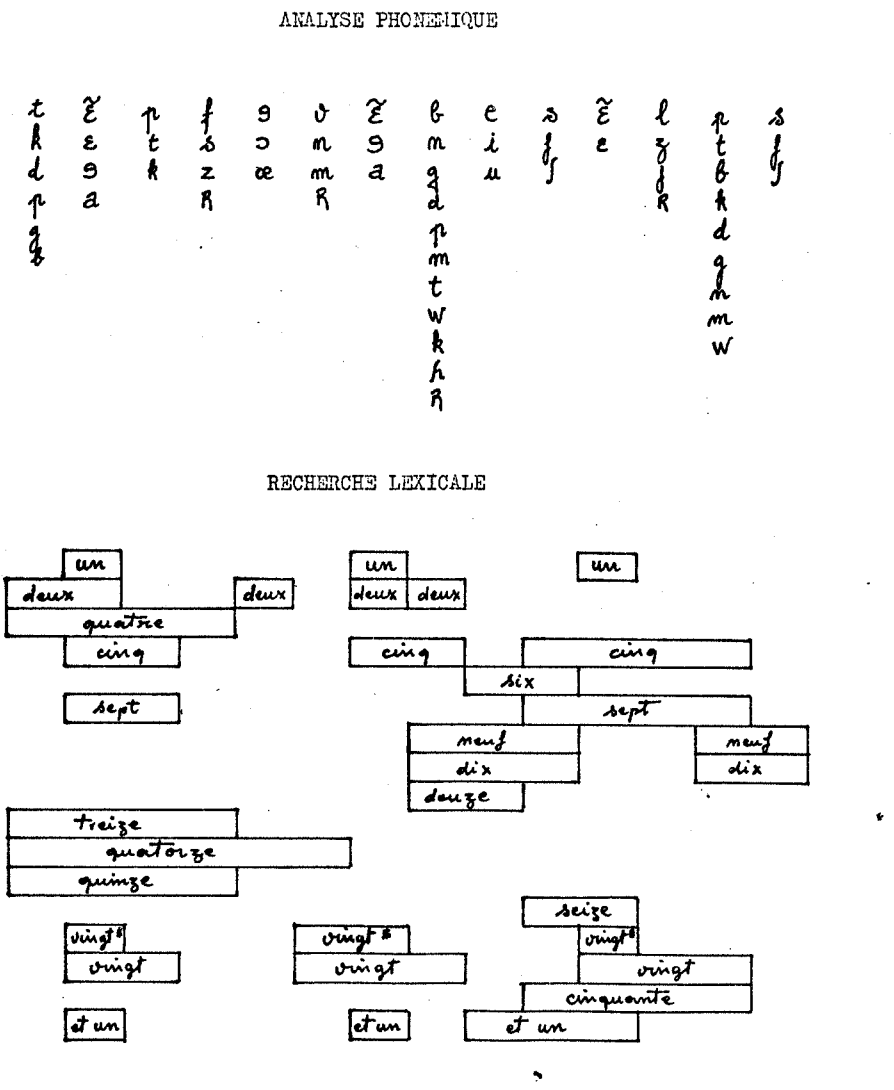


fig. 1

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

SYNTHESE PARAMETRIQUE DE LA PHRASE INTERROGATIVE EN FRANCAIS

(Question totale)

Louis-Jean BOË & Michel CONTINI

Institut de Phonetique de Grenoble

RESUME : Dans un premier temps sont présentés les résultats relatifs à l'analyse de la phrase interrogative (question totale) dont la nature est liée essentiellement à l'intonation. Un corpus de 45 phrases a été enregistré par 10 locuteurs, les tracés de la fréquence laryngienne permettent de dégager un contour type, positionné à l'aide de 3 niveaux (probabilité d'occurrence de 68%), référencés par rapport à la fréquence laryngienne moyenne de la phrase énonciative.

Ensuite une série de tests perceptifs a été menée à l'aide de stimuli synthétiques générés par un ensemble vocodeur/ordinateur (département ETA du CNET Lannion) afin de déterminer avec quelles latitudes il est possible de donner à une phrase énonciative un caractère interrogatif et vice-versa. Les résultats confirment et précisent l'importance de l'évolution de la fréquence laryngienne dans sa partie finale, mais aussi de celle de l'intensité de la dernière syllabe.

Cette étude permet de définir les différents paramètres permettant de générer une phrase interrogative, une série de stimuli sont présentés pour illustration.

SUMMARY :

In previous papers three pitch levels have been specified for statement in relation to the mean laryngeal frequency as frequency bands in which the attack, the maximum and the end of the intonational pattern have an occurrence probability. In this paper the same method has been used for yes-no question. In a second part the results of the analysis have been tested with synthetic speech produced by a vocoder to determine an archetypal pattern for synthesis by rule.

SYNTHESE PARAMETRIQUE DE LA PHRASE INTERROGATIVE EN FRANCAIS
(Question totale)

Louis-Jean BOË & Michel CONTINI

INTRODUCTION

Cette étude s'inscrit dans un programme de recherche concernant l'intonation du français; nous ne reviendrons pas en détail sur les hypothèses de départ adoptées ni sur la méthodologie utilisée, plusieurs rapports ayant déjà été publiés*. Nous rappelerons seulement la première étape de notre démarche qui peut se résumer ainsi :

1. enregistrement d'un corpus lu de phrases d'une catégorie donnée et étude des tracés des variations de la fréquence laryngienne F_ℓ;
2. établissement de niveaux intonatifs spécifiés statistiquement sur une échelle logarithmique et référencés par rapport à la fréquence laryngienne moyenne $\bar{F}_\ell$ de la phrase énonciative; détermination d'un modèle de contour intonatif positionné par rapport à ces niveaux et dégagé des caractéristiques intrinsèques;
3. élaboration de stimuli synthétiques comportant des modifications de F_ℓ par rapport à cet archétype mais gardant, dans la mesure possible, l'information spectrale intacte;
4. mise en évidence, à l'aide de tests, des influences perceptives de ces variations, c'est-à-dire de l'importance relative des éléments du contour.

Nous présentons ici un travail concernant la question totale, type de phrase pour lequel c'est l'intonation seule qui porte la marque interrogative. C'est d'ailleurs le modèle de loin le plus courant pour le français parlé (HOLDER, 1969; TERRY, 1970). Traditionnellement (GREVISSE, 1964) et par opposition à l'évolution décroissante de la partie finale du contour de la phrase énonciative, la question totale est associée à l'intonation montante; FOULET (1921) avait même avancé que ... "toute interrogation s'accompagne d'un ton montant" ... mais que celui-ci n'est pas obligatoirement à la même place. Un certain nombre de travaux perceptifs ou instrumentaux ont été effectués pour cette catégorie d'interrogation et des méthodes d'enseignement du français aux étrangers proposent des contours mélodiques modèles (COUSTENOBLE & AMSTRONG, 1934; DELATTRE, 1960, 66, 67, 69; METTAS, 1963, 64; LEON, 1966; LEON & MARTIN, 1969; FONAGY & BERARD, 1973; GRUNDSTROM, 1973). Mais pour l'essentiel, ces études relèvent du domaine qualitatif et ne sont pas directement utilisables pour l'établissement systématique de règles de synthèse et la caractérisation de leurs limites de validité.

PROCEDURE ET RESULTATS

1. Analyse

Le corpus se compose de 45 phrases (placées dans un ordre aléatoire) dont certaines avaient été réalisées de façon énonciative dans un travail précédent (CONTINI & BOË, 1973), leur nombre de syllabes varie de 4 à 13 (la moyenne est légèrement supérieure à 7); pour l'essentiel nous avons utilisé un sujet se présentant sous la forme d'un syntagme nominal. Afin de n'avoir à apporter que le minimum de corrections relatives aux caractéristiques intrinsèques, les phrases commencent et

* BOË & RAKOTOFIRINGA 1971, 72, 75; BOË, 1973 a,b; CONTINI & BOË, 1973, 75; LARREUR & BOË, 1973; BOË & LARREUR, 1974; BOË, CONTINI & RAKOTOFIRINGA, 1975.

se terminent le plus souvent par la voyelle [a] . Les locuteurs (5 hommes et 5 femmes) sont ceux qui avaient enregistré le corpus des phrases énonciatives. La lecture des phrases détachées a été précédée par celle d'un texte suivi de deux minutes environ ne comportant que des phrases énonciatives. Cette procédure de routine a été utilisée pour permettre au locuteur de bien positionner sa voix avant d'entamer le corpus proprement dit et de servir pour un relevé de F_ℓ utilisé comme référence pour positionner les niveaux*.

L'enregistrement a été effectué en chambre sourde (microphone A.K.G.D.24, NAGRA 4.2, bande Scotch 175). Les phrases jugées peu naturelles ou mal réalisées ont été réenregistrées après une première lecture. Les tracés de la fréquence laryngienne ont été obtenus avec l'appareillage de l'Institut de Phonétique qui permet de mesurer, en temps réel, le nombre de périodes d'une séquence donnée, la durée de voisement et donc la fréquence laryngienne moyenne $\overline{F}_\ell$. Des appareils annexes ont été utilisés : générateur Hewlett-Packard 3300A et tiroir 3302A, compteur Hewlett-Packard 5246L et S.E.F. 4905, inscripteur à jet d'encre Oscillomink standard Siemens. Les calculs statistiques ont été effectués avec les moyens informatiques du Département E.T.A du C.N.E.T. Lannion**. et du Centre Universitaire de Calcul de Grenoble auquel l'Institut de Phonétique est relié par un terminal.

Sur les tracés de F_ℓ nous avons relevé les valeurs de l'attaque, du maximum intermédiaire, de la valeur finale (en Hz), de la durée de la montée (en cs) et de la montée (en tons) : voir figure 1. Pour les locuteurs masculins nous avons noté, pour chaque phrase la durée voisée et le nombre de périodes ce qui permet de préciser la $\overline{F}_\ell$ de chaque énoncé.

Comme nous l'avions fait pour les phrases énonciatives, nous avons vérifié que pour les questions la distribution des variables relevées pouvait être considérée comme normale, tout au moins en première approximation. Pour ce faire, nous avons tracé, en repère gausso-arithmétique, les fonctions de répartition; à titre d'exemple nous les présentons, à la figure 2 pour le locuteur 8.

Au tableau I, nous avons noté, par locuteur, les valeurs de l'attaque, du maximum intermédiaire et de la fin du contour et celles de leur écart type qui permettent de préciser une bande de fréquence à $\pm\sigma$ que par définition nous appelons niveau. Nous ne faisons pas de moyennes sur ces résultats, mais sur ceux du tableau II dans lequel sont exprimées en tons les bandes B précédentes, les valeurs de la montée et son écart type, sa durée (en cs.) et son écart type.

Comme ce qui avait été observé pour la phrase énonciative, on peut observer de légères différences entre les locuteurs masculins et féminins :

---

* . en fait la fréquence laryngienne moyenne des phrases énonciatives lues détachées est légèrement plus élevée que celle d'un texte ne comportant que ce même type d'énoncé : nous avons relevé des différences moyennes de l'odre de 1/3 de ton (CONTINI & BOË, 1973).

**. nous tenons à remercier Monsieur LORAND, le responsable du département et P. GRAILLOT pour son aide.

. la m[illegible]tée est un peu plus importante pour les locuteurs masculins, [illegible],5 [illegible]ns au lieu de 3,3 tons.

. pour les locuteurs féminins, l'attaque de la question coïncide avec la $\bar{F}_\ell$ des phrases énonciatives (ce qui est aussi le cas pour les phrases énonciatives) alors qu'il y a une différence de presque 1 ton pour les locuteurs masculins.

A part ceci, les résultats sont assez homogènes. L'attaque et la fin sont toujours très stables (68% des cas dans un intervalle de 1 à 1,3 ton). Par contre le maximum intermédiaire est deux fois plus fluctuant qu'il ne l'était pour la phrase énonciative (2 tons au lieu de 1) ce qui laisse suspecter sa validité en tant que niveau intonatif. L'importance des écarts type de la montée et de sa durée donne à penser qu'au niveau de la production, ce n'est pas la pente de la partie finale qui se présente comme un invariant, mais bien la valeur finale du contour de F_ℓ.

Aux figures 3, nous avons présenté par sexe, les contours de la question totale tels qu'ils se présentent à l'analyse. Sont précisés, en tons, les différents écarts par rapport à $\bar{F}_\ell$ de la phrase énonciative et les bandes de fréquence $\pm \sigma$. Pour les locuteurs masculins nous avons fait figurer de plus la valeur de $\bar{F}_\ell$ de la phrase interrogative.

Nous avons relevé (tab. III) le coefficient de corrélation entre les différentes variables et ceci pour les 5 locuteurs masculins : on peut constater qu'il n'apparaît aucune relation systématique si ce n'est pour 4 locuteurs sur 5 entre la valeur finale du contour et celle de la montée (la taille de chaque échantillon étant 45, pour atteindre un seuil de signification de 0,01 ou 0,05 il faut que le coefficient de corrélation soit supérieur à 0,38 ou 0,28). Ceci met en évidence l'indépendance de l'évolution des variables à l'intérieur des limites précisées précédemment

2. Synthèse

La réalisation et la perception des faits prosodiques mettent en jeu pour l'essentiel, des phénomènes de fréquence laryngienne, d'intensité, de durée segmentale et plus généralement d'organisation temporelle du discours (que l'on peut appeler le rythme). Pour le français, il a été mis en évidence, que F_ℓ constitue, à elle seule, la variable la plus importante et c'est pour cette raison que nous lui avons consacré l'essentiel de l'analyse. Aussi la deuxième partie de ce travail a pour but :

- de déterminer les limites perceptuelles des moyennes issues de l'analyse et de les comparer avec les écarts observés à la production,
- de tester les règles permettant, à l'aide de la synthèse, de passer d'une phrase énonciative à une question totale et vie-versa.

C'est un phénomène connu depuis longtemps des phonéticiens, et dans une étude précédente nous l'avons spécifié quantitativement pour la phrase énonciative, la partie finale du contour de F_ℓ, porte, à elle seule, l'essentiel de la marque intonative.
Pour la phrase interrogative, nous nous sommes donc limités à des stimuli ne comportant que des modifications de ce segment.

Nous sommes partis d'une même phrase, ne comportant que des sons voisés, *Annie a gagné*, réalisée dans les deux corpus sous forme énonciative et interrogative. Nous avons choisi un locuteur masculin, dont les résultats correspondent à peu près aux valeurs moyennes. Les deux énoncés ont été analysés à l'aide d'un ensemble vocodeur / ordinateur*, puis nous avons opéré un certain nombre de modifications. Dans un premier temps nous avons généré, à partir de chaque phrase, 13 contours de F_0 (y compris celui de la phrase originale) correspondant à des évolutions différentes de la dernière syllabe. Cependant un test préliminaire nous a conduit à partir d'énoncés plus neutres. En effet avec la phrase énonciative comme base de départ, les contours montants paraissent peu naturels, c'était la même impression avec les stimuli issus de la phrase interrogative et affectés d'un contour descendant. En définitive (figure 4) :

- nous avons généré, à partir de la phrase interrogative, 13 contours de F_0 , différents avec une dernière syllabe dont l'intensité a été diminué d'une dizaine de dB,
- avec la phrase énonciative nous avons les mêmes contours mais l'accent de l'avant-dernière syllabe a du être gommé : le maximum de F_0 , situé avant la chute finale, a été supprimé, l'intensité diminué d'environ 10 dB; par contre l'intensité de la dernière syllabe a été augmentée de la même quantité.

En fait ces corrections correspondent à des phénomènes bien observables à l'analyse : avec une question totale, il ne peut y avoir un accent très net sur l'avant-dernière syllabe, puisque c'est sur la dernière que va porter toute l'information intonative. D'autre part, et nous serons amenés à le préciser quantitativement, cette montée finale de F_0 entraîne (par interaction physiologique) une augmentation de l'intensité alors que c'est le contraire avec la phrase énonciative.
Les 26 stimuli synthétiques ont été placés dans un ordre aléatoire et présentés à 12 auditeurs qui avaient le choix entre 3 catégories : énonciation, interrogation, ni l'une ni l'autre des deux précédentes. Les réponses figurent au tableau IV où la 3e catégorie a été regroupée sous l'étiquette indécision. Si l'on détermine des classes à partir de 75% des réponses concordantes, on constate que ce sont les mêmes quelle que soit la phrase initiale. Les contours 0, 1, 2, 3, 4 provoquent une perception d'énonciation et les contours 10, 11, 12 une impression de question. Ce sont les 7 et 8 qui sont les plus étrangers à ces deux classes.
Entre les contours 12 et 10 l'écart est de 2,2 tons et entre le 0 et 4 de 4,4 tons ce qui est, dans les deux cas, nettement plus important que ce que l'on peut relever à l'analyse (environ 1,5 ton, puisque l'écart-type correspond à 68% des cas).

---

*. Celui du département E.T.A. du C.N.E.T. Lannion; nous tenons à remercier J.J. LUCAS pour sa collaboration.

CONCLUSION

Les résultats de ce travail sur la question totale permettent d'avancer :

- l'attaque et la fin de l'évolution de F_0 constituent deux niveaux qui s'étendent sur 1 ton environ (si l'on choisit comme définition 68% des occurrences, c'est-à-dire à $\pm \sigma$). Le niveau de l'attaque de la question correspond, à peu de chose près, à celui de la phrase énonciative. Ceci confirme les travaux de DELATTRE (1966).
- l'absence de corrélation entre durée de la montée et montée elle-même et l'importance des écarts-type de ces deux variables donnent à penser que ce n'est pas la pente de la partie finale, mais bien la valeur finale elle-même, qui est significative.
- le maximum intermédiaire, qui se produit essentiellement sur le syntagme nominal sujet ne constitue pas un niveau.

Les tests perceptuels mettent en évidence, et cela avait déjà été constaté pour la phrase énonciative, une plus grande latitude pour les contours.

Les modèles de phrases énonciative et interrogative que nous avons relevés nous permettent de passer à la synthèse d'une type à l'autre mais il faut aussi intervenir au niveau de l'intensité des deux dernières syllabes, faute de quoi les stimuli générés ne sont pas naturels.

Références

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

BOË, L.J. (1973b), Etude de l'interaction source laryngienne-conduit vocal dans la détermination des caractéristiques intrinsèques des consonnes du français (fréquence laryngienne). Mesures de la durée. Bulletin de l'Institut de Phonétique de Grenoble 2, pp. 1-24.

BOË, L.J., CONTINI, M. & RAKOTOFIRINGA, H. (1975), Etude statistique de la fréquence laryngienne. Application à l'analyse et à la synthèse des faits prosodiques du français. Phonetica 32, pp. 1-23.

BOË, L.J. & LARREUR, D. (1974), Synthesis by Rule of Enonciative Sentence in French. Preliminary Study. Speech Communication Seminar. Stockholm.

BOË, L.J. & RAKOTOFIRINGA, H. (1971), Exigences, réalisation et limite d'un appareillage destiné à l'étude de l'intensité et de la hauteur d'un signal acoustique. Revue d'Acoustique 4, pp. 103-113.

BOË, L.J. & RAKOTOFIRINGA, H. (1972), Une méthodologie systématique de la mesure de la fréquence laryngienne, de l'intensité et de la durée de la parole. Bulletin de l'Institut de Phonétique de Greroble 1, pp. 1-9.

BOË, L.J. & RAKOTOFIRINGA, H. (1975), A Statistical Analysis of Laryngeal Frequency : its Relationship to Intensity Level and Duration. Language and Speech 18, pp. 1-13.

CONTINI, M. & BOË, L.J. (1973), Contribution à l'étude quantitative de l'évolution de la fréquence laryngienne dans la phrase énonciative en français. Bulletin de l'Institut de Phonétique de Grenoble 2, pp. 77-92.

CONTINI, M. & BOË, L.J. (1975), Etude quantitative de l'intonation en français. Premiers résultats. 8th Int. Congr. Phonet. Sci., 17-23 Aug. Leeds.

COUSTENOBLE, H.N. & ARMSTRONG, L.E. (1934), Studies in French Intonation, Heffer, Cambridge.

DELATTRE, P. (1960), Un cours d'exercices structuraux et de linguistique appliquée. The French Review 33, pp. 591-603.

DELATTRE, P. (1966), Les dix intonations de base du français. The French Review 40, pp. 1-14.

DELATTRE, P. (1967), La nuance de sens par l'intonation. The French Review 41, pp. 326-339.

DELATTRE, P. (1969), L'intonation par les oppositions. Le Français dans le Monde 64, pp. 6-13.

FERRIEU, G., PERSON, J.M. & CARTIER, M. (1969), A.S.P.I.C. : Analyseur et synthétiseur de parole à informations codées (système C.N.E.T.). L'Onde Electrique 49, pp. 376-377.

FONAGY, I. & BERARD, E. (1973), Questions totales simples et implicatives en français parisien. Studia Phonetica 8, pp.19-51.

FOULET, L. (1921), Comment ont évolué les formes de l'interrogation. Romania XLVII, pp. 243-348.

GREVISSE, M. (1964), Le bon usage. Hatier, Paris.

GRUNDSTROM, A. (1973), L'intonation des questions en français standard. Studia Phonetica 8, pp. 19-51. Didier, Montréal, Paris.

HOLDER, M. (1969), Etude sur l'intonation comparée de la phrase énonciative en français canadien et en français standard. in Recherches sur la structure phonique du français canadien. Studia Phonetica 1, pp. 175-191.

LEON, P. (1966), Prononciation du français standard. Didier, Paris.

LEON, P. & MARTIN, P. (1969), Prolégomènes à l'étude des structures intonatives. Didier, Montréal, Paris, Bruxelles.

METTAS, O. (1963), Etude sur les facteurs ectosémantiques de l'intonation en français. Travaux de Linguistique et de Littérature 1, pp. 143-154.

METTAS, O. (1964), Etude sur l'intonation en français. Travaux de Linguistique et de Littérature 2, 1, pp. 99-105.

TERRY, R.M. (1970), Contemporary French Interrogative Structure. Editions Cosmos, Montréal.

locuteur	att.	σ(att.)	att.±σ	max.	σ(max.)	max.±σ	fin	σ(fin)	fin ±σ
1	234	17,8	216 - 252	305	27,4	278 - 332	444	26,6	417 - 471
2	197	12,3	185 - 209	265	15,9	249 - 281	308	23,2	285 - 331
3	195	8,6	186 - 204	230	14,6	215 - 245	301	18,4	283 - 319
4	216	12,6	203 - 229	252	25,2	227 - 277	330	32,6	297 - 363
5	234	10,5	223 - 244	268	17,9	250 - 286	307	24,5	282 - 331
6	124	5,9	118 - 130	168	15,5	152 - 183	191	15,5	175 - 206
7	148	15,0	133 - 163	239	23,2	216 - 262	260	14,8	245 - 275
8	124	10,1	114 - 134	196	16,3	180 - 212	223	19,2	204 - 242
9	116	9,1	107 - 125	167	30,7	136 - 198	262	16,4	246 - 278
10	119	7,2	104 - 118	169	21,2	148 - 190	218	17,7	200 - 236

Tableau I

Valeurs en Hz : - relevées de l'attaque, du maximum intermédiaire, de la fin du contour de la fréquence laryngienne F_ℓ

- calculées : de leur écart-type et des bandes à ±σ

locuteur	B(att.)	B(max.)	B(fin)	montée	σ (montée)	att./F_ℓ	max./F_ℓ	fin/F_ℓ	durée	σ(durée)
1	1,3	1,5	1,0	5,3	1,0	0,0	2,2	5,5	19	3,3
2	1,1	1,0	1,3	2,8	0,7	-0,4	2,2	3,5	15	4,2
3	0,8	1,1	1,1	3,4	0,9	0,0	1,4	3,7	15	3,7
4	1,0	1,7	1,7	4,2	0,8	-0,2	1,1	3,5	16	3,5
5	0,8	1,1	1,4	3,1	0,7	0,1	1,3	2,5	16	4,2
moyenne	1,0	1,3	1,3	3,3	0,8	-0,1	1,6	3,7	16	3,8
6	0,8	1,6	1,4	2,8	0,8	-1,6	1,0	2,1	15	4,4
7	1,8	1,7	1,0	4,5	0,9	0,2	4,3	5,0	18	3,9
8	1,4	1,4	1,5	5,1	0,9	0,1	4,1	5,2	17	4,0
9	1,3	3,2	1,1	5,4	1,2	-2,7	0,4	4,3	14	3,9
10	1,1	2,2	1,4	4,5	0,9	-0,4	2,6	4,8	22	4,2
moyenne	1,3	2,0	1,3	4,5	0,9	-0,9	2,5	4,3	17	4,1

Tableau II - Par locuteur et moyenne par sexe :

Valeurs en tons (calculées à partir du tableau I) des bandes (à ±σ) de l'attaque, du maximum intermédiaire et de la fin ainsi que les écarts entre les valeurs de l'attaque, du maximum, de la fin référencés par rapport à la F_ℓ des phrases énonciatives.
En cs les durées de la montée et leur écart-type.

	1	2	3	4	5
att. & max	0,31	0,19	0,48	0,42	0,22
att. & fin	0,16	-0,16	0,06	0,03	0,12
att. & durée	-0,04	0,03	0,01	-0,06	0,15
att. & montée	-0,44	-0,27	0,16	-0,12	0,10
max. & fin	0,09	-0,12	0,00	-0,18	-0,13
max. & durée	0,24	0,23	-0,12	-0,16	-0,11
max. & montée	-0,20	-0,05	-0,03	-0,30	-0,26
fin & durée	0,30	-0,02	0,02	0,27	-0,05
fin & montée	-0,22	0,50	0,64	0,60	0,57
durée & montée	-0,05	0,29	0,48	0,54	0,18

Tableau III

Coefficients de corrélation entre les différentes variables pour les cinq locuteurs masculins

	PHRASE DE BASE					
	énonciative			interrogative		
Perception %	énonciation	question	indécision	énonciation	question	indécision
contour 0	100	0	0	92	0	8
1	92	0	8	92	0	8
2	100	0	0	83	8	8
3	100	0	0	83	0	17
4	100	0	0	75	0	25
5	66	0	33	67	0	33
6	50	8	42	50	0	50
7	17	8	75	33	17	50
8	8	67	25	25	0	75
9	8	67	25	0	67	23
10	0	100	0	8	83	8
11	0	100	0	0	100	0
12	0	100	0	0	100	0

Tableau IV

Pourcentage des réponses selon les énoncés d'origine et les contours modifiés.

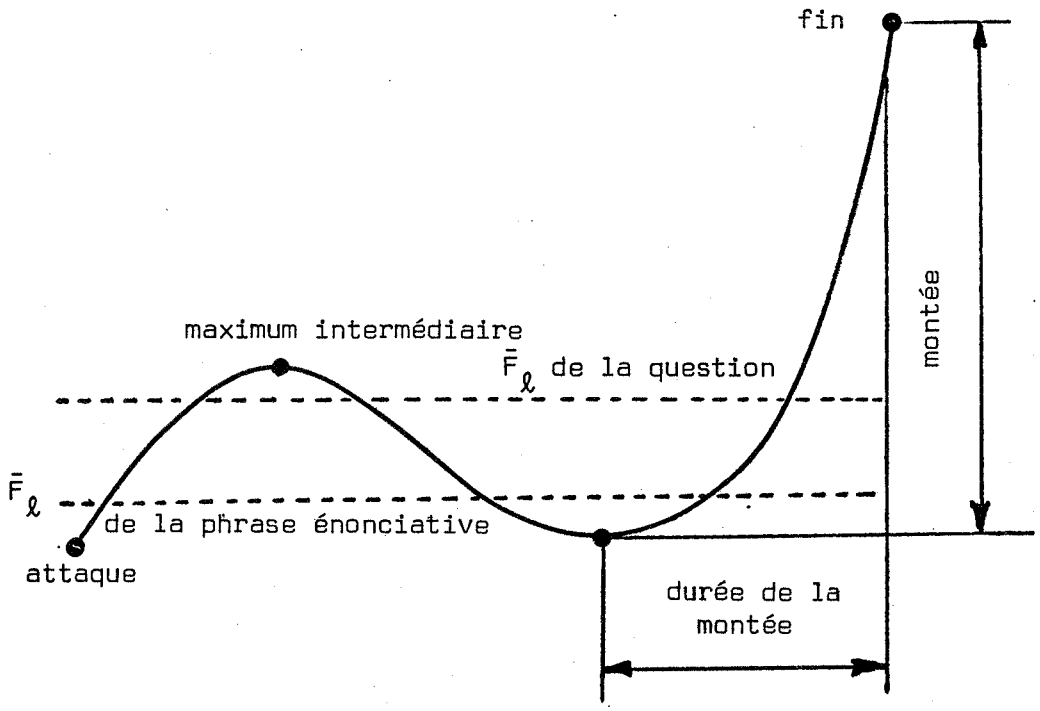


Figure 1. Mesures relevées sur le contour intonatif de la question, référencées par rapport à la fréquence laryngienne moyenne de la phrase énonciative.

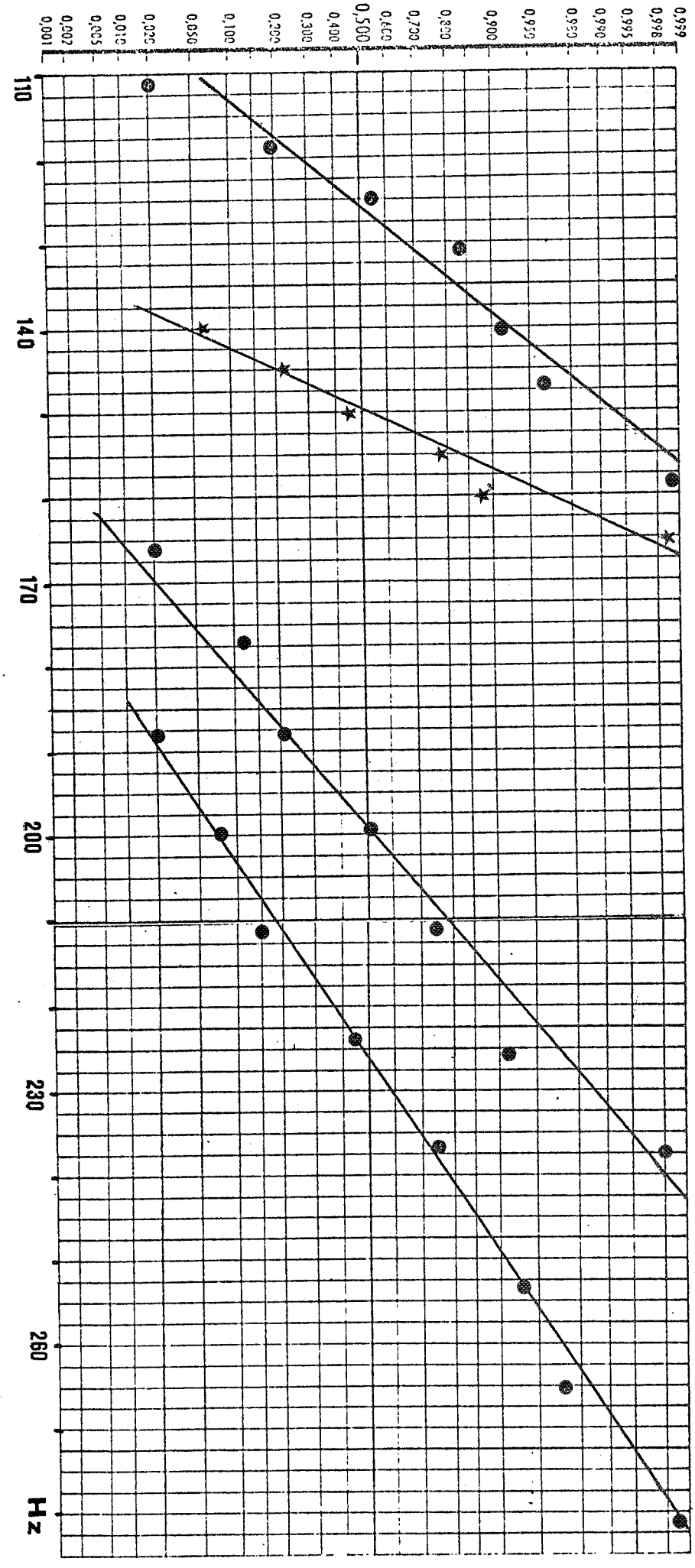


Figure 2. Fonction de répartition de l'attaque, du maximum intermédiaire et de la fréquence laryngienne moyenne pour les 45 phrases du corpus et le locuteur 8.

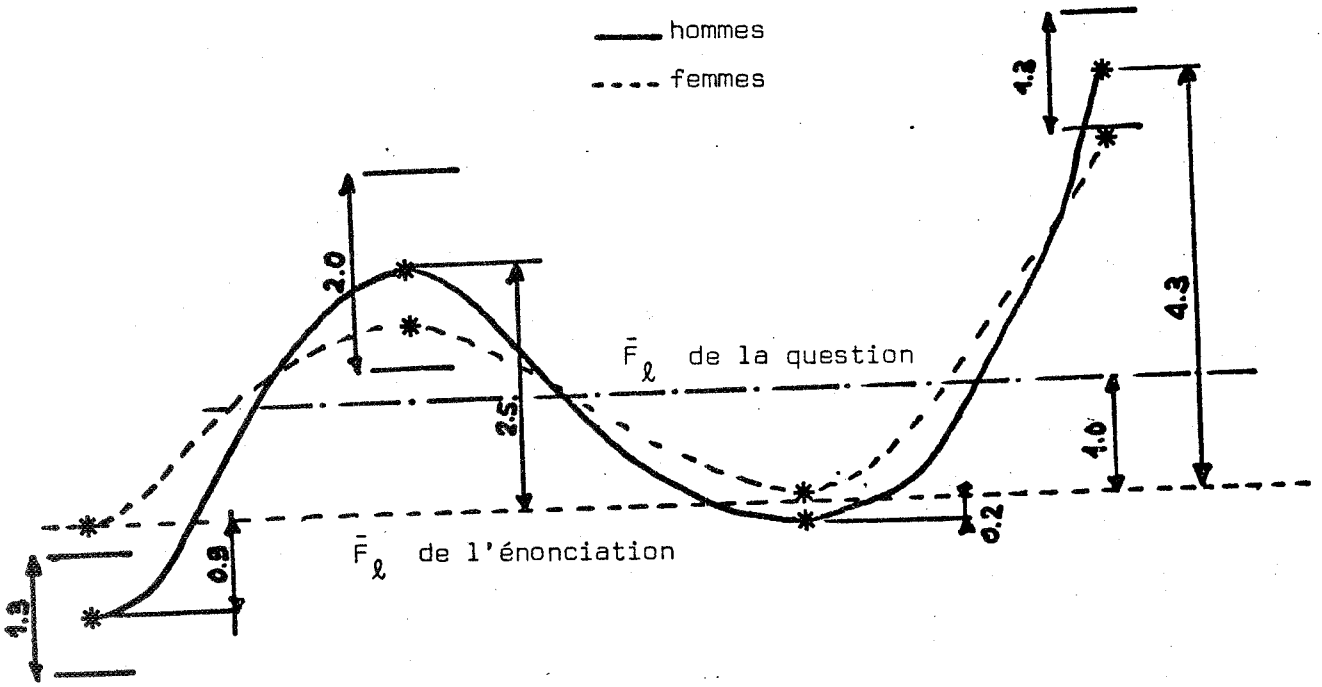


Figure 3. Contours type de la question totale, pour les locuteurs masculins et féminins référencés par rapport à la fréquence laryngienne moyenne de la phrase énonciative, les valeurs sont données en tons.

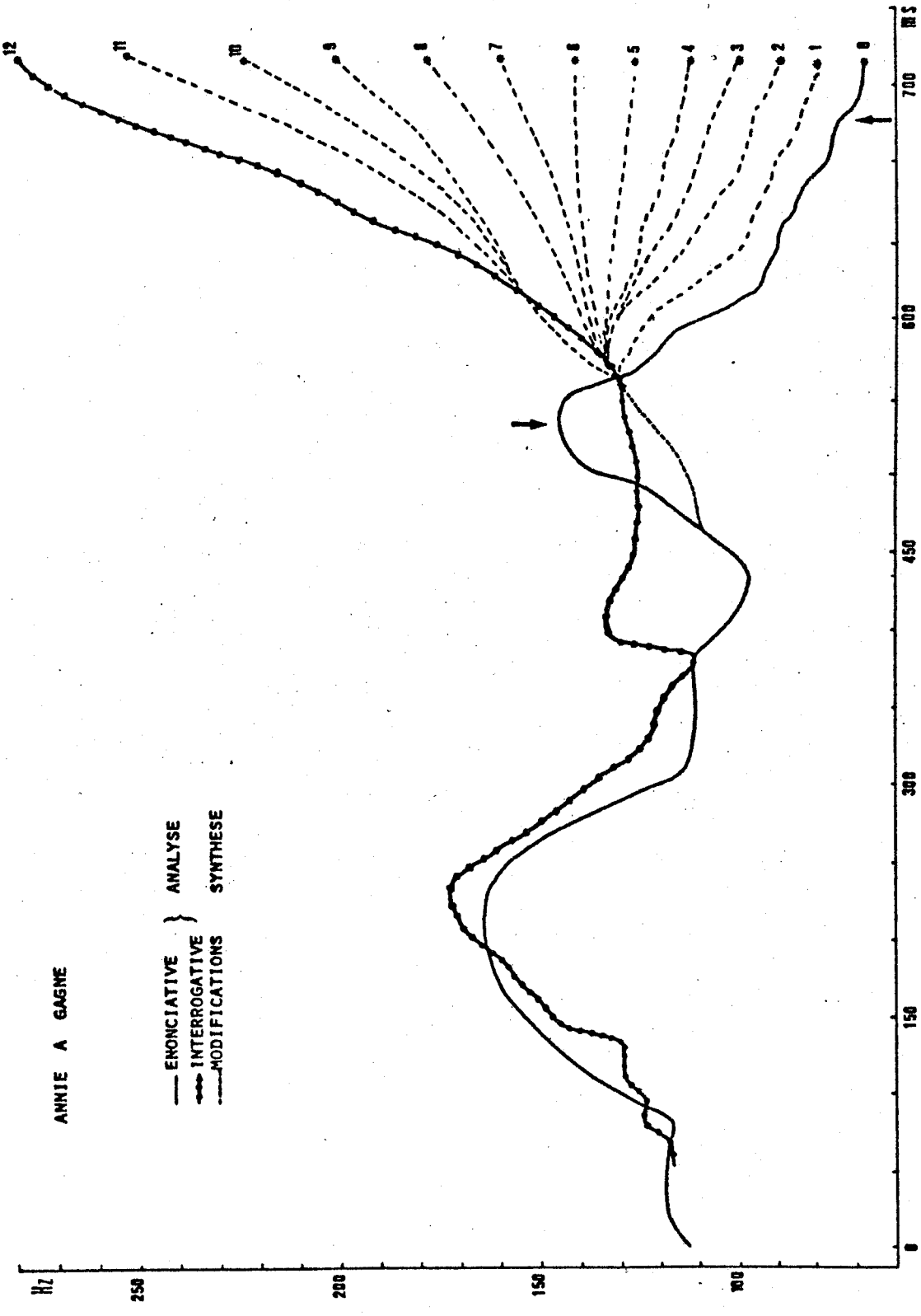


Figure 4. Phrases énonciatives et interrogatives analysées et modifiées à la synthèse (locuteur 8).
Les flèches indiquent les corrections qui ont été apportées à l'intensité des deux dernières syllabes.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

LE PROBLEME DES SOURCES DANS LA SIMULATION DYNAMIQUE DU TRACTUS VOCAL

par

A. BOURJAULT - A. CHEVILLARD - F. LHOTE

(Laboratoire d'Automatique de Besançon)

RESUME : Dans la plupart des modèles analogiques de l'appareil phonatoire les "sources" sont traitées à part et interviennent généralement de manière additive.Nous proposons une approche différente de ce problème en réalisant un analogue dynamique et en temps réel , rendant compte globalement des phénomènes phonatoire.

SUMMARY : In the main vocal tract analog models,the sources are treated aside and they appear as an addition.We propose here a different approach to this problem:we realize a dynamic analog in real time, taking lump into account the phenomenae of phonation.

LE PROBLEME DES SOURCES DANS LA SIMULATION DYNAMIQUE DU TRACTUS VOCAL

par

A. BOURJAULT - A. CHEVILLARD - F. LHOTE

(Laboratoire d'Automatique de Besançon)

LES PHENOMENES PHONATOIRES :

Il est généralement admis que la genèse des sons de la parole fait intervenir trois modes fondamentaux d'ébranlement : [12]

- périodique, pour les sons voisés
- aléatoire, pour les constrictives
- transitoire, pour les occlusives

(les constrictives et occlusives sonores étant obtenues par combinaison d'un ébranlement aléatoire et périodique d'une part, transitoire et périodique d'autre part).

Ces différents ébranlements sont modelés grâce aux mouvements articulatoires ; dans le cas des sons voisés, il s'agit simplement de la déformation des cavités buccales ; pour les occlusives et constrictives, engendrées par le rapprochement de deux organes, la nature de ces organes et le point d'articulation déterminent leur contenu informationnel.

Ces deux fonctions - la génération d'un ébranlement sonore brut et son "modelage" - ne sont pas dissociées l'une de l'autre lors de la phonation et agissent de pair.

LES ANALOGUES (ELECTRIQUES) DU TRACTUS VOCAL :

Dans le cadre des études sur la synthèse et la reconnaissance de la parole, nombreux sont ceux qui réalisent des analogues de l'appareil phonatoire.

L'analogie électrique utilisée au départ n'est valable que pour un conduit acoustique passif et présentant une forme fixe ; les modèles évolutifs sont souvent obtenus en faisant varier dans le temps les paramètres correspondants [1].

Cette démarche artificielle peut sembler impropre à la représentation des régimes transitoires, puisque les équations statiques sur lesquelles est basée l'analogie électrique sont inaptes à rendre compte des évolutions dans le temps de la section du canal vocal. Des

travaux plus récents permettront peut être de lever cette incompatibilité [2].

Une autre conséquence de ce procédé est que, dans tous ces modèles, les deux fonctions décrites précédemment sont traitées séparément.

"La source vocale", dont le comportement peut être rapproché de celui d'un oscillateur à relaxation, est simulée à part, soit sur un ordinateur [3], soit analogiquement [4]. Elle est généralement obtenue par analogie directe [5] ou à partir de l'étude d'un analogue mécanique (modèle à une masse ou à deux masses) [6]. Quant aux ébranlements transitoires et aléatoires, ils sont généralement induits par une source de pression, aux caractéristiques déterminées, ajoutée en parallèle. Enfin, la commande en temps réel présente dans ce cas d'importantes difficultés, qui se traduisent souvent par une relative complexité.

On peut donc penser que ces analogues sont en partie inaptes à rendre compte de l'ensemble des phénomènes phonatoires ; les "sources", traitées à part, n'apparaissent que sous forme additive dans la plupart des modèles.

NOUVELLE APPROCHE :

Notre contribution à ce problème des sources est de proposer une approche différente permettant une simulation effectivement dynamique et en temps réel.

En effet, il est naturel de considérer l'appareil phonatoire comme un processus globalement générateur de signaux, sans faire de distinction à priori entre les mouvements nécessaires à l'élaboration de l'ébranlement sonore brut (périodique, transitoire ou aléatoire) et ceux de sa modulation.

Ceci impose donc une simulation de tout l'ensemble larynx-conduit vocal ; celle-ci est basée sur un modèle mathématique, composé d'un système de deux équations aux dérivées partielles, rendant compte directement des phénomènes d'écoulement d'air dans un conduit de section A(x,t), non uniforme et variable dans le temps (avec l'hypothèse simplificatrice d'un écoulement monodimensionnel) [7] :

$$(1)\qquad \begin{cases} -\dfrac{\partial p}{\partial t} = \gamma p \dfrac{\partial u}{\partial x} + \gamma p\, u \dfrac{1}{A}\dfrac{\partial A}{\partial x} + \gamma \bar{p} \dfrac{1}{A}\dfrac{\partial A}{\partial t} \\ -\dfrac{\partial u}{\partial t} = \dfrac{c^2}{\gamma p}\dfrac{\partial p}{\partial x} \end{cases}$$

p(x,t) et v(x,t) représentant respectivement la pression et la vitesse en un point d'abcisse x et à l'instant t.

A partir de ce modèle mathématique, nous ne cherchons pas à traduire les phénomènes mécaniques eux-mêmes, mais le processus par lequel ces derniers conditionnent la génération et la modulation des sons.

La fonction d'aire A(x,t) intervient à deux niveaux :

- d'une manière multiplicative, par le terme $\frac{1}{A}\frac{\partial A}{\partial x}$, qui traduit notamment les accidents qui se présentent à un instant donné le long du canal vocal ;

- d'une manière additive, par le terme $\frac{1}{A}\frac{\partial A}{\partial t}$, qui représente la vitesse de variation de la section ; ce terme est généralement omis dans les autres modèles, mais nous avons montré [8] que dans certains cas, il jouait un rôle essentiel.

Les équations (1) ont été discrétisées par rapport à la variable d'espace x, afin de leur substituer un système d'équations différentielles ordinaires, susceptibles d'être traitées par des techniques analogiques;

Un ensemble hybride, composé d'un ordinateur et d'un calculateur analogique rapide mis au point au laboratoire [9] permet de réaliser ces équations et leur commande en temps réel. Cette commande s'effectue à partir de la fonction d'aire A(x,t), mémorisée sur l'ordinateur puis distribuée localement à chaque tranche [10].

Les ébranlements sonores périodiques ou transitoires apparaissent généralement en des points du tractus où la section peut momentanément s'annuler, ce qui introduit des coefficients infinis dans les équations (1). Pour lever cette difficulté, nous avons pensé donner à ces coefficients les plus grandes valeurs possibles permises par les techniques de simulation, grâce à un calibrage adéquat ; il semble d'ailleurs raisonnable de supposer que la section n'est jamais tout à fait nulle.

Des expériences ont été effectuées à partir d'un modèle composé de 13 tranches. Dans un premier temps, elles ont porté sur la simulation des voyelles, notamment les voyelles [i] et [e], sur la base des données de Chiba [11] . Nous passons actuellement aux divers types de consonnes ; les premiers résultats obtenus sur les occlusives telles que [p] sont très encourageants.

Une autre solution possible consisterait à considérer le tractus vocal comme étant formé de deux tuyaux distincts avec des conditions de raccordement adéquats au point où la section s'annule. Mais nous n'avons pas encore à ce jour expérimenté ce principe.

En ce qui concerne la source de bruit - aléatoire - qui peut prendre naissance lors d'une constrictive, nous pensons qu'il suffit de simuler les circonstances de sa génération, plutôt que les causes (qui sont intraduisibles par des équations macroscopiques).

En effet, il existe dans ce cas une interaction certaine entre l'air et les parois, que nous traduisons au niveau du modèle par des vibrations rapides et de faibles amplitudes de la section, au voisinage du point de constriction.

CONCLUCTON :

Ces premières expériences nous prouvent la validité de notre modèle, plus proche, nous semble-t-il, des phénomènes naturels. En liant les deux fonctions fondamentales de la production de parole, nous ne traitons plus les sources à part ; ce qui, outre la simplification apportée au niveau de la réalisation et de la commande en temps réel, permet d'obtenir un analogue réellement dynamique et permet d'attendre une parole plus naturelle.

B I B L I O G R A P H I E :

[1] H.K. Dunn .The calculation of vowel resonances and an electrical vocal tract . J A S A . 22 ,740-753, (1950)

G. Fant .Acustic theory of speech production . 's Gravenhage. Mouton and Co .

R. Jakobson,G. Fant,M. Halle . Preliminaries to speech analysis. M.I.T. Press (1961)

G. Rosen . Dynamic analog speech synthetizer . JASA . 30 ,201-209 (1958)

K.N. Stevens,R.P. Bastide and C.P. Smith .Electrical synthetizer of continuous speech. J A S A . 27, 207 (A) (1955)

K.N. Stevens , S. Kasovski , G. Fant . An electrical analog of the vocal tract . J A S A . 25 ,734-742 , (1953)

[2] B. Guerin . Simulation du conduit vocal . 3ièmes J.E.P. Lannion (??)

J. Genin . Simulation du conduit vocal . 3ièmes J.E.P Lannion (??)

J.L. Courbon . Simulation du conduit vocal en technologies analogiques. 6ièmes J.E.P. Toulouse (75)

[3] J.L. Flanagan , L. Landgraf . Self oscillating source for vocal tract synthetizer . IEEE Trans. Audio. and Elec. AU 16 , 1 , 57-64 . (1968)

4] T.H. Crystal . Model of larynx activity during phonation . MIT.QPR 78 . July 15 , 212-219 (1965)

[5] J. Paillé . Source vocale pour synthétiseurs à formants . Revue d'acoustique , 6 , 111-114 . (1969)

[6] K. Ishizaka , J.L. Flanagan . Synthesis of voiced sounds from a two-mass model of the vocal cords . The Bell System Technical Journal . Vol. 51 , N° 6, July 72.

[7] A. Bourjault . A Chevillard . Nouvelle approche pour la simulation dynamique du canal vocal . Revue d'acoustique;34, 25-29 . (1975)

[8] A. Chevillard. F. Lhote . A. Bourjault . Modèle dynamique du tractus vocal . 8th ICA .Londres (1974)

[9] P. André . A Bourjault . A. Chevillard . J.M. Henrioud . Calculateur analogique rapide pour la simulation en temps réel des phénomènes de phonation . Simulation'75 . Zürich (1975)

[10] A. Bourjault. A. Chevillard . F. Lhote . Hybrid simulation of vocal tract in real time . 8th AICA Congess , Delft (1976)

[11] T. Chiba . M. Kajiyama . The vowel:its nature and stucture . Tokyo . Society phonetic of Japan (1958)

[12] P. Simon . Les consonnes françaises . (1967)

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

SIMULATION EN TEMPS REEL DU CONDUIT VOCAL

J.L. COURBON, J. GENIN
C.N.E.T. - LANNION

RESUME : La présente réalisation est une simulation câblée, fonctionnant en temps réel du conduit vocal. Les paramètres de commande sont de nature physiologique ; ils décrivent les configurations du larynx et du conduit vocal. Un circuit permet l'insertion automatique de bruit si les conditions requises sont réalisées. Un programme fonctionnant actuellement en temps différé permet la commande articulatoire du modèle.

SUMMARY : This realization is a hardware real time simulation of vocal tract. The control parameters are physiological ones. They describe the configuration of larynx and vocal tract. An added circuit inserts noise automatically when necessary conditions are effective. Using a computer program, we can control the model with articulatory parameter.

SIMULATION EN TEMPS REEL DU CONDUIT VOCAL *

J.L. COURBON, J. GENIN

C.N.E.T. LANNION

I - INTRODUCTION

Des dispositifs des plus divers ont été utilisés à fin de synthèse de la parole. Ils sont généralement composés de deux parties simulant, l'une, la fonction de transfert du conduit, l'autre, le fonctionnement des cordes vocales. Bien sûr, une telle schématisation est très grossière car certains phénomènes physiquement associés à une part du modèle peuvent parfois être pris en compte par l'autre. C'est souvent le cas des zéros apparaissant dans le spectre de l'onde glottique par exemple.

De telles méthodes donnent de bons résultats pour des applications en transmission ou en "analyse-synthèse" mais l'interprétation et l'utilisation de paramètres de nature physiologique est difficile.

Dans cette communication, nous décrivons notre système de simulation du conduit vocal. La partie câblée du système se compose du simulateur de conduit vocal qui a été présenté aux 6èmes Journées d'Etudes à TOULOUSE [1] et d'un simulateur de source glottique. L'intérêt de cette structure est de tenir compte des interactions existant entre ces deux éléments. Nous avons ajouté un nouveau circuit permettant l'insertion automatique de pertes et de bruit au niveau des constrictions. Nous citons enfin rapidement l'élaboration d'un programme de commande articulatoire fonctionnant actuellement en temps différé.

II - LES MODES DE SIMULATION DE LA SOURCE GLOTTIQUE

Il est généralement admis que la forme de l'onde produite par les cordes vocales a une importance pour la qualité de la parole de synthèse [2] . Les premières études dans ce sens étaient semblables à celles de J.PAILLE [3] qui cherchait à construire un circuit électronique simple fournissant un signal d'excitation plus réaliste que la classique séquence d'impulsions fines. Cependant ce modèle est trop simple pour s'accomoder de la nature physiologique des phénomènes.

La réalisation de ROTHENBERG et al [4] est d'une conception voisine. Les auteurs ont choisi trois paramètres de commande, paramètres physiques décrivant la forme du signal obtenu, qu'ils pensent bien corrélé avec des paramètres physiologiques. Le problème est comment les commander à partir de la connaissance de l'état des cordes vocales. Le modèle présenté par FLANAGAN, MATSUDAIRA et ISHIZAKA [5, 6,7] est satisfaisant sur le plan physiologique. Ses paramètres de commande ne sont pas fréquence,

* Cet article a été présenté sous une forme voisine à : 1976 IEEE conference on Acoustics, Speech and Signal Processing (PHILADELPHIE, avril 1976)

amplitude et facteur de forme du signal émis, mais pression sous-glottique, ouverture au repos des cordes vocales et effort musculaire de tension.

Notre simulateur de cordes vocales est une réalisation câblée du "modèle à deux masses" simulé sur calculateur par FLANAGAN et al. Nous allons rapidement décrire cette réalisation.

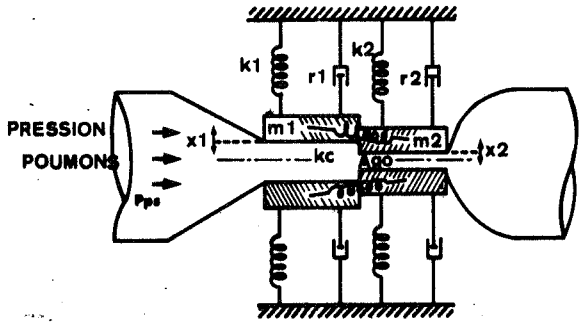


Fig 1. Le modèle à deux masses.

III - LA PARTIE MECANIQUE

L'intérêt du modèle à deux masses est de tenir compte des différences de phase observées entre les mouvements des parties haute et basse des cordes vocales.

Telle que représentée sur la figure 1, la partie mécanique se ramène à deux couples de masses M, M'_1, M_2, M'_2, reliées à la "masse" par les ressorts k_1, k_2, les amortissements r_1, r_2, également reliés entre eux par le ressort k_{12}. Les équations du modèle mécanique sont :

$$m_1 \frac{d^2x_1}{dt^2} + r_1 \frac{dx_1}{dt} + k_1 x_1 = -k_{12}(x_1 - x_2) + f_{p1} \qquad m_2 \frac{d^2x_2}{dt^2} + r_2 \frac{dx_2}{dt} + k_2 x_2 = k_{12}(x_1 - x_2) + f_{p2}$$

où f_{p1}, f_{p2} sont les forces appliquées à ce système par les phénomènes aérodynamiques. Dans ces équations le second membre représente l'action de couplage entre les mouvements des deux couples de masse. La simulation du fonctionnement mécanique de chaque couple de masses est effectué par le circuit représenté sur la figure 2.

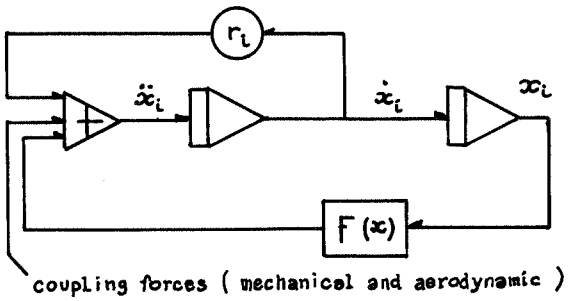


Fig. 2 Simulation du comportement d'un couple à deux masses.

Le coefficient d'amortissement r_i reçoit deux valeurs différentes selon que les masses sont écartées ou accolées. F (x), force appliquée par les ressorts est un terme non linéaire qui obéit aux formules (2)

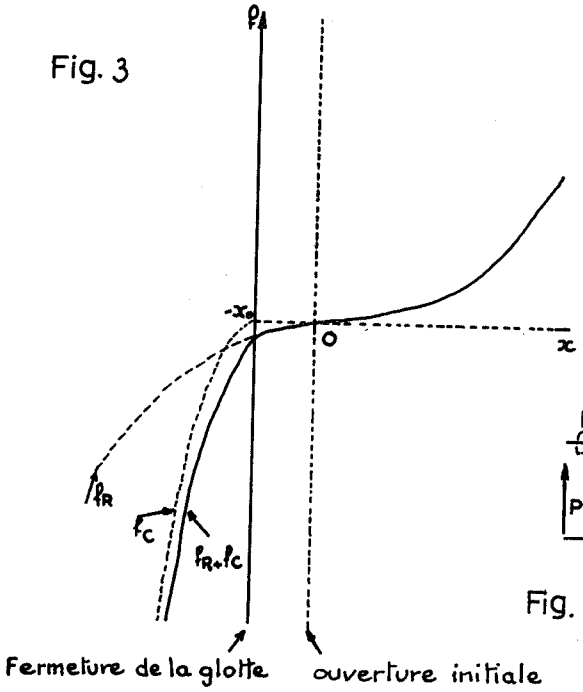
$$F(x) = F_R(x) + F_c(x) \qquad F_c(x) = 0 \quad \text{pour } x \geq -x_o$$

$$F_R(x) = k\,x\,(1 + nk\,x^2) \qquad F_c(x) = k'(x+x_o)(1+nk'(x+x_o)^2) \quad \text{pour } x < -x_o$$

k k' valent approximativement 50 à 100 dyne/cm
nk, nk' : 100 cm^{-2}
xo est l'écart au repos des cordes vocales
x est ici la mesure du déplacement des masses par rapport à leur position de repos.

La figure 3 montre l'évolution de la force F(x) en fonction de l'écart des masses. Les circuits non linéaires nécessaires utilisent des générateurs à diodes.

Fig. 3



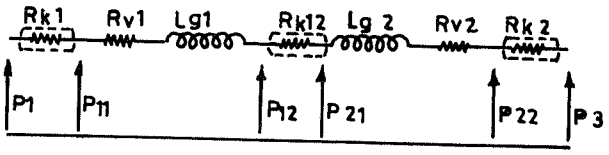


Fig. 4 Schéma équivalent du phénomène aérodynamique.

Forces appliquées à chaque couple de masses

IV - LA PARTIE AERODYNAMIQUE

La partie aérodynamique peut être vue comme une boîte noire dont les entrées sont :

Ag1, Ag2, écarts de chaque couple de masses.
P_1, P_3, pression d'air en amont et en aval du système.

Le circuit fournit comme résultat :

P_{m1}, P_{m2}, pression d'air moyenne sous chaque couple de masses

u , débit volumique à travers le système.

Le circuit équivalent est montré sur la figure 4. Il obéit aux équations (3). Dans ces formules, ρ est la masse volumique de l'air, μ son coefficient de viscosité et S_3 l'aire d'entrée du conduit vocal. f_{p1} et f_{p2} sont les forces aérodynamiques appliquées à chaque couple de masses. L est la longueur de la fente glottique.

Rk représente l'effet de BERNOUILLI aux différents changements de section,

Rv les pertes par frottement et Lg l'effet de l'inertie de l'air. Le circuit de simulation est schématisé sur la figure 5.

$$(Rk_1 + Rk_2 + Rk_{12} + Rv_1 + Rv_2)\,U + (Lg_1 + Lg_2)\frac{dU}{dt} + P_3 - P_1 = 0$$

$$Rk_1 = 1{,}37\,\frac{\rho |U|}{2\,Ag_1^2} \qquad Rk_2 = -2N(1-N)\,\frac{\rho |U|}{2\,Ag_2^2} \qquad N = \frac{Ag_2}{S_3}$$

$$Rv_1 = \frac{12\,\mu\, d_1 L^2}{Ag_1^3} \qquad Rv_2 = \frac{12\,\mu\, d_2 L^2}{Ag_2^3} \qquad Rk_{12} = \frac{\rho}{2}|U|\left(\frac{1}{Ag_2^2} - \frac{1}{Ag_1^2}\right)$$

$$Lg_1 = \frac{\rho\, d_1}{Ag_1} \qquad Lg_2 = \frac{\rho\, d_2}{Ag_2}$$

$$f_{p1} = Ld_1 Pm_1$$

$$Pm_1 = \frac{P_{11} + P_{12}}{2} = P_1 - Rk_1 U - \frac{Rv_1 U + Lg_1\, dU/dt}{2} \qquad f_{p2} = Ld_2 Pm_2$$

$$Pm_2 = P_3 + Rk_2 U + \frac{Rv_2 U - Lg_2\, dU/dt}{2}$$

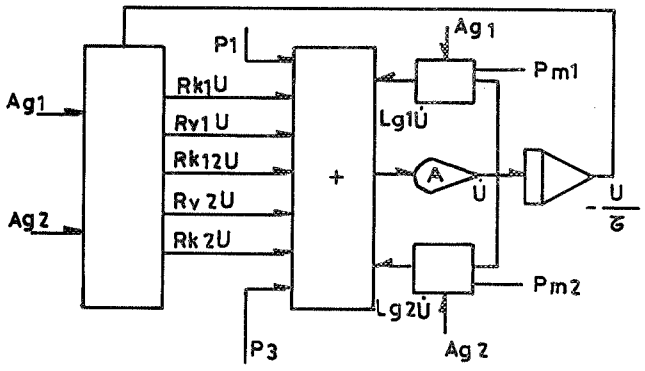


Fig. 5 La partie aérodynamique

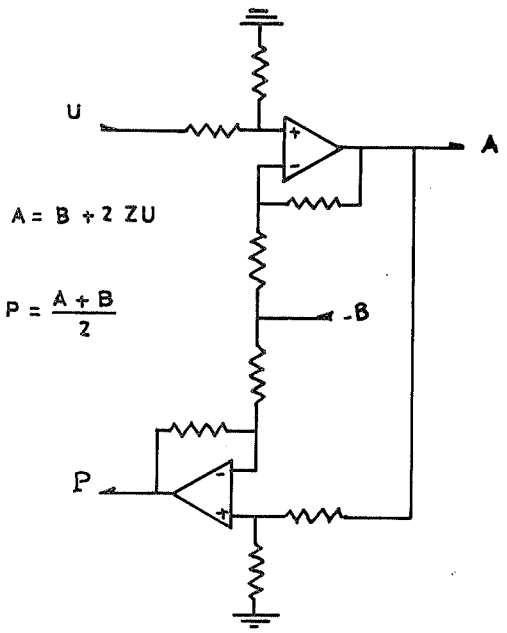


Fig. 6 Interface entre source et simulateur du conduit vocal.

V - INTERACTIONS AVEC LE CONDUIT VOCAL

Commandé par la valeur de la pression de l'air dans les poumons et l'état physiologique des cordes vocales, le simulateur fournit la valeur du débit volumique d'air qui traverse le système.Le conduit vocal doit fournir en retour la valeur de la pression qu'il oppose à ce fonctionnement.

Notre modèle de conduit vocal a déjà été décrit [1]. Il s'agit d'une ligne de retard bidirectionnelle composée de 12 sections. Le signal incident est :

$$A = p + Zu$$

le signal réfléchi est :

$$B = p - Zu$$

où Z est l'impédance d'entrée du conduit vocal(dépendant de l'aire de la première section).La figure 6 schématise le circuit d'interface utilisé entre conduit vocal et simulateur de cordes vocales.

VI - FONCTIONNEMENT DES CORDES VOCALES, PREMIERS RESULTATS

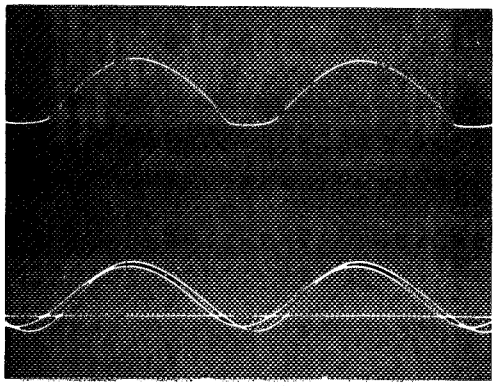
La figure 7 montre quelques résultats obtenus sans interaction dans différentes configurations physiologiques. Sont représentés le débit volumique et les élongations des deux couples de masses. Pour négliger l'effet du conduit vocal il suffit de ne pas appliquer la pression P_3.

La figure 8 montre l'effet de la forme du conduit vocal. Sont représentés les débits volumiques à l'entrée et à la sortie du conduit vocal.

La figure 9 montre le résultat de la première expérience de fonctionnement dynamique. Tout d'abord, nous avons fixé les paramètres de la source vocale à des valeurs provoquant normalement l'oscillation. Nous avons ensuite fermé le conduit vocal à l'extrêmité des lèvres. Dans ces conditions, il n'y a pas de chute de pression à travers la glotte (pression P_3 ramenée par le conduit vocal), pas de débit d'air, pas d'oscillation. D'un seul coup, nous ouvrons le conduit vocal et lui donnons la configuration propre à la voyelle `a'.La pression tombe en aval de la glotte,et l'oscillation commence. La figure 9 montre l'évolution du débit glottique et de la pression d'air au niveau des lèvres. On remarquera l'explosion initiale et l'évolution de la période du fondamental, de l'amplitude et de la forme de l'onde glottique. Le son obtenu est la syllabe `pa'. En répétant une deuxième fois l'expérience le système dit son premier mot : "papa" ! ...

VII - L'INSERTION DE PERTES DE CHARGE ET DE BRUIT AU NIVEAU DES CONSTRICTIONS DU CONDUIT VOCAL

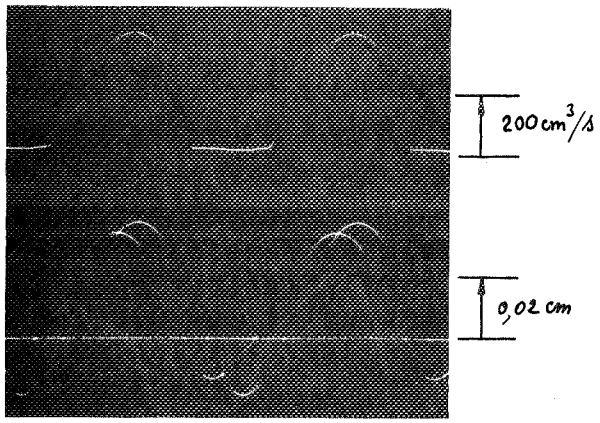
Le simulateur a été doté d'un système de détection des conditions de naissance de bruit et de son d'injection de manière à pouvoir produire les fricatives lors d'une constriction (langue contre le palais ou les dents) ou lors d'une ouverture brusque. Dans ce cas, des pertes énérgétiques causées par les tourbillons du flux sont également à insérer : elles se traduisent par une chute de pression additionnelle (résistance à l'écoulement) [8].



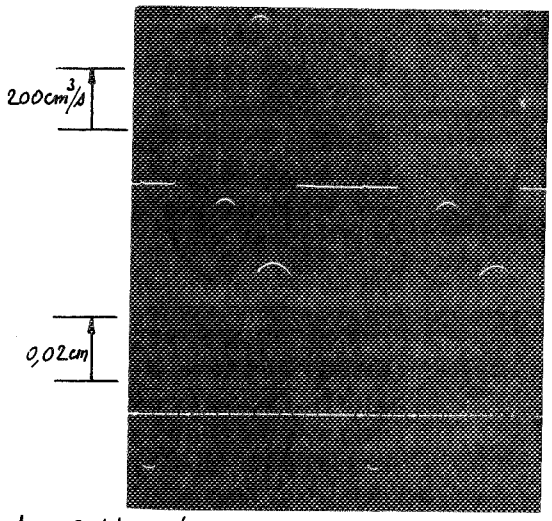
$k_1 = 80$ kdyne/cm
$k_2 = 8$ kdyne/cm
$k_{12} = 62,5$ kdyne/cm
$p_1 = 8$ cm H_2O
$X_0 = 0,02$ cm
$F = 134$ Hz

FORMES DU DEBIT ET DU MOUVEMENT DES MASSES.

Fig. 7



$k_1 = 125$ kdyne/cm $k_2 = 8$ kdyne/cm
$k_{12} = 12,5$ kdyne/cm $p_1 = 8$ cm H_2O
$X_0 = 0,02$ cm $F = 143$ Hz



$k_1 = 80$kdyne/cm $p_1 = 10$ cm H_2O
$k_2 = 8$ kdyne/cm $X_0 = 0,02$ cm
$k_{12} = 12,5$ kdyne/cm $F = 138$ Hz

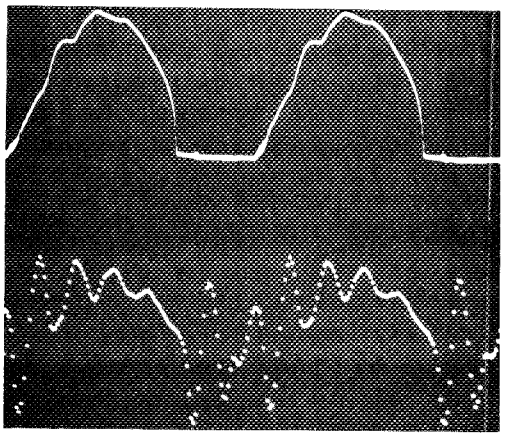
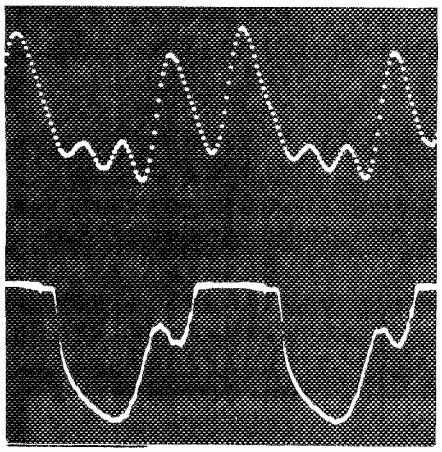


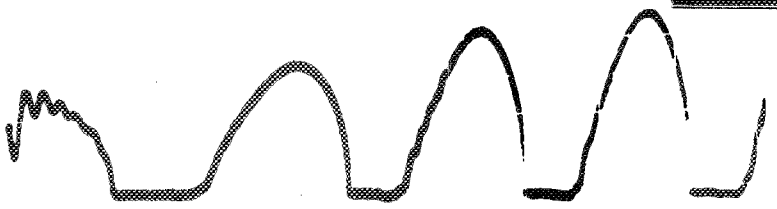
Fig. 8

Débits à La glotte et aux lèvres pour les voyelles:

A

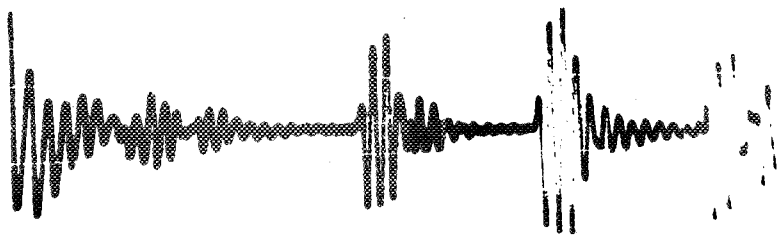


O



Vitesse volumique au niveau de La glotte.

Fig. 9



Pression aux lèvres

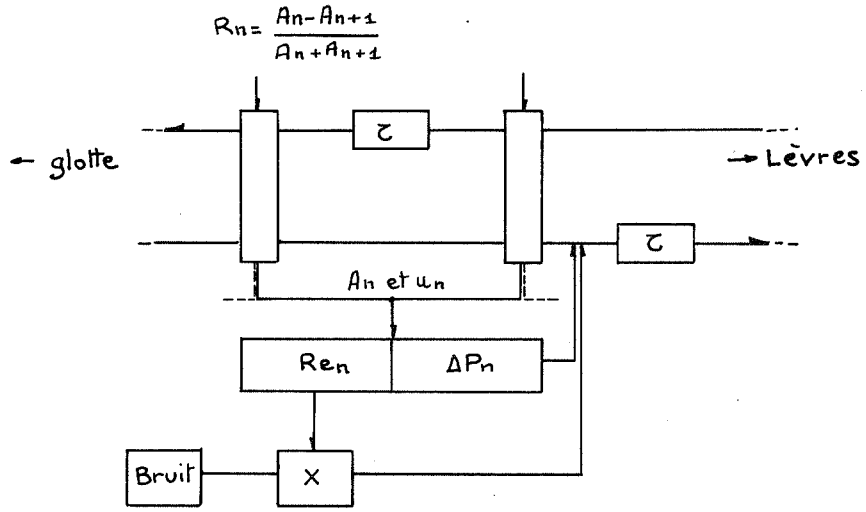


Fig. 10 Production du bruit dans le conduit vocal.

On utilise le nombre de Reynolds $Re = \frac{\rho u.l}{\mu}$ ou u est la vitesse d'écoulement, l la largeur du conduit, ρ la densité de l'air et μ le coefficient de viscosité, pour détecter le passage de l'écoulement laminaire à l'écoulement turbulent [9]. Lorsque ce nombre, calculé en chaque point du conduit vocal, dépasse un certain seuil Rec, on insère en aval du point de décision une puissance de bruit $Ps = K(Re^2 - Rec^2)$ dont le spectre est constant. De plus la pression est affaiblie d'une quantité $\Delta P = \rho u^2/2$.

La réalisation est entièrement analogique et ne nécessite aucun paramètre de commande : le fonctionnement est automatique à partir des valeurs de débit et d'aire existant à tout moment et en chaque point du conduit vocal.

VII - MODELE ARTICULATOIRE DE COMMANDE

L'intérêt de la simulation du conduit vocal pour la synthèse de parole réside dans l'utilisation d'un modèle plus proche de la nature physiologique des phénomènes. Aussi est-il important de tenir compte de contraintes apportées par la constitution des organes phonatoires en utilisant un modèle articulatoire. Celui-ci, commandé par des paramètres tels que la position de la langue, des machoires..., permet de calculer la fonction d'aires correspondante pour le conduit vocal.

Notre travail dans ce sens aboutit à l'implantation d'un programme conversationnel permettant l'utilisation et la modification de la coupe sagittale du conduit vocal. Ce programme permettra prochainement l'envoi direct au simulateur des fonctions d'aires calculées. La figure 11 montre un exemple de schéma fourni sur une console de visualisation. Les modifications graphiques sont définies en utilisant le réticule commandable à l'aide de deux potentiomètres.

CONCLUSION

Nous disposons à ce jour de trop peu de résultats d'ensemble. Les éléments dont il vient d'être question sont en effet des circuits indépendants que nous commençons seulement à relier les uns aux autres.

Cependant, nous voyons ici, non pas un organe de synthèse de parole automatique économique, mais un appareil assez puissant d'analyse des phénomènes. Notre but est en effet de rechercher la simulation la plus fidèle et d'obtenir la meilleure qualité possible, quitte à simplifier les matériels pour conduire à des réalisations économiques. Par ailleurs, ces différents éléments répondant à une certaine modularité, il nous paraîtrait intéressant de les utiliser en relation avec d'autres équipements de synthèse élaborés par d'autres équipes. Ici, nous pensons tout particulièrement au simulateur de la source glottique dont la conception, l'état d'achèvement (et de volume) sont mieux adaptés à des déplacements...

Mis à part le travail des auteurs, le présent article rassemble le fruit du travail de M. CARCAUD (stagiaire de 3ème cycle) pour la source vocale, M. et Mme GILLET, Melle SAILLARD (stagiaires) pour le modèle articulatoire, M. DUHAMEL (stagiaire) pour l'insertion de bruit, M. ROUMIGUIERE (CNET-ETA) à différents niveaux, et le service de technologie du Département pour les réalisations câblées.

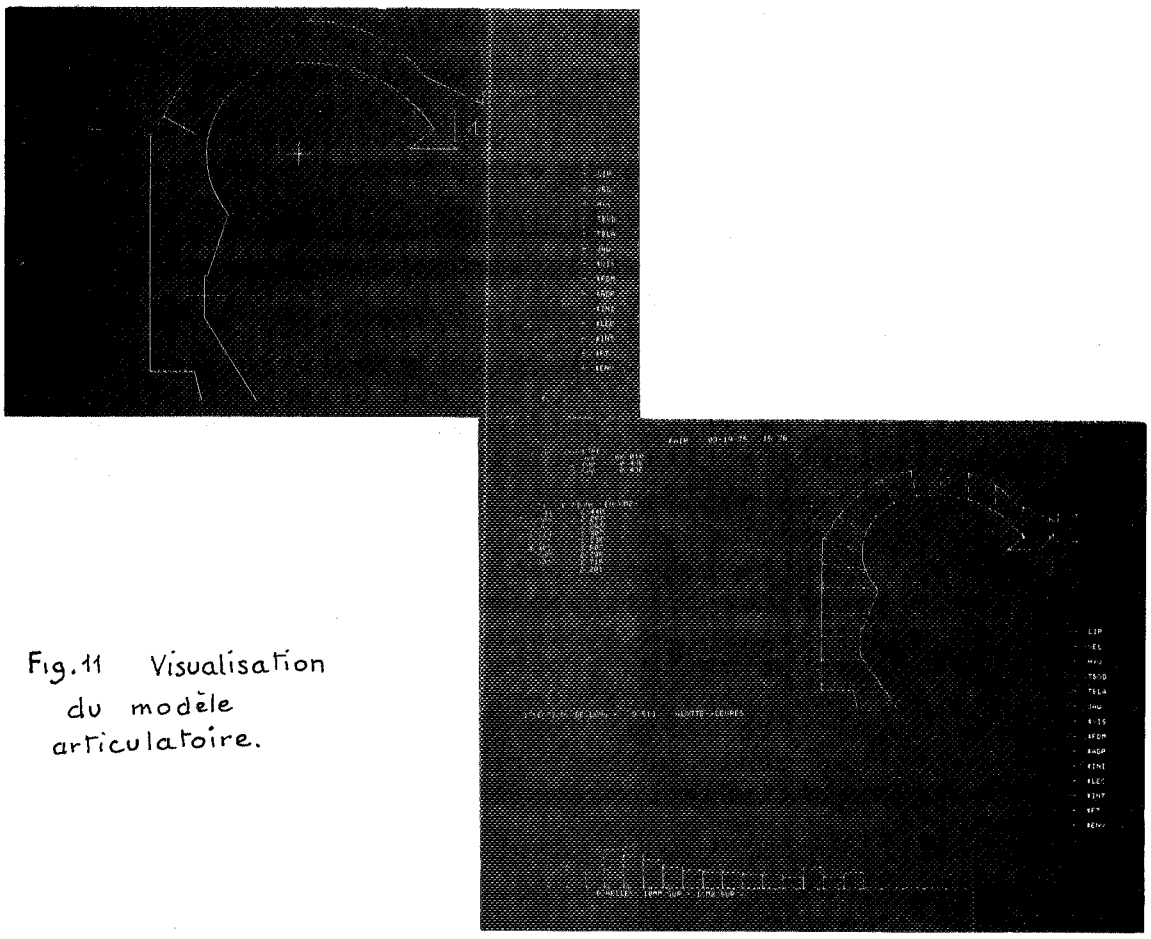


Fig.11 Visualisation du modèle articulatoire.

BIBLIOGRAPHIE

1. J.L. COURBON - "Simulation du conduit vocal en technologies analogiques" - 6ème journées d'études sur la parole organisées par le G.A.L.F. - Toulouse 28-30 mai 1975 pp. 327-338.

2. A.E. ROSENBERG - "Effect of glottal pulse shape on the quality of natural vowels". J.A.S.A. Vol. 29 n° 5 (1967) pp. 626-631

3. J. PAILLE - "Source vocale pour synthetiseurs à formants" Revue d'Acoustique n° 6 (1969) pp. 111-114

4. M. ROTHENBERG, R. CARLSON, B. GRANSTROM, J. LINQVIST-GAUFFIN - "A three-parameter voice source for speech synthesis".

5. J.L. FLANAGAN, L.L. LANDGRAF - "Self oscillating source for vocal tract synthesizers." - I.E.E.E. transactions on audio and electroacoustics. - Vol. AU-16 n° 1 March 1968.

6. K. ISHIZAKA, M. MATSUDAIRA - "What makes the vocal cords vibrate" - 6th International congres on acoustics - TOKYO August 1968 pp. B1-3.

7. K. ISHIZAKA, J.L. FLANAGAN - "Synthesis of voice sounds from a two-mass model of the vocal cords". B.S.T.J. Vol. 51 n° 6 July-August 1972.

8. R. GUERIN - "Simulation du conduit vocal" - Compte rendu des journées d'études sur la parole 1972 C.N.E.T.-LANNION pp. 129-141

9. MEYER EPPLER - "Zum Erzeugungs mechanismus der Geräuschlaute." - Phonetik 7 - 1953 - pp. 196-212.

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 au 2' mai 1976

Synthétiseur numérique à prédiction linéaire pour unité de réponse vocale

par

I.I. EL MALLAWANY

CNET - LANNION

RESUME :

La détermination des conditions optimales de fonctionnement d'un vocoder fondé sur la prédiction linéaire est un préalable à toute réalisation matérielle de la partie synthétiseur. L'intégration de ce dernier dans une unité de réponse vocale suppose que l'on ait trouvé le moyen d'atteindre les faibles débits (< 5 Kbit/s) sans dégradation notable de la qualité de la parole de synthèse. Les conditions optimales de fonctionnement d'un tel vocoder sont résumées. Les caractéristiques et l'organisation d'un synthétiseur travaillant en temps réel sont décrites.

SUMMARY :

The optimal working conditions of a linear predictive vocoder have to be determined prior to the hardware implementation of the synthesizer subsystem. In order to insert the latter in a vocal response unit one must have found the means of attaining low transmission rates (< 5Kbit/s) without significant deterioration of the synthesized speech quality. The optimal working conditions of the vocoder are briefly summarized. The characteristics and organization of a real-time synthesizer are described.

TITRE : SYNTHETISEUR NUMERIQUE A PREDICTION LINEAIRE POUR UNITE DE REPONSE VOCALE

AUTEUR : I.I. EL MALLAWANY

I - INTRODUCTION

Le support le plus naturel pour la communication homme-machine est la parole humaine. Dans ce contexte, le poste téléphonique représente le terminal le moins cher et le plus répandu et dont l'utilisation est simple. Ces considérations ont motivé l'étude d'organes de synthèse et de reconnaissance de la parole. Les différents aspects du traitement du signal qui se présentent dans ces deux organes se trouvent réunis dans un système vocoder. Les vocoders à canaux et à formants étant bien connus, il nous a paru utile d'étudier un système d'analyse-synthèse fondé sur la prédiction linéaire [1] , à titre de comparaison, et ce d'autant plus que la partie synthétiseur ne fait appel qu'à des calculs simples.

Cette étude de vocoder mettant en oeuvre la prédiction linéaire a eu trois objectifs :

- choisir la méthode qui est de réalisation la plus simple donc la plus économique et susceptible de fonctionner en temps-réel,
- définir les conditions optimales de fonctionnement de ces vocoders,
- parvenir à une parole de synthèse de qualité satisfaisante aux faibles débits (≤ 5 Kbit/s).

II - LA PREDICTION LINEAIRE

Le principe de la méthode [1] est qu'en l'absence temporaire d'une source d'excitation, un échantillon de parole, s_n, peut être prédit avec exactitude à l'aide d'une somme pondérée linéairement de p échantillons le précédant immédiatement. Dans cette hypothèse la fonction de transfert de l'appareil vocal peut être représentée par un filtre numérique récursif ne possédant que des pôles [1] de la forme

$$H(z) = G/(1 - \sum_{k=1}^{p} a_k z^{-k}) \qquad (1)$$

Le détail des hypothèses et des calculs intervenant dans l'évaluation des coefficients, a_k , est donné ailleurs [1]. On note que le filtre récursif (1) peut être transformé en d'autres formes canoniques, notamment, des filtres à structure en échelle dont les paramètres $\{k_i\}$ sont directement dérivables des $\{a_k\}$ [1], l'inverse étant vrai. Les $\{k_i\}$ représentent les coefficients de réflexion aux jonctions consécutives entre les sectionscylindriques d'égale longueur constituant un conduit acoustique équivalent à l'appareil vocal. Si A_i représente l'aire de la section i, il vient :

$$k_i = (B_i - 1)/(B_i + 1) \qquad (2)$$

où

$$B_i = A_i / A_{i+1} \qquad (3)$$

= rapport d'aire entre sections adjacentes.

III - VOCODER A PREDICTION LINEAIRE

L'organisation d'un vocoder dépend du modèle de production de la parole qui est retenu et qui représente les fonctions que doit remplir le synthétiseur. Le modèle le plus général est représenté sur la figure :1. De cette structure on peut relever les paramètres que l'analyseur devra évaluer pour la commande du synthétiseur. Ces grandeurs sont : un indicateur de voisement ou non, la période de mélodie, le facteur de gain, les coefficients du filtre modèle et la mesure de la pente spectrale reflétant l'influence de la source d'excitation et du rayonnement. Le nombre de coefficients, p, du filtre dépend de la cadence d'échantillonnage, f, et de la longueur du conduit vocal [1].

La figure : 2 représente le schéma fonctionnel du vocoder dont la description détaillée est donnée ailleurs [1]. Les caractéristiques principales sont les suivantes :

- Calcul de la fonction de transfert : la méthode adoptée est la prédiction linéaire fondée sur l'autocorrélation et le mode d'analyse est asynchrone avec apodisation. Cette solution est bien adaptée au codage du type global de la parole et a le mérite d'être d'une mise en oeuvre très simple [1].

- Passe-bas avant convertisseur : caractéristique la plus plate possible sur toute l'étendue de la bande utile (0 à f/2) et coupure très raide hors de bande. Un filtre elliptique est bien adapté à ce cas. En effet, la méthode d'autocorrélation réalise une optimisation des coefficients, a_k, telle que le produit des spectres du signal et du filtre inverse $[1/H(z)]$ soit le plus plat possible dans la bande 0 à f/2. Par conséquent, si le filtre coupe sévèrement avant f/2, il s'ensuit que plusieurs paramètres, a_k, seront consacrés à la réalisation de l'effet inverse de le pente de coupure et donc perdus pour l'approximation de l'enveloppe spectrale du signal de parole.

- Codage : Le codage du facteur de gain et de la mélodie est fondé sur la minimisation de l'erreur relative de quantification sur la totalité de l'intervalle de variation de la grandeur (soit un codage logarithmique).

L'ensemble de paramètres retenu pour représenter la fonction de transfert doit garantir la stabilité du filtre après quantification, avoir un ordre naturel afin de permettre l'exploitation des propriétés statistiques de la distribution de chaque coefficient, et avoir un coefficient de sensibilité de la déviation spectrale par rapport à chaque paramètre qui soit constant. De l'examen des différents ensembles candidats on trouve [1] que seuls les logarithmes des rapports d'aire

réunissent ces trois propriétés.

- Préaccentuation : L'objet de ce prétraitement est d'égaliser l'effet de la source d'excitation ce qui permet de réduire le nombre de coefficients nécessaire à la représentation de la fonction de transfert globale de l'appareil vocal. Cependant, il a été constaté [1] que la préaccentuation redresse la pente du spectre à court terme et a pour conséquence la diminution de la dispersion des valeurs Log B_i. L'analyse avec deux préaccentuations l'une fixe et l'autre conditionnelle permet de réduire de 17% la somme totale des intervalles de variations des Log B_i par rapport à une préaccentuation fixe unique [1].

- Filtre de synthèse : La structure la mieux adaptée est celle d'un filtre en échelle à un ou deux multiplieurs. En effet, on évite ainsi les calculs récurrents pour calculer les a_i à partir des k_i et on peut lisser les k_i sans compromettre la stabilité du filtre. Dans le cas où la précision des calculs est grande le choix se portera sur le modèle à un multiplieur. Cependant, ce modèle présente des inconvénients dans le cas de calcul avec précision réduite :

a) Le facteur de gain est fortement dépendant des $\{k_i\}$ et dans l'hypothèse d'une préaccentuation d'ordre un ou zéro, ce facteur varie dans des proportions considérables et dans un laps de temps très court. Ces variations peuvent engendrer un léger bruit dans la parole de synthèse. Avec deux zéros de préaccentuation les $\{k_i\}$ sont moins sélectifs et les valeurs de gain sont nettement moins dispersées [1], mais les variations brutales subsistent.

b) Afin que la précision des calculs du filtre soit constante indépendemment du gain, il faut que la prise en compte de ce facteur n'intervienne qu'après le filtrage. Or dans ce cas les risques de débordement des calculs du filtre sont considérables.

Au regard de ces inconvénients le choix s'est porté sur la structure en échelle dite à deux multiplieurs.

Les études de simulation ont montré que l'on peut atteindre avec ces vocoders des débits de 2 Kbit/s sans dégradation sensible de la qualité.

IV - LE SYNTHETISEUR

La réalisation du synthétiseur a été envisagée en vue de son raccordement à un organe de synthèse vocale dont le cahier des charges est fixe. Plus précisément :

	Débit en ligne	=	4800 bits/s
	Fréquence d'adaptation	=	75 Hz
	Nombre de bits pour la commande	=	64 bits
dont	Mélodie	=	8 bits
	Fonction de transfert	=	48 bits
	Indications prosodiques	=	8 bits

Les contraintes de réalisation sont le temps-réel et la simplicité des opérations à effectuer (s'écarter des conditions optimales). Le prototype doit se prêter facilement à des modifications.

Dans la simulation du système dans les conditions réelles de fonctionnement avec une cadence de 8 KHz, l'attention s'est portée sur les différents aspects d'ordre pratique que présente une réalisation matérielle :

A) Format des paramètres de commande : La période de mélodie a une définition imposée de 8 bits. Il reste 48 bits à répartir parmi les différents paramètres de la fonction de transfert. Afin de simplifier les opérations de prélèvement et les calculs nous avons retenu un nombre fixe de bits pour les 8 coefficients de réflexion du filtre de synthèse, à savoir 8 x 5 bits, avec 7 bits pour le facteur de gain et un bit pour la désaccentuation. Il faut noter que dans l'étude précédente nous avions montré que le codage de la fonction de transfert est optimale sur le log. des rapports d'aires. Cependant, dans un souci de simplification du synthétiseur nous avons retenu un codage linéaire des coefficients de réflexion. Ce choix n'est possible que dans le cas d'une préaccentuation à deux zéros, dans quel cas les valeurs des coefficients de réflexion dépassent rarement 0.75. Or nous savons que la variation de ce coefficient avec le log. du rapport d'aire est presque linéaire de 0 à 0.7. Ainsi la dégradation qui en découle n'est pas très sensible. Les distributions des valeurs des k_i sont représentées sur la figure 3.

B) Précision des calculs : Des études statistiques [2] ont été entreprises pour évaluer la précision nécessaire aux calculs intervenant dans le filtre de synthèse dans différents cas d'application. Il en ressort qu'en général l'erreur quadratique moyenne, EMQ, entre les spectres du signal de synthèse obtenu avec des calculs précis et celui obtenu avec des calculs en virgule fixe portant sur Nbits décroît très rapidement avec N. Les résultats indiquent, par ailleurs, que, pour atteindre un taux donné de EMQ, avec préaccentuation au niveau de l'analyse il faut 3 bits de moins qu'en l'absence de préaccentuation. Ces données statistiques montrent que dans notre cadre d'application il suffit de N=16 bits en virgule fixe pour que l'EMQ soit inférieur à 0,02 en dB^2 pour des fréquences d'échantillonnage inférieure à 12 KHz.

C) Description fonctionnelle

Le schéma fonctionnel du synthétiseur est présenté sur la figure 4. On notera qu'à chaque instant deux ensembles de paramètres de commande sont disponibles au synthétiseur. Ces données de commandes sont les résultats de deux analyses consécutives du signal de parole décalées dans le temps de 13,33ms. Le contenu de ces mémoires est, par conséquent, raffraichi tous les 13,33ms. On qualifie les données de la mémoire à droite d'antérieures et celles de la mémoire à gauche d'actuelles. Afin de réduire le bruit de transition entre les valeurs consécutives d'un même paramètre on effectue une interpolation au temps intermédiaire soit après 6,667ms du dernier raffraîchissement du contenu des mémoires. Dans ce cas, il s'agit de réaliser une opération de moyenne (somme avec décalage à droite d'une position binaire). Ce lissage est appliqué à tous les paramètres, Cependant, il faut

tenir compte des cas particuliers suivants :

- Mélodie : lors de la transition voisée-non voisée, la période de mélodie est maintenue. Dans le cas d'une transition inverse le signal synthétisé reste non-voisé pendant les 13,33ms.

- Désaccentuation : au temps intermédiaire le pôle de désaccentuation conditionnelle prend la valeur 0.625 lors d'une transition dans le mode de désaccentuation.

Le décodage du gain est effectué par inversion de la loi de codage qui est une approximation d'une loi logarithmique (type MIC) avec trois bits poids forts représentant la caractéristique et les 4 bits poids faibles donnant la mantisse.Le calcul du gain appliqué à la sortie du filtre tient compte de la valeur décodée $\hat{g}$ et de l'énergie (fixe) de l'excitation, donc également du type d'excitation. La valeur ajustée du gain devient dans ce cas :

- sons sonores : $G = \hat{g} * \sqrt{M}$)
  où M = période de mélodie) $G = G/2$
- sons sourds : $G = \hat{g} \times 3.45197 \times 4$)

La cadence de fonctionnement du synthétiseur est dans ce cas de 125 µs.

Il existe deux modes d'excitation :

- sons sonores : générateur d'impulsions. Cependant afin de maintenir un signal d'excitation de moyenne nulle, la séquence d'excitation comprend un échantillon de valeur 4 096 et (N-1) échantillons de $\{-4096/(N-1)\}$ ou N = nombre d'échantillons dans une période de mélodie. Cette séquence étant répétée autant de fois que nécessaire. Le choix de 4 096 représente un compromis entre la précision souhaitée des calculs et le souci d'éviter le débordement des calculs.

- sons sourds : générateur de nombres pseudo-aléatoires. Ce générateur est réalisé à l'aide d'un registre à décalage de 16 bits avec une boucle d'introduction d'un bit en poids faibles. La valeur obtenue est ramenée dans la fourchette ± 512, par décalage de 6 positions binaires avec propagation du bit de poids fort.

L'inhibition de l'excitation se produit dans le cas d'un gain nul.

Le filtre de synthèse est représenté sur la figure 5. Les additions et soustractions sont effectuées sur des mots de 16 bits. Les multiplications portent sur des mots de 6 bits x 16 bits, où les 6 bits sont utilisés pour représenter les coefficients de réflexion lissés.

Le facteur de gain est introduit par multiplication avec la sortie du filtre soit 12 x 16 bits.

La désaccentuation consiste en l'affaiblissement des hautes fréquences. La figure 6 illustre les opérations prises en compte. Les deux mémoires ont chacune 16 bits et les coefficients p et q sont de 6 bits. Le pôle p représente la désaccentuation conditionnelle et peut prendre les valeurs {0,0.625,0.875}. Le pôle q représente une désaccentuation fixe et prend la valeur 0.9375.

V - REALISATION DU PROTOTYPE

L'étude de la réalisation matérielle du synthétiseur est destinée à étudier la faisabilité d'un tel système de synthèse de la parole et de sa mise au point. Trois principes en ont dicté la réalisation :

- Facilité de mise en oeuvre du système ;
- Adaptabilité à différents types de calculateur ;
- Modification aisée de la périodicité des calculs du nombre de cellules du filtre et du format des données.

Il doit pouvoir effectuer les opérations suivantes :

- Multiplication 16 x 16 en complément à 2 en moins de 3 µs
- Division 13/8 en binaire
- Extraction d'une racine carrée
- Interpolations et lissages toutes les 6,6ms.

Les trois principes adoptés précédemment nous obligent à utiliser un processeur microprogrammé (pour la souplesse d'emploi) rapide (à cause du filtre de synthèse). Le choix s'est donc porté sur un système à 2 niveaux organisé autour d'un microprocesseur.

Le microprocesseur effectuera :

- La gestion de la liaison avec le calculateur
- Le lissage des données
- Le programme de tests du synthétiseur.

Un séquenceur rapide microprogrammé effectuera les calculs de la source et du filtre de synthèse.

La division et l'extraction de la racine carrée seront faites par traduction.

VI - REFERENCES

[1] I.I. EL MALLAWANY : "Etude de vocoders à prédiction linéaire..." Thèse de Docteur Ingénieur, Grenoble (19 septembre 1975).

[2] J.D. MARKEL et A.H. GRAY : "Fixed point truncation arithmetic implementation of a linear prediction autocorrelation vocoder". IEEE Trans., Vol. ASSP-22,No4, (August 1974) pp. 273-282.

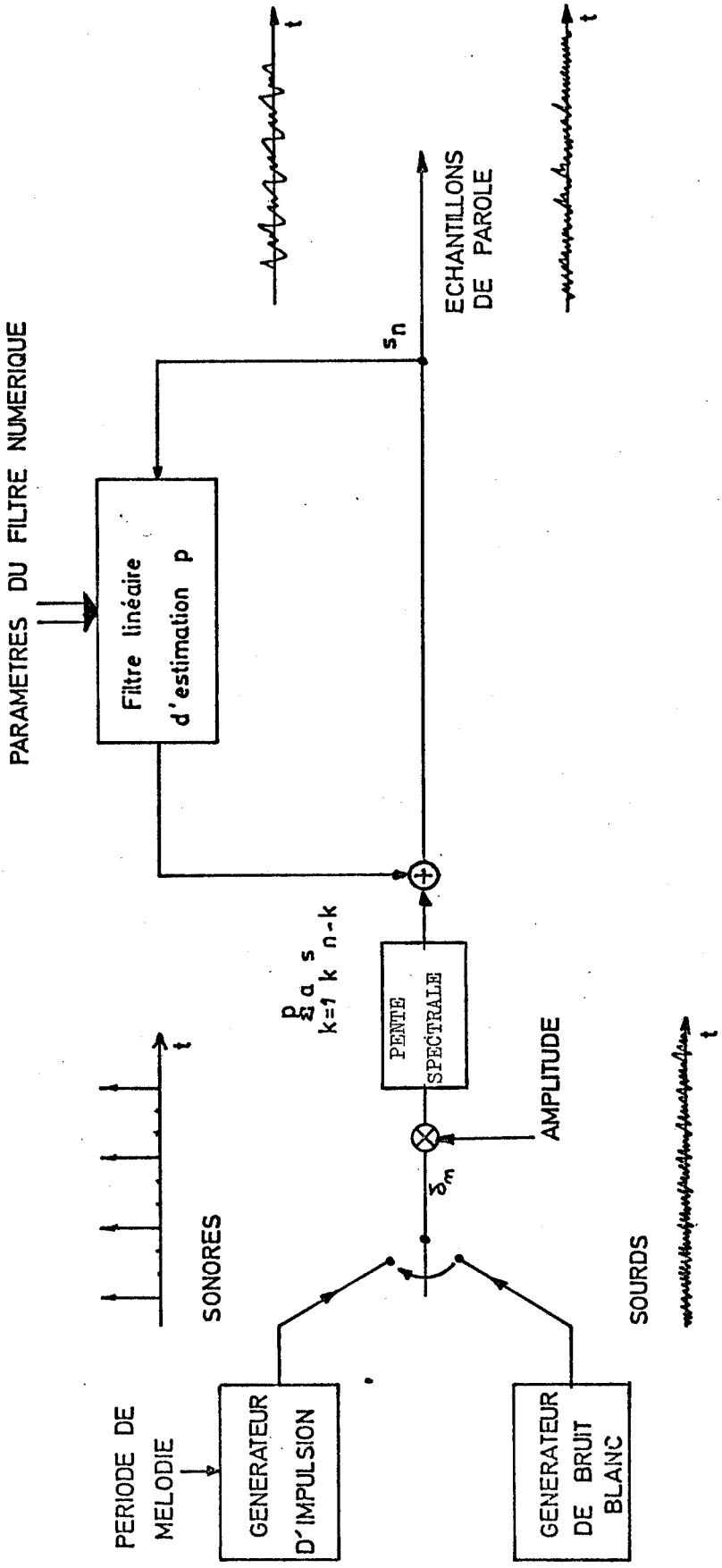


Figure 1 MODELE NUMERIQUE DE PRODUCTION DE LA PAROLE

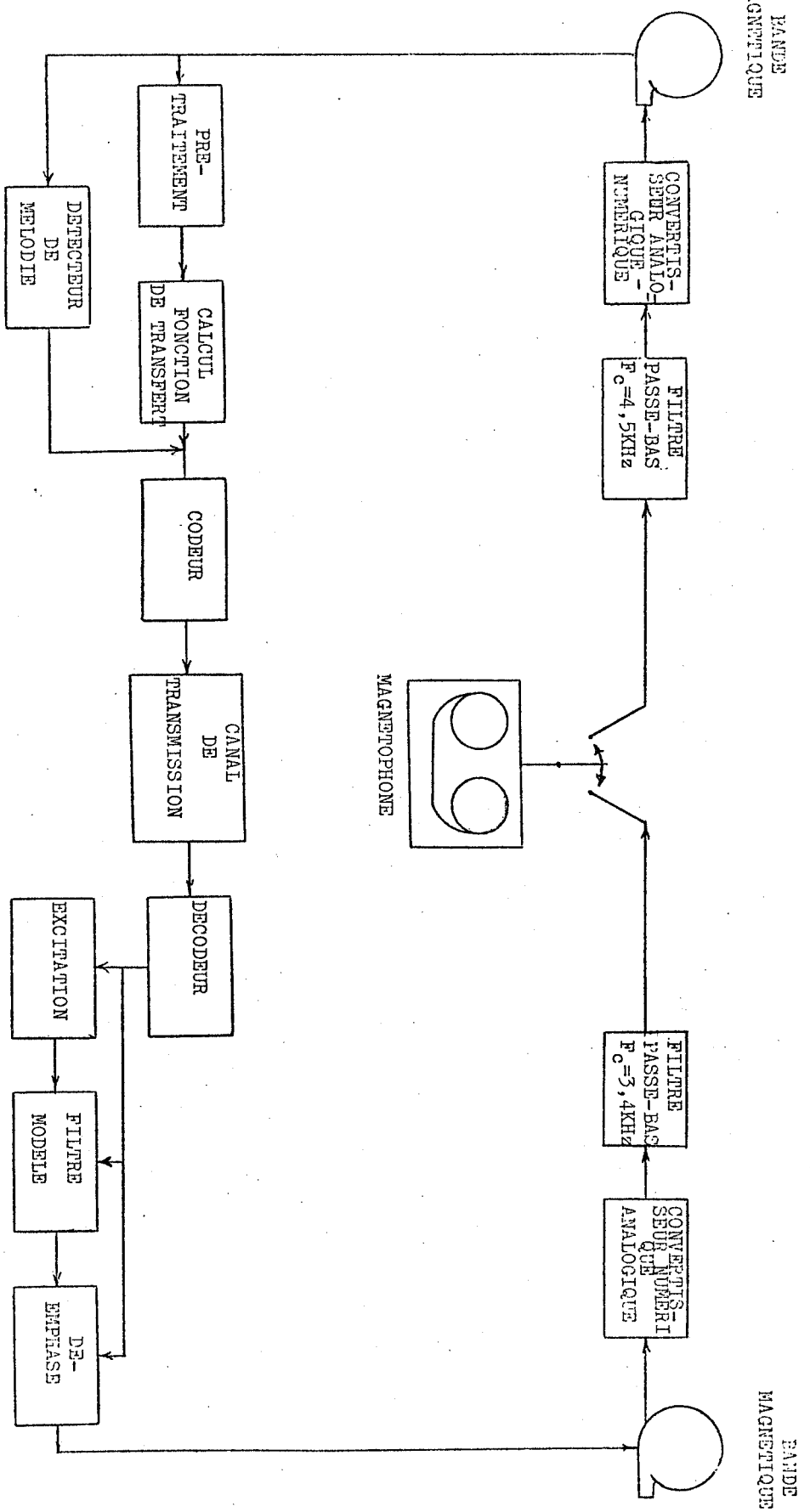
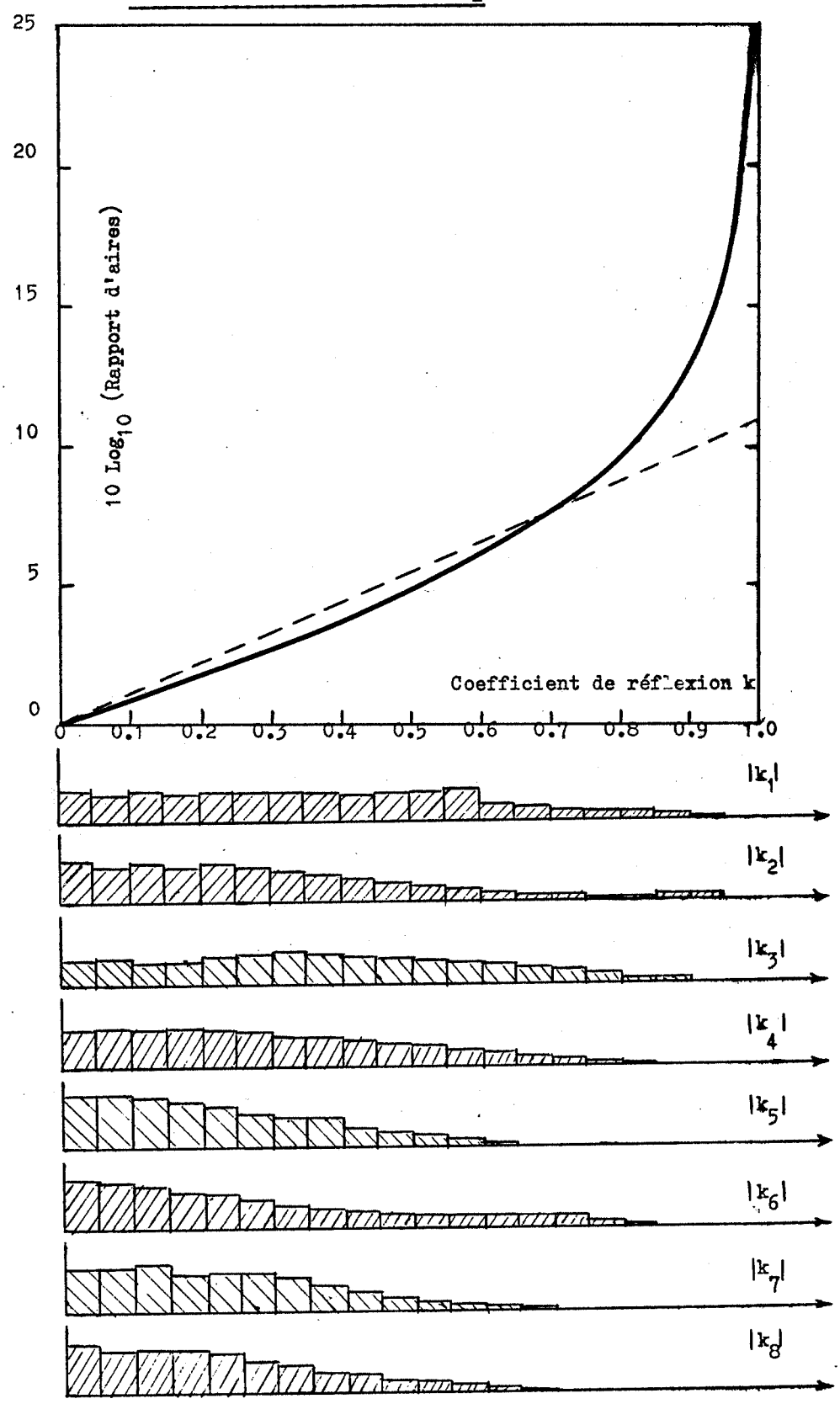


Figure : .2 SIMULATION D'UN VOCODER A PREDICTION LINEAIRE

Figure 3 - Histogrammes des $|k_i|$



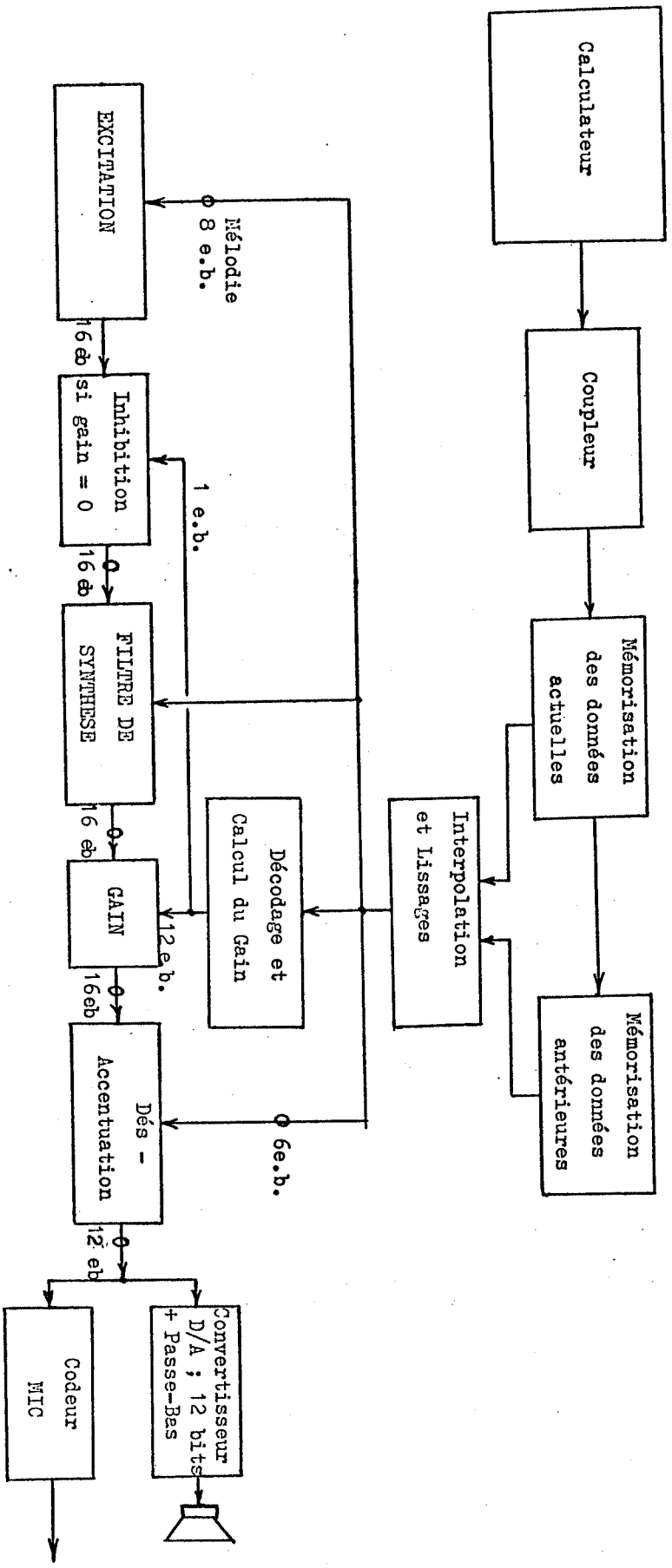


Figure : 4 - Schéma Fonctionnel du Synthétiseur.

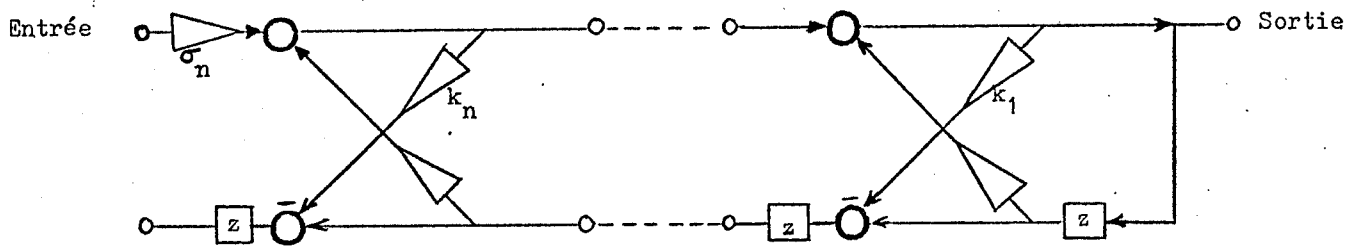


Figure : 5 - Filtre de Synthèse (Structure en échelle à 2 multiplieurs)

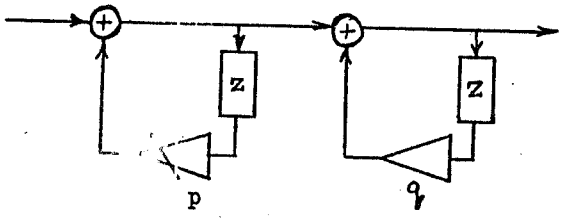


Figure : 6 - Désaccentuation

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 au 21 mai 1976

Synthèse par diphones et Traitement automatique de la prosodie

Françoise EMERARD et Danielle LARREUR
Centre National d'Etudes des Télécommunications
22301 LANNION

RESUME : Ce traitement automatique de la prosodie utilise un nombre limité de règles simples indépendantes de la complexité de la structuration grammaticale.

L'application des schémas mélodiques dépend de la seule position des mots dans la phrase et ne cécessite que la notation d'un certain nombre de points sur la chaine phonétique.

SUMMARY : A simple set of prosodic rules, which are not taking the complexity of syntactic sentence structuration into account is described.

The application of this automatic and real time intonation processing requires only a limited number of marks throughout the phonetic string, in order to know boundaries and position of words inside each pattern of sentences.

Synthèse par diphones et Traitement automatique de la prosodie

---

Françoise EMERARD et Danielle LARREUR
Centre National d'Etudes des Télécommunications
22301 LANNION

I. INTRODUCTION

La langue que l'on se propose de produire dispose de 33 réalisations phonémiques (plus le silence noté #) ; Il faut donc, déduction faite des juxtapositions de phonèmes non réalisables de par la nature physique des organes phonatoires, stocker en mémoire à peine plus de 1000 diphones représentant toutes les possibilités, pour avoir les moyens de composer n'importe quel message.

La constitution de la bibliothèque est réalisée par l'intermédiaire d'un vocodeur à canaux qui analyse actuellement les données numériques relatives aux diphones à 4 800 e.b./secondes.

Chacun des 14 filtres non recouvrants recueille pour sa bande de fréquence une intensité codée logarithmiquement à l'aide de 16 niveaux. L'enveloppe spectrale est échantillonnée toutes les 13,3 ms, ainsi que la période du fondamental (pitch) quantifiée sur 256 niveaux pour les sons voisés (pour les sons non voisés cette période alors inexistante est codée Ø).

Ce qui représente pour chaque échantillon vocodeur :

- concernant le spectre :
  14 canaux x 4 e.b./canal = 56 e.b.
- concernant le pitch = 8 e.b.

soit au total (56 + 8) 75 échantillons/s = 4800 e.b./s.

Le synthétiseur à canaux parallèle à l'analyseur offre le double avantage d'un stockage économique des données et d'une commande aisée du rythme et de l'intonation sur la parole obtenue par simple assemblage des diphones stockés dans la mémoire du calculateur.

.../...

Cependant si la simple juxtaposition de diphones permet d'obtenir une parole intelligible, celle-ci demeure tout à fait dépourvue de naturel : en effet, l'ensemble du message est émis sur le même ton, et aucun élément ne facilite l'opération prioritaire de démarcation de la chaîne parlée en mots ou en groupes de mots et de sens, nécessaires pour réaliser le décodage du message et aider la compréhension de l'auditeur.

Difficultés de méthode

Le problème de la taille mémoire oblige à ne prendre en considération qu'une seule configuration spectrale et temporelle pour chaque diphone. Or il est évident que dans la parole naturelle, chaque unité minimale a une durée et un spectre fonction de son environnement phonétique, et fonction de sa position dans le message.

La durée relevée pour chaque diphone est dans la plupart des occurences excessive par rapport à celle que l'on constate dans la parole continue, les réalisations consenantiques étant le plus sensible à cet allongement. Olive et Nakatani précisent la nécessité de réduire les mots prononcés de façon isolée de 20 % de leur durée quand ils sont utilisés dans la parole naturelle /1/.

Le corpus des mots dont sont issus les segments, ainsi que les procédures d'enregistrement doivent être définis de façon à éliminer au mieux des niveaux et des variations d'intensité dont les différences ne seraient pas dues aux seules caractéristiques des diphones.

Les mots qui ont servi à l'enregistrement des données ont été prononcés par le locuteur sur un ton volontairement monocorde - donc obligatoirement mais inconsciemment avec un timbre plus grave (ce phénomène a été constaté systématiquement avec notre locuteur et est assez facilement généralisable)- pour éviter des variations d'intonation importantes, c'est-à-dire la possibilité de discontinuités de la fréquence fondamentale aux frontières de diphones lors de l'assemblage. Il est par conséquent évident que l'on ne peut conserver des informations fournies sur celle-ci que ce qui revèle les caractéristiques intrinsèques des unités phonétiques, et que l'on doit superposer entièrement le message mélodique en le recomposant.

II. LES TRAITEMENTS

A. Traitement du rythme et de l'intensité

1. Procédure d'élaboration du dictionnaire

Afin d'obtenir une parole de synthèse au débit moins heurté, plus fluide, une nouvelle bibliothèque a été réalisée à partir non plus de mots artificiels (titi, fafa, koukou ...) mais de mots pour la plupart usuels et dont les segments voyelle-consonne ont été extraits de syllabes fermées. Cette procédure permet de diminuer sensiblement la sensation auditive d'un découpage syllabique du flux de la parole due vraisemblablement au fait que les diphones (voyelle-consonne) étaient précédemment extraits de réalisations syllabiques ouvertes, ce qui augmentait la durée des transitions de la voyelle à la consonne.

Contraints -nous l'avons dit- pour une question de taille de bibliothèque à ne conserver qu'une seule configuration pour chaque diphone, nous avons établi pour chacun d'eux, une "durée moyenne" convenable et unique qui tient compte des caractéristiques intrinsèques propres aux deux éléments de chaque segment :

- La variation de durée en fonction de la nature des consonnes se manifeste dans les diphones [consonne/voyelle] : en effet, à durée de voyelle identique, les segments [consonne/voyelle] dont l'élément consonantique est sourd, ont une durée supérieure à ceux dont la consonne est voisée.
- La variation de durée selon les voyelles est fixée dans les diphones [voyelle/consonne] en fonction de la consonne subséquente : on sait par exemple que la voyelle qui précède une consonne voisée est plus longue que la même moyelle précédant une consonne sourde.

On peut cependant signaler que la durée moyenne d'une voyelle orale a été fixée aux alentours de 105 ms, et que la durée d'un diphone avoisine 120 ms. Quant aux réalisations les plus longues, soit les segments [voyelle / voyelle] , par exemple [e-ɥ] dans [réinvestir] , elles-ne peuvent -pour des raisons de programmation- excéder 220 ms.

L'analyse d'un corpus enregistré par cinq locuteurs a permis d'établir une différence importante :

- entre l'allongement d'une syllabe finale par rapport à celui d'une syllabe de fin de syntagme,

- entre l'intensité recueillie dans la dernière syllabe de la phrase par rapport à celle que l'on relève dans la dernière syllabe des mots de fin de syntagme non finaux de phrase (Fig. 1).

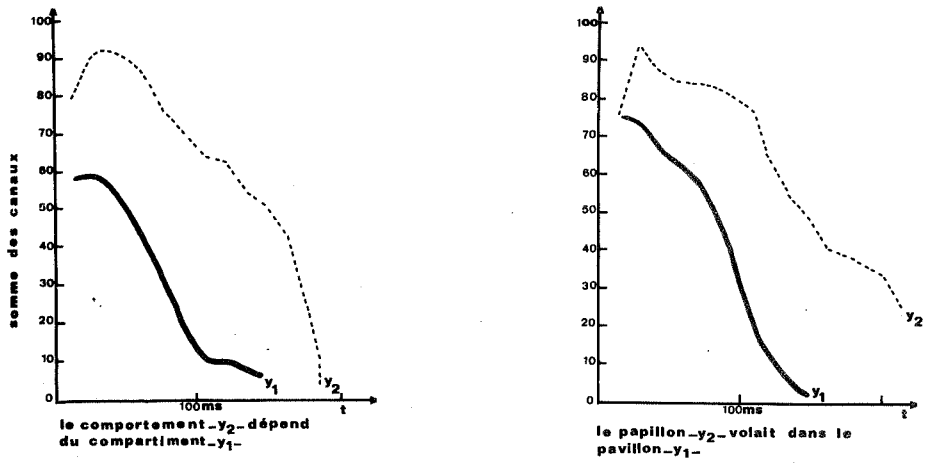


Fig. 1 : Configuration de la voyelle finale en fin de groupe (y_2) et en fin de phrase (y_1)

Pour respecter ce phénomène, tout en préservant l'objectif d'un système de synthèse en temps réel, nous avons choisi d'introduire dans le dictionnaire une double configuration spectrale et temporelle pour les diphones [consonne/silence] :

- Une configuration -x 1- qui provient du dernier segment des mots prononcés avec l'intonation descendante propre aux fins de phrase énonciative.

- Une configuration -x 2- propre aux fins de mots qui caractérisent les groupes de continuité (groupes de continuité majeure pour Delattre /2/, groupes de continuité en général pour Léon et Léon /3/).

Mais on stocke seulement une configuration - y2- spécifique aux fins de mots des groupes de continuité en ce qui concerne les diphones [voyelle-silence] (Voir cas 2a).

2. Traitement au niveau de la synthèse

a. Traitement du rythme et de l'intensité dans le cadre des mots

Parce que les diphonèmes sont extraits de mots isolés bien articulés, il est nécessaire dans la parole continue d'accélérer la fréquence d'échantillonnage en fonction de la longueur du mot. Ce problème pour l'instant n'a pas été résolu.

En ce qui concerne le problème particulier des fins de mots des "groupes de continuité" et de phrase, quatre cas doivent être envisagés :

1. Fin de groupe de continuité,
   essentiellement . fin de syntagme nominal sujet
   . fin de syntagme verbal

   se terminant par
   - 1a) Syllabe ouverte (configuration voyelle X2)
   - 1b) Syllabe fermée (configuration consonne X2, ə X2)

2. Fin de phrase

   se terminant par
   - 2a) syllabe ouverte (configuration voyelle X2)
   - 2b) syllabe fermée (configuration consonne X1, ə X1)

Résolution du traitement

- cas_1a/ : Il ne nécessite aucun traitement particulier, les diphones utiles étant tous stockés en mémoire.

- cas_1b/ : Pour respecter l'allongement final de dernière syllabe, on double tous les échantillons de la voyelle, mais on laisse intacte la configuration X2 de [consonne#] ou de [ə#] . Aucun traitement n'est effectué sur l'intensité puisque l'on constate pour les voyelles une faible différence entre l'image d'évolution spectrale_Y2_ obtenue à l'analyse, et l'image obtenue par simple doublement des échantillons de la voyelle.

- cas_2a/ : On est obligé d'utiliser pour les diphones [voyelle#] , la configuration_Y2_des fins de groupes de continuité car elle seule s'adapte sans grande discontinuité au diphone [consonne/voyelle] précédent. Par conséquent on utilise un double marquage inscrit sur tous les diphones [consonne/voyelle] : d'une part au début de la consonne et d'autre part au début de l'élément vocalique, et on diminue l'intensité de 1/2 dB pour chaque échantillon de la consonne puis de 1 dB pendant toute la durée de la voyelle finale.

- cas_2b/ : On double tous les échantillons de la dernière voyelle, on effectue le même traitement d'intensité qu'en /2a/, et on laisse intact le dernier élément de la syllabe (consonne X1 ou /ə/1).

Ex. /cable/ → /#k/ka/ab/bl/lə/ə #/
1 2 3

1 = - 1/2 dB
2 = doublement des échantillons, - 1 dB
3 = configuration X1

b. découpage temporel de l'énoncé

On tient compte du fait que la parole n'est absolument pas un phénomène continu, que 40 à 50 % du temps de parole est occupé par des pauses /4/.

Afin de faire face à une double nécessité :

- assurer au message un flux de parole qui tienne compte des nécessités physiologiques de respiration phonatoire.

- effectuer un découpage de la phrase qui reflète et signale l'essentiel de sa structure afin de renforcer la cohésion de sens entre les mots d'un groupe /5-6/. On a prévu des marques graphiques (qui ont en même temps une fonction d'indicateurs de schémas mélodiques) :

- Celles ci imposent l'insertion de pauses de différentes longueurs :

  . D'une part des pauses en fin de syntagme nominal et verbal, ainsi qu'à la fin de groupes de sens qui ne correspondent pas en même temps à la fin d'un syntagme. La durée de ces pauses est variable (comprise entre 60 et 800ms) selon le nombre de mots et la complexité des groupes qu'elles concluent.

  . D'autre part, des pauses plus longues associées aux marques usuelles de délimitation (points, points virgules, points de suspension)

- Elles assurent l'indépendance de la démarcation des mots situés en frontière de syntagme ou de groupe, car elles permettent d'éliminer le diphone de transition entre deux unités. Par exemple, dans la phrase "le petit chat a mangé des souris", l'élément /aa/ est supprimé et les deux syllabes frontières de syntagme sont entièrement prononcées (... ʃa/a # / # a/am/mɑ̃/...) comme s'il s'agissait de syllabes de fin et de début de phrase.

B. Traitement de la fréquence fondamentale

Le procédé d'enregistrement et d'obtention des diphones permet de connaître toutes les variations de Fo relatives aux caractéristiques intrinsèques des phonèmes -et plus particulièrement des consonnes voisées- variations micromélodiques indépendantes de la volonté du locuteur, et dont l'influence sur l'intelligibilité a été étudiée par ailleurs /7/.

1. Traitement au niveau intrinsèque

a. Nous laissons en mémoire toutes les informations relatives à l'enveloppe spectrale des diphones, mais en ce qui concerne la fréquence fondamentale des consonnes voisées (les articulations consonantiques sont beaucoup plus évolutives que les réalisations vocaliques) nous gardons, non pas les valeurs absolues telles qu'elles sont délivrées par le vocodeur, mais des valeurs relatives qui, par leur écarts, respectent à la fois les schémas des variations micromélodiques et la relation Fo-énergie ; de manière générale à la plus petite valeur d'énergie correspondra la valeur minimale de la fréquence fondamentale. Afin de respecter les schémas micromélodiques spécifiques aux consonnes voisées, c'est-à-dire un tracé de la fréquence fondamentale de type descendant-montant, on impose à la période fondamentale des diphones voyelle-consonne, des valeurs relatives positives, et au contraire aux segments consonne-voyelle des valeurs relatives négatives (car on constate une dégradation de l'intelligibilité si les consonnes ont un schéma mélodique supérieur à celui de la voyelle subséquente). De la même façon, les valeurs absolues correspondant aux voyelles n'ont pas été utilisées ; on les remplace par une valeur unique sur chaque échantillon et on délimite le début de la réalisation vocalique par un marquage particulier qui servira de point de repère lors de l'attribution de patrons intonatifs.

b. A partir de l'analyse d'un corpus enregistré par seize locuteurs, il s'est dégagé deux configurations différentes pour la réalisation consonantique /r/.

. /r/ entouré de deux phonèmes voisés, est lui-même voisé,
. /r/ précédé par des consonnes sourdes (occlusives et constrictives) est sourd.

Or, pour chaque diphone, une seule configuration en mémoire est envisageable. D'autre part, l'inclusion d'un élément de programme qui viendrait étudier lors du décodage du message à synthétiser l'environnement phonémique de /r/ nécessiterait un temps de calcul qui risquerait de compromettre le temps réel. Par conséquent, il s'est avéré nécessaire de faire à la suite d'un certain nombre de tests sur diphone, un choix entre les deux réalisations possibles : nous avons synthétisé un message comportant le le phonème /r/ dans tous les contextes phonémiques possibles et en avons réalisé deux versions : la première dans laquelle cette consonne était toujours non voisée, la seconde dans laquelle elle était sonore. les résultats perceptuels montrent que /r/ sourd est bien toléré dans n'importe quel environnement, sauf en finale de mots, où le trait /sourd/ de ce phonème provoque une impression "gutturale" désagréable ; alors que /r/ sonore entraîne dans certains cas (tr, fr...) une "résonance"

responsable d'une diminution d'intelligibilité sur les syllabes en question. Par conséquent, nous avons adopté une solution intermédiaire satisfaisante qui consiste à laisser en mémoire tous les diphones [consonne sourde-/r/] entièrement sourds, et les diphones [consonne sonore-/r /], [/r/ -voyelle] , et [/r/ - consonne sonore] sonores.

2. Traitement au niveau intonatif

On aborde ici l'étude de l'intonation dans sa fonction d'aide au décodage du message, c'est-à-dire linguistique. Ce traitement concerne essentiellement les voyelles, car les consonnes n'interviennent pas dans la perception du message mélodique.

On a fait enregistrer par un locuteur puis analysé au vocodeur un important corpus de phrases de complexités diverses : depuis des phrases courtes et simples (sujet + verbe + complément) jusqu'aux phrases comportant de multiples expansions dans chacun de ces trois éléments.

- a. A la suite de tests comparatifs sommaires sur l'intelligibilité et l'agrément de deux "voix synthétiques", notre choix s'est délibérément porté sur un locuteur féminin.

  Cette décision n'est pas véritablement surprenante car déjà les différents tests réalisés sur des listes de logatomes ont montré que les pourcentages d'intelligibilité sont supérieurs en moyenne de 4 % avec les voix de femme .

  Grâce à l'analyse , on a pu définir la tessiture du locuteur qui a servi à l'enregistrement des données, en l'espèce on peur situer les limites de son registre fréquentiel entre 136 et 380 Hz, soit une étendue d'environ une octave et demie. La valeur de la fréquence laryngienne moyenne se situe vers 195 Hz, le niveau maximal est atteint pendant la dernière syllabe du mot portant l'inflexion interrogative -ce peut être ou bien la dernière syllabe de phrase ou bien la dernière syllabe de mot interrogatif-, et la valeur minimale pendant la dernière syllabe des phrases énonciatives /8/.

- b. On fixe dans le message un certain nombre de points importants qui sont marqués différemment et qui correspondent à certaines décisions au niveau de l'attribution des schémas mélodiques. Le décodage du message s'effectue à six niveaux successifs.

  1. Reconnaissance du type de phrase

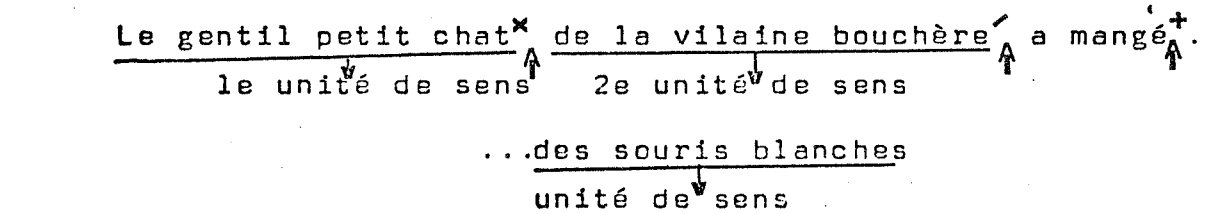
     . phrase énonciative (marque de type /./)
     . phrase impérative (marque de type /!/)
     . phrase interrogative (marque de type /?/). Ce type de phrases pose des problèmes particuliers car on y trouve une grande variété de structures syntaxiques différentes.

2.Détermination des dernières syllabes de syntagme

Dans chacun de ces trois types de phrase, on situe la fin du syntagme nominal sujet et des possibles groupes de sens qui le composent, la fin de syntagme verbal, et la fin du syntagme complément qui correspond également à la fin de phrase. On donne à ces trois points des significations mélodiques différentes selon le type de phrase auquel ils appartiennent.

3.Si la composition des syntagmes est complexe, on délimite à l'intérieur de chacun d'eux les frontières des différentes expansions qu'il comporte (jonctures externes). Cela signifie que l'on intervient pas à l'intérieur des groupes dont les éléments possèdent une grande cohésion et que la délimitation ne se situe qu'au moment où cette cohésion, cette unité de sens, prend fin.

Exemple :



×Fin d'un groupe de sens dans le syntagme nominal sujet
/Fin du syntagme nominal sujet
+Fin du syntagme verbal
·Fin de phrase énonciative

Un schéma descendant est attribué en fin de groupe de sens du syntagme nominal sujet, et un schéma montant en fin de groupe de sens du syntagme complément.

4. Détermination des jonctures internes

Il faut définir où se situent les frontières syllabiques entre deux mots pour savoir à quel moment appliquer le schéma spécifique (schéma descendant) au mot non marqué, c'est-à-dire à un mot associé très directement et très profondément au mot qui le suit : soit dans l' l'exemple ci-dessus entre /gentil/ et /petit/, entre /petit/ et /chat/

5.Décomposition des mots en syllabes

Les schémas mélodiques de mots dépendent directement du nombre de syllabes qu'ils comportent, la position que le mot occupe dans la phrase intervenant surtout au niveau de l'attribution du schéma de sa dernière syllabe (ou de ses deux dernières syllabes dans le dernier mot de phrase).

6. En fonction du nombre de syllabes, on opère une distinction selon les mots : tous les mots non marqués qui ne possèdent qu'une syllabe sont considérés comme des proclitiques, ou mots-outils (articles, relatifs, possessifs...) porteurs d'une faible information sémantique. Et tous les mots supérieurs ou égaux à deux syllabes sont envisagés comme non -clitiques (adjectifs, substantifs...).

Cette distinction est arbitraire mais légitime, compte tendu du choix méthodologique que nous avons fait de ne pas recourir à une analyse syntaxique : on peut en effet observer que tous les mots -en majorité mono-syllabiques- de faible contenu sémantique, possèdent en général des valeurs de fondamental plus basses que les autres, une énergie plus faible d'environ 23 % (résultat calculé sur la somme des canaux) et une durée plus courte ; cette réalité justifie l'application pour tous les segments monosyllabiques d'un schéma mélodique unique et simplifié, quelle que soit leur localisation dans la phrase.

Quant aux non -clitiques (lexèmes nominaux), on leur attribue un schéma mélodique fonction de leur longueur et des valeurs de fréquence fondamentale fonction de leur position dans le message. Cependant, on relève, si l'on ne prend pas en considération le schéma de dernière syllabe,

- certaines constantes dans les schémas de mots, par exemple un schéma descendant sur toutes les voyelles d'avant dernière syllabe des mots,
- et un phénomène particulier, non généralisable, d'évolution de la fréquence fondamentale, observé sur plusieurs locuteurs :

  . Schéma montant sur la voyelle de syllabe initiale ; si cette voyelle est précédée d'une consonne sourde, son schéma micromélodique est d'abord descendant pendant environ 30 ms avant d'être montant.

  . La voyelle de deuxième syllabe (schéma descendant) a pour point de départ une valeur fréquentielle supérieure à la valeur d'arrivée de la voyelle de première syllabe. L'apparition de ce double phénomène n'est pas systématique et semble ne se produire que pour manifester une insistance sémantique particulière sur le mot. C'est en tous cas une impression auditive d'accentuation qui est perçue lors de cette réalisation.

  . Toutes ces voyelles, sauf les voyelles finales de syntagme et de phrase connaissent une intensité maximale au centre de leurs réalisations.

Ainsi, pour chaque mot, il suffit de fixer un nombre de valeurs absolues ainsi qu'une pente algébrique qui correspondent au nombre de syllabes décomptées : alors, connaissant les valeurs et la forme du schéma que l'on veut voir attribué à chaque voyelle, il est très aisé, par soustraction des valeurs relatives, de connaître la valeur qu'il faut imposer au premier échantillon de la consonne dans les diphones consonne-voyelle pour parvenir sans dis-

continuité de fréquence fondamentale jusqu'à la fin du diphone voyelle consonne subséquent. On comprend donc que la rupture de Fo, si elle existe, ne peut se produire qu'au centre d'une réalisation consonantique à un moment où justement une concavité positive se forme dans la configuration des consonnes voisées, et est rendue impossible à tout autre moment où elle aurait des conséquences perceptuelles néfastes. La succession de ces niveaux d'analyse permet de produire automatiquement les patrons mélodiques des messages quelle que soit la complexité de leur structuration.

Les règles énoncées ci-dessus sont simples, précises, et applicables avec une facilité qui permet au traitement de la prosodie d'être effectué en temps réel ; en effet, une fois connu le type de phrase à traiter, il suffit de suivre logiquement le message phonétique de gauche à droite et d'appliquer le schéma mélodique qui correspond à un marquage particulier ; la position du mot dans la phrase se révèle plus pertinente que sa fonction pour l'attribution de ces schémas /9/.

Il reste cependant que ce traitement de la prosodie peut et doit être affiné, et que nous envisageons dans une phase ultérieure de dépasser le cadre strictement linguistique de la prosodie pour étudier un traitement relatif à la fonction expressive (phonostylistique) de l'intonation. En ce sens, l'effet des émotions semble donner priorité à la modification du timbre de la voix, à la répartition des pauses, au rôle de la durée.

BIBLIOGRAPHIE

/1/ OLIVE J.P., et NAKATANI, L.H., Rule Synthesis of Speech by word concatenation, a first step. J.A.S.A. - Vol. 55 n° 3, pp. 660-666, 1974.

/2/ DELATTRE, P., (1966) Les dix intonations de base du français, French Review, 40 (1).

/3/ LEON, P.R., et LEON, M., (1964) (2è éd. 1966) Introduction à la phonétique corrective, Hachette, Paris.

/4/ GOLDMAN-EISLER, F., Psycholinguistics : Experiments in spontaneous speech, New York : Academic (1968)

BOE, L.J., CONTINI, M. RAKOTOFIRINGA, H., in Etude statistique de la fréquence laryngienne, Phonetica 32 : 1 - 23 (1975)

/5/ LIEBERMAN, P. intonation, and Language. Cambridge : MIT Press (1967)

/6/ RUDER, K.F., and JENSEN, P.J., Fluent and hesitation pauses as a function of syntactic complexity Journal of Speech and Hearing, vol. 15 n° 1, March 1972 pp. 49-60.

/7/ BOE, L.J., et LARREUR, D, Etude de l'influence des variations de la fréquence laryngienne sur l'intelligibité et la qualité des consonnes sonores générées par vocodeur - in Bulletin de l'Institut de Phonétique de Grenoble, Volume II, 1973, p. 103-123.

/8/ LARREUR D., et BOE, L.J. Synthèse parametrique de la phrase énonciative en français - 5ème Journées d'Etudes du Groupe Communication parlée - Vol. II; Orsay, Mai 1974.

/9/ OLIVE, J.P., Fundamental frequency rules for the synthesis of simple declarative English sentences, J.A.S.A., Vol. 53 n° 2, p. 476-482, 1975.

==========

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

COMPORTEMENT D'UN MODELE DE LA SOURCE VOCALE CHARGE PAR L'IMPEDANCE D'ENTREE DU CONDUIT VOCAL

B. GUERIN, M. DELOS, M. MRAYATI
Laboratoire de la Communication Parlée
E.N.S.E.R.G. 23 Rue des Martyrs
38031 GRENOBLE CEDEX

RESUME :

On décrit un modèle de l'impédance d'entrée du conduit vocal constitué de deux circuits résonnants R.L.C. accordés sur les fréquences des deux premiers formants.

Une simulation de la source vocale chargée par ce modèle de l'impédance d'entrée permet d'étudier d'une part l'influence de la configuration du conduit vocal sur l'onde de débit de la source et d'autre part l'évolution de la fréquence fondamentale intrinsèque.

Les variations de cette fréquence dues au couplage source-conduit ont trois causes principales. On montre que le couplage avec l'impédance d'entrée contribue pour sa part à des variations contraire de la fréquence fondamentale à celles observées dans la parole naturelle.

SUMMARY :

A model of the driving point impedance of the vocal tract is described. This model consist of two R.L.C. tuned circuits which are controlled by the first two formant frequency.

By means of a simulation of a vocal source loaded by this equivalent circuit, the interaction between the volume velocity generated by the source and the configuration of the vocal tract is studied. The relation between the intrinsic fundamental frequency and the configuration of the vocal tract is also investigated.

Three types of source - tract coupling can operate but in this case studied above, the variations of the fundamental frequency are small compared to the intrinsic pitch and opposite to the variations measured on natural speech.

COMPORTEMENT D'UN MODELE DE LA SOURCE VOCALE CHARGE PAR L'IMPEDANCE D'ENTREE DU CONDUIT VOCAL

B. GUERIN, M. DELOS, M. MRAYATI
Laboratoire de la Communication Parlée
E.N.S.E.R.G. 23 Rue des Martyrs
38031 GRENOBLE CEDEX

1 - INTRODUCTION

L'étude des caractéristiques acoustiques de l'onde de débit de la source vocale sur la parole naturelle est difficile. En effet, on ne sait pas bien dissocier, dans le signal image de la parole, les contributions dues à la source vocale de celles dues à la fonction de transfert du conduit vocal. On est alors ammener, pour étudier précisément l'onde de débit, à faire une simulation numérique de la source vocale. Mais cette source est couplée, sur le plan acoustique, au conduit vocal par l'intermédiaire de l'impédance d'entrée de celui-ci. Une simulation complète devra tenir compte de cette impédance d'entrée qui devra être connue avec précision.

Dans la suite de ce papier, nous allons décrire comment nous avons représenté l'impédance d'entrée du conduit vocal en fonction des différentes configurations de celui-ci. Puis dans une première étude, en supposant connue la variation de l'aire d'ouverture des cordes vocales, nous analyserons les influences du couplage source-conduit, au niveau temporel et au niveau spectral, de l'onde dé débit. Dans une deuxième étude, à l'aide d'un modèle complet de la source vocale, modèle à une ou deux masses chargé par l'impédance d'entrée équivalente du conduit vocal, nous calculerons les variations de la fréquence fondamentale suivant la configuration du conduit vocal considéré.

2 - MODELE DE L'IMPEDANCE D'ENTREE DU CONDUIT VOCAL

Afin d'étudier et d'incorporer certains aspects du couplage source vocale - cavités supraglottiques, une connaissance précise de l'impédance d'entrée du conduit vocal est nécessaire. Comme les mesures de cette impédance sont difficiles sur des conduits réels, nous l'avons calculer par simulation du conduit vocal sur ordinateur. En tenant compte des différentes source de perte avec précision, nous avons pu évaluer cette impédance d'entrée (MRAYATI, 1976). Les domaines de variation de cette impédance sont les suivants :

38 - 380 ohm - acoustic c.g.s. pour le 1er formant

20 - 320 " " " " " 2ème "

20 - 450 " " " " " 3ème "

Les valeurs indiquées ci-dessus sont celles de la partie réelle de l'impédance d'entrée aux fréquences de formants. On peut montrer, en effet(MRAYATI,GUERIN et BOE, 1976) que cette valeur est maximale pour ces fréquences et que la partie imaginaire s'annule. Les variations de $X(\omega)$, partie imaginaire, et $R(\omega)$, partie réelle de l'impédance d'entrée sont données figure 1.

On constate que ces courbes correspondent à celle d'un circuit de FOSTER à résonances multiples avec amortissement. Par ailleurs, l'étude des variations de cette impédance avec les différents termes de pertes dans le conduit vocal a permis d'établir des relations simples entre les valeurs des éléments du circuit de FOSTER et les fréquences des deux premiers formants (MRAYATI, 1976). On obtient finalement le circuit de la figure 2.

3 - INFLUENCE DU COUPLAGE SOURCE - CONDUIT VOCAL SUR LA FORME DE L'ONDE DE DEBIT DES CORDES VOCALES

3-1 Introduction

Nous avons d'abord étudier les perturbations apportées au niveau temporel et spectral par le couplage source - conduit. On a supposé que l'aire d'ouverture des cordes vocales suivait la même loi pour toutes les voyelles étudiées. Le circuit utilisé pour cette simulation est donné figure 3.

Sur cette figure, les éléments R_V et R_K représente les composantes cinétiques et visqueuses de la résistance glottique, L_g étant l'inertance de la glotte. Les valeurs de ces éléments sont commandées par l'aire d'ouverture A_g de la glotte, aire variable avec le temps que nous supposons connu.

L'étude théorique du comportement de ce circuit est difficile car ce circuit est non linéaire. Une simulation sur ordinateur permet de calculer le débit $U(t)$. Les effets de l'impédance d'entrée du conduit vocal et des éléments de l'impédance dynamique de la glotte seront étudiés tant sur le plan temporel que spectral.

Si nous ne prenons en compte que la plage du 1er formant, les équations de ce circuit s'écrivent :

$$P_S = U(t).(R_k+R_v) + \frac{d}{dt}(Lg.U(t)) + V$$

$$U(t) = \int \frac{V}{L_1}\, dt + C_1 \frac{dV}{dt} + \frac{V}{R_1}$$

C'est un système différentiel non linéaire à coefficients non constants. On note que L_g dépend du temps et que le terme $\frac{d}{dt}(L_g.U(t))$ donne, en le développant, deux termes :

$$L_g \frac{dU(t)}{dt} \quad \text{et} \quad U(t) \frac{dL_g}{dt}$$

Le système d'équation sera résolu en utilisant la méthode de Runge KUTTA.

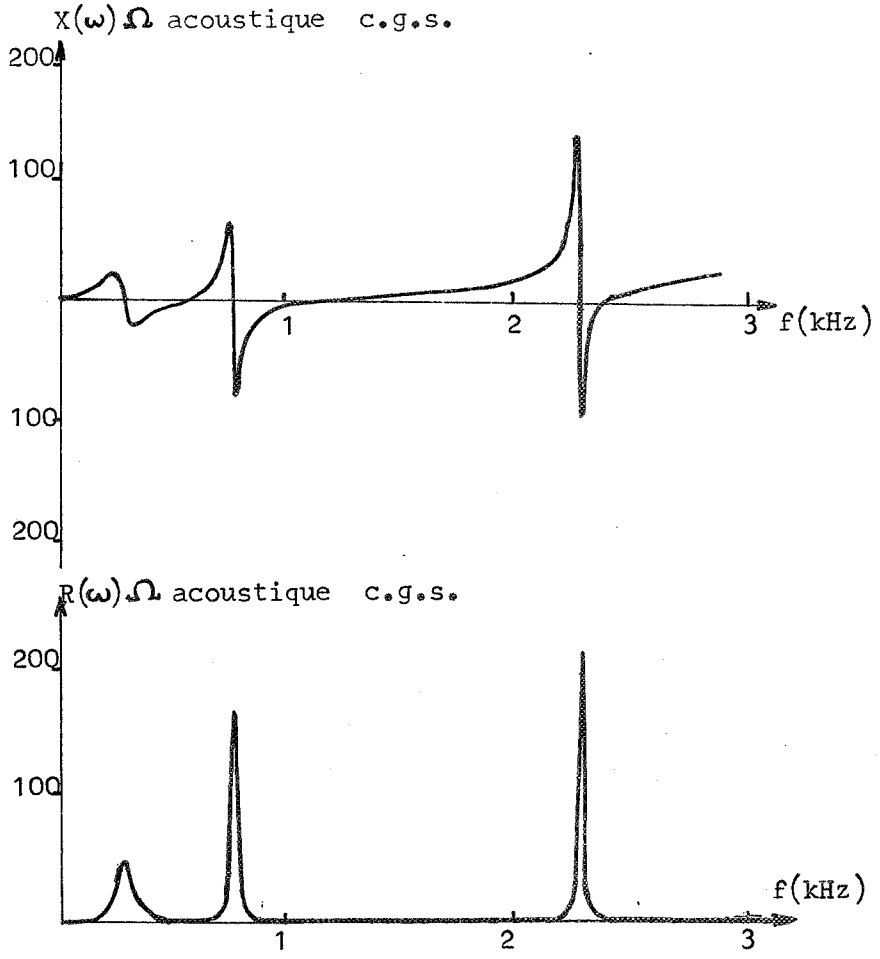


Figure 1 - Partie imaginaire $X(\omega)$ et partie réelle $R(\omega)$ de l'impédance d'entrée du conduit vocal pour la voyelle [u]

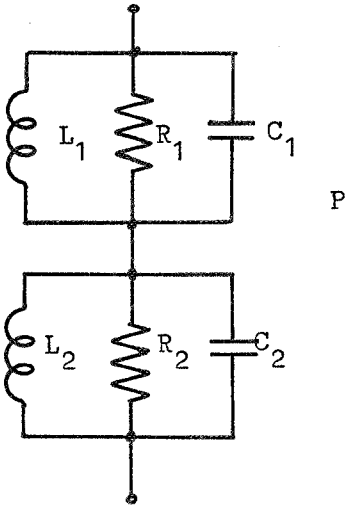


Figure 2 - Circuit équivalent de l'impédance d'entrée du conduit vocal pour les 2 premiers formants

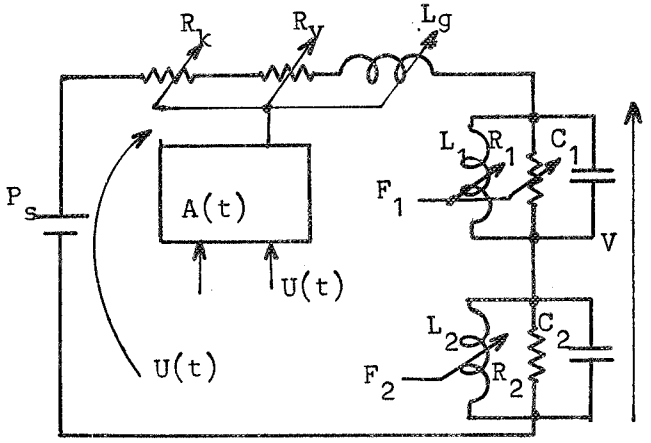


Figure 3 - Modèle de la source glottique chargé par un circuit équivalent à l'impédance d'entrée du conduit vocal

3-2 Résultats (GUERIN, MRAYATI, CARRE, 1976)

De la simulation du circuit de la figure 3, nous avons tiré deux séries de résultats : une série où on ne tenait pas compte du terme en $U(t).dL_g/dt$ et une série où ce terme était inclu.

La figure 4 donne la forme des signaux U(t) obtenue pour 3 voyelles : [a] , [o] et [u] . On remarque, en comparant les différentes formes d'onde de débit avec celle de A(t), que les ondes U(t) ont une fluctuation additionnelle à une fréquence proche de celle du premier formant de la voyelle considérée. Ces ondulations constituent un effet de l'intéraction source-conduit vocal.

En fait, l'impédance d'entrée du conduit vocal est importante pour des fréquences proches de celles d'un formant. L'amplitude de la partie réelle peut être du même ordre ou même plus grande que la résistance différentielle équivalente de la glotte. En conséquence, l'amplitude des composantes spectrales proche du premier formant sera diminuée. La figure 5 donne le spectre de l'onde U(t) comparés à celui de la fonction d'aire A(t). On remarque que l'impédance d'entrée du conduit vocal introduit un "zéro" supplémentaire à une fréquence proche de celle du premier formant. Ce phénomène doit être présent pour les autres formants mais atténué. Pour bien illustrer cette dépendance du "zéro" avec l'impédance d'entrée du conduit vocal, nous avons augmenté artificiellement le couplage source conduit en augmentant la résistance R_1. La figure 6 montre que plus élevée est la valeur de R_1, plus important sera le "zéro" supplémentaire. Nous pouvons donc conclure que le couplage source - conduit vocal introduit un "zéro" dans le spectre du débit de la source, à une fréquence proche de celle du premier formant.

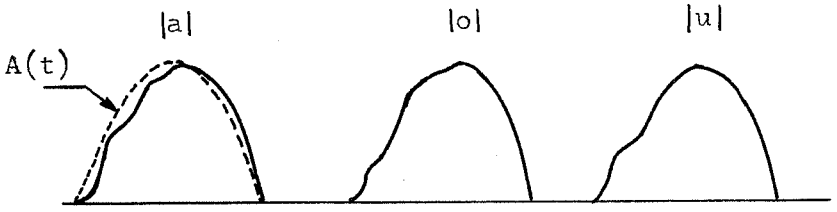


Figure 4 - Onde de débit U(t) calculé pour 3 voyelles [a] , [o] et [u] ($(dL_g)dt$ n'est pas inclus)

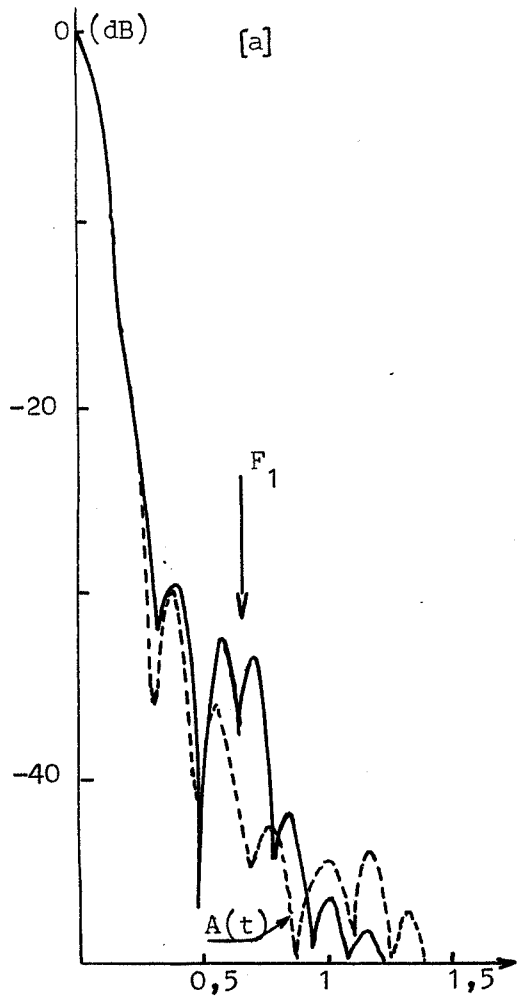


Figure 5 - Spectre d'amplitude de l'onde de débit comparé à celui de A(t)

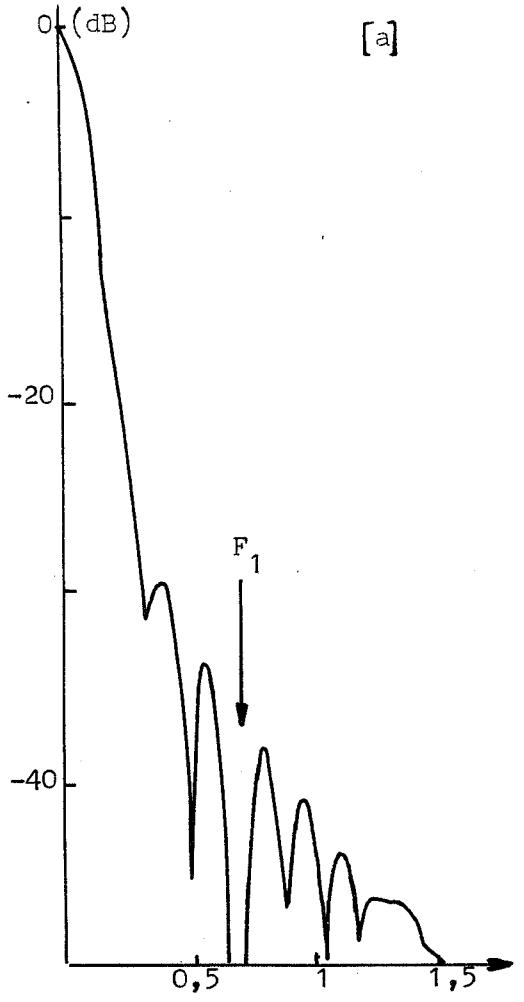


Figure 6 - Sensibilité du "zéro" avec R_1. **Par rapport à la figure 5, R_1 est 5 fois plus grand**

3-3 Influence de l'Inertance de la glotte

La dépendance de L_g avec le temps introduit le terme dL_g/dt dans le système d'équations différentielles de la source vocale. On peut reprendre les études précédentes en tenant compte de ce terme. Les formes d'onde de débit alors obtenues sont données figure 7. Sur cette figure, on peut faire les remarques suivantes :

- des perturbations à l'ouverture et à la fermeture de la glotte peuvent apparaître (dLg/dt est alors important). Ces perturbations accroissent les composantes hautes fréquences ;

- les ondulations de l'onde U(t) sont plus marquées.

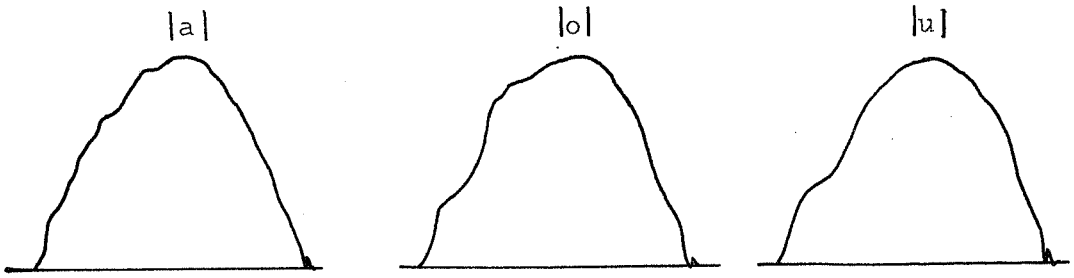


Figure 7 - Onde dé débit U(t) calculé pour trois voyelles [a] , [o] , [u] , en tenant compte du terme dL_G/dt

4 - INFLUENCE DE L'IMPEDANCE D'ENTREE SUR LA FREQUENCE INTRINSEQUE DE LA SOURCE VOCALE

4-1 Introduction

Il est bien connu que les voyelles "hautes" ont une mélodie intrinsèque plus élevée que les voyelles "basses". Généralement, ce phénomène est expliqué par l'interdépendance mécanique entre la configuration supraglottique et la position du larynx. Les mouvements du larynx changent les caractéristiques mécaniques des cordes vocales et par conséquence leur fréquence d'oscillation (KINTZING, SONESSON, 1967). Mais deux autres types de couplage peuvent opérer :

1 - Pour les voyelles très fermées, la constriction du conduit vocal introduit une résistance au flux d'air qui tend à augmenter la pression à travers la glotte. Ceci introduit une légère diminution de la fréquence fondamentale pour certaines voyelles, mais cette variation est opposée à celle observée sur la parole naturelle ;

2 - L'impédance d'entrée des cavités supraglottiques est une charge acoustique qui affecte le comportement mécanique des cordes vocales et par là, leur fréquence d'oscillation propre. Nous avons étudié, par simulation, les effets de ce couplage. Nous utiliserons pour ceci un modèle complet de la source vocale à 1 ou 2 masses (FLANAGAN et LANDGRAF 1968, ISHIZAKA , FLANAGAN 1972). Une expérience sur des voyelles naturelles viendra confirmer les résultats de la simulation.

4-2 Simulation numérique. Résultats

L'étude de la fréquence fondamentale a été faite en utilisant le circuit de la figure 3 auquel nous avons incorporé le modèle mécanique des cordes vocales à 1 ou 2 masses. L'impédance d'entrée du conduit vocal est simulé avec précision avec les valeurs déterminées par MRAYATI (1976). Les résultats obtenus sont donnés figure 8. On observe que pour le modèle à 1 masse, les variations sont importantes, environ 40 Hz, alors que pour le modèle à 2 masses, elles sont seulement de 7 Hz, mais les deux courbes gardent la même forme. Le modèle à 1 masse est donc très sensible aux variations de l'impédance d'entrée alors que le modèle à 2 masses donne une amplitude plus proche de la réalité. On a vérifié d'autre part que le couplage source-conduit par l'intermédiaire de l'impédance d'entrée

était principalement lié à la fréquence de 1er formant (figure 9). Le sens de variation de la fréquence de mélodie est trouvé opposé à celui observé sur la parole naturelle (figure 10). On peut donc en conclure que le couplage physiologique entre le conduit vocal et la source glottique est le plus important car son effet est opposé aux deux autres types de couplage décrits plus haut et donne en définitive une variation en accord avec les mesures sur la parole naturelle.

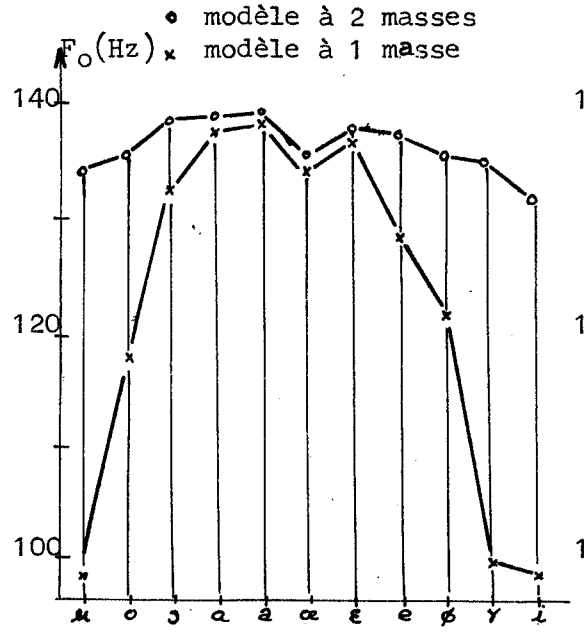


Figure 8 - Fréquence fondamentale calculée pour les voyelles du français

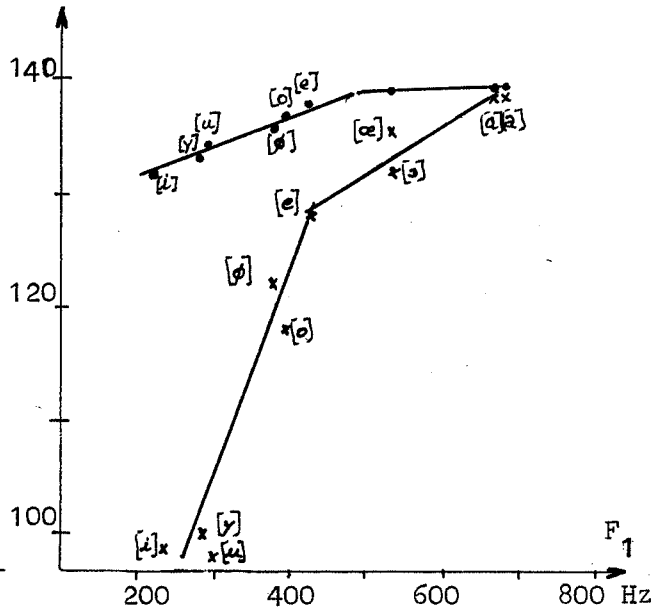


Figure 9 - Fréquence fondamentale **calculée en fonction de la valeur de la fréquence** du 1er formant

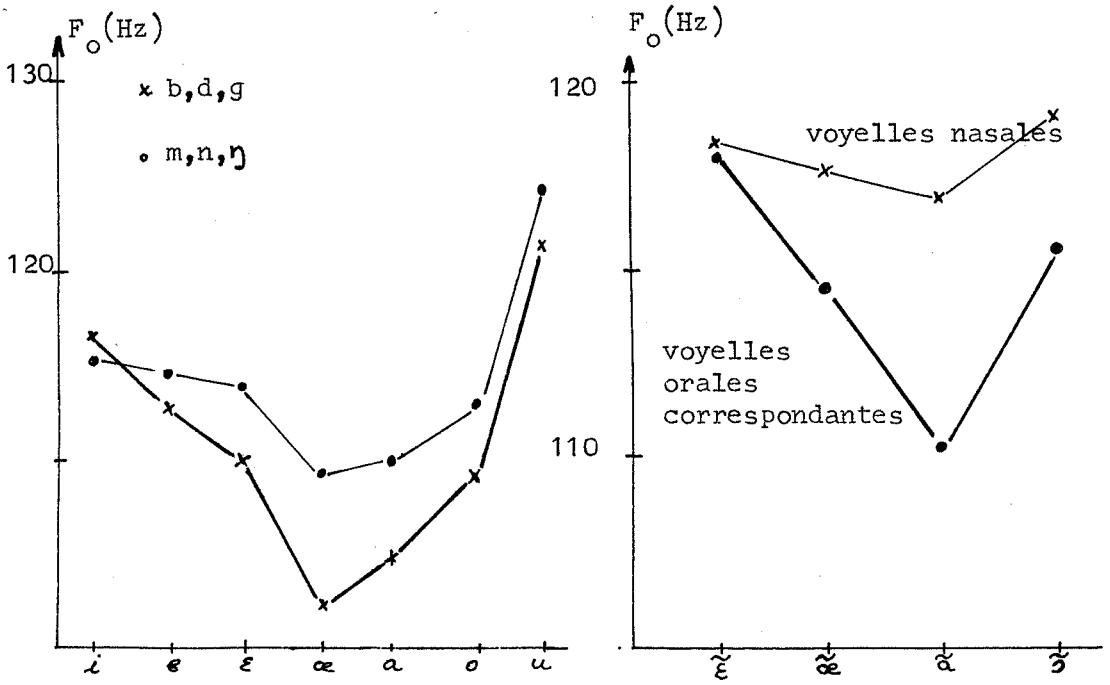


Figure 10 - Fréquence fondamentale intrinsèque des voyelles mesurée sur la parole naturelle suivant l'environnement consonnantique ou la nasalisation

4-3 Vérification expérimentale et justification théorique de la fréquence fondamentale

Une vérification expérimentale des effets dus à la charge acoustique présentée par le conduit vocal sur la fréquence de mélodie a été faite en mettant un tube d'environ 10 cm de longueur et 5 cm de diamètre devant la bouche d'un locuteur prononçant la voyelle [a] . Cette voyelle tend à se transformer en un son proche de [ɔ] . Par analyse des deux sons, on observe une diminution des fréquences de formant et de la fréquence fondamentale. Les résultats obtenus sont :

- pour [a] : F_o=104,7 Hz, F_1=850 Hz, F_2=1302 Hz et F_3=3190 Hz
- pour le son comme [ɔ] : F_o=102,4 Hz, F_1=693 Hz, F_2=1027 Hz et F_3=2918 Hz

Ces résultats sont en accord avec ceux que l'on a obtenus par simulation d'un modèle à 1 ou 2 masses. Pour les voyelles [a] et [ɔ] (voir figure 9). D'une manière générale, on constate, sur cette figure 9, que la fréquence fondamentale est liée fortement à la fréquence du premier formant.

On peut étudier le lien entre la fréquence d'oscillation et la valeur de l'impédance d'entrée. En effet, le fonctionnement mécanique des cordes vocales peut être approché par un circuit électrique du type oscillateur à relaxation (cas du modèle à 1 masse) (PAILLE, 1971). Sur la figure 11, on voit que cet oscillateur est chargé par le circuit équivalent à l'impédance d'entrée du conduit vocal ; les composantes réactives et passives de ce circuit de charge modifieront la fréquence d'oscillation propre de l'oscillateur. Or, aux fréquences habituelles de F_o, entre 80 et 250 Hz, l'impédance d'entrée équivalente du conduit vocal a un comportement selfique, c'est-à-dire qu'elle peut être représenté par une inductance en parallèle avec une résistance. Cette inductance vient en série avec l'inductance qui représente la masse mécanique des cordes vocales, l'ensemble aura donc une fréquence d'oscillation propre inférieure à celle que l'on aurait sans charge. Il y aura donc une liaison étroite entre: la valeur de la partie réactive et de la partie passive de l'impédance d'entrée, et la fréquence d'oscillation des cordes vocales. On vérifie cette dépendance en traçant la courbe représentant les valeurs de F_o que l'on a calculé en fonction du rapport $X(\omega)/R(\omega)$, à F_o = 136 Hz (X et R sont les valeurs de la réactance et de la résistance de l'impédance d'entrée à 136 Hz). La figure 12 montre que cette relation $F_o = f(X/R)$ est généralement bien vérifiée. ISHIZAKA et FLANAGAN (1972) décrivent une expérience où on ajoute un tube cylindrique de longueur variable devant la bouche d'un locuteur. On mesure alors la fréquence fondamentale et on obtient une évolution qui peut être expliquée par le lien F_o-impédance d'entrée du conduit vocal décrit plus haut. En effet, le comportement capacitif de l'impédance d'entrée aura l'effet contraire, sur F_o, à celui de son comportement inductif.

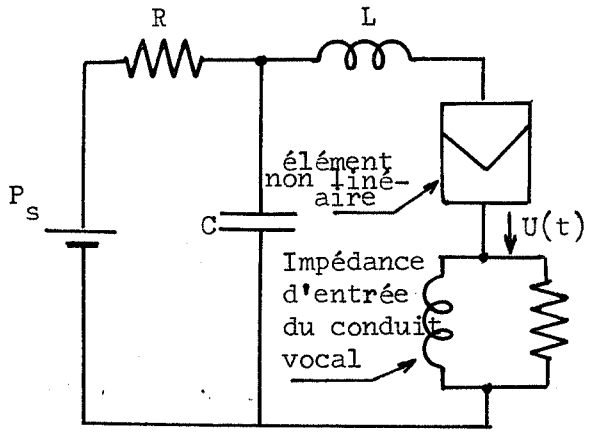


Figure 11 - Modèle électrique équivalent aux cordes vocales

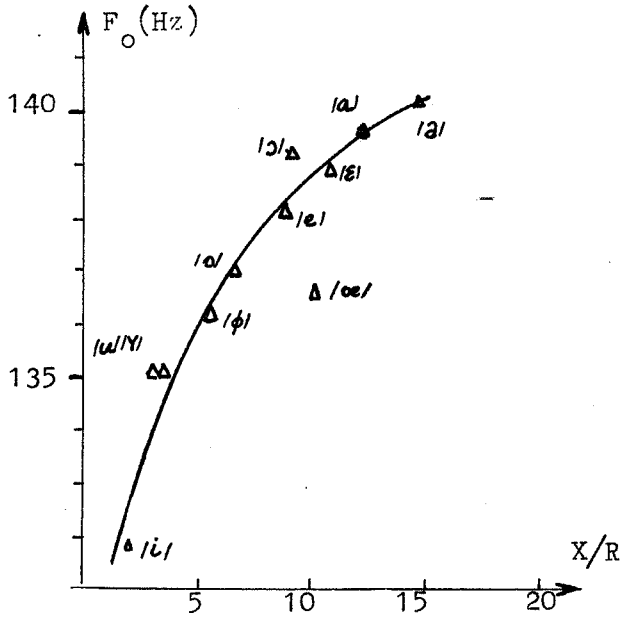


Figure 12 - Variation de la fréquence fondamentale avec la valeur de l'impédance d'entrée

5 - CONCLUSIONS

Les différentes simulations de la source vocale décrites ci-dessus ont permis d'étudier deux aspects du couplage source - conduit vocal.

Le premier aspect se situe au niveau de la forme temporelle du signal de débit de la glotte. On a constaté que la configuration du conduit vocal, par son impédance d'entrée acoustique variable vue de la glotte, modifiait notablement la forme et donc le spectre du signal de débit de la source. On a montré que le premier formant joue un rôle important dans ce couplage en ajoutant un zéro supplémentaire dans le spectre du débit. Une étude de l'effet de ce couplage au niveau perceptif est projetée.

Le deuxième aspect se situe au niveau de la valeur de la fréquence fondamentale d'oscillation de la source. La charge acoustique présentée par le conduit vocal modifie légèrement cette fréquence d'oscillation, mais dans un sens contraire à celui que l'on observe, pour les différentes voyelles françaises, sur la parole naturelle. On peut justifié ce sens de variation en considérant la valeur de l'impédance d'entrée aux fréquences proches de F_o. On déduit de cette étude que le couplage physiologique qui existe entre les cordes vocales et la configuration du conduit vocal est le principal responsable des différentes valeurs de la fréquence fondamentale intrinsèque des voyelles françaises. On note, sur la figure 10, que la fréquence de mélodie intrinsèque dépend aussi de l'environnement consonnantique (CARRE, 1975). Dans le cas de consonnes nasales, le couplage physiologique ne peut pas expliquer complètement les différences constatées. L'impédance d'entrée joue peut-être un rôle plus important dans le cas où il y a nasalisation.

BIBLIOGRAPHIE

CARRE R. (1975)
Acoustic characteristics of vowel nasalization
Q.P.R. N° 115, Res. Lab. of Electronics, M.I.T., 270-274

FLANAGAN J., LANDGRAF L. (1968)
Self-oscillating source for vocal tract synthesizers
I.E.E.E. Trans., AU-16, n° 1, 57-64

GUERIN B., MRAYATI M., CARRE R. (1976)
A voice source taking account of coupling with the supraglottal cavities
I.E.E.E. International Conference on Acoustics, Speech and signal Processing - Philadelphia

ISHIZAKA K., FLANAGAN J. (1972)
Synthesis of voiced sounds from a two-mass model of the vocal cords
B.S.T.J., 51, 1233-1268

KINTZING P., SONESSON B. (1967)
Shape and shift of the laryngeal ventricle during phonation
Acta Oto-Laryngologica, 63, 479-488

MRAYATI M. (1976)
Contribution aux études sur la production de la parole
Thèse de docteur d'état - Université de Grenoble

MRAYATI M., GUERIN B., BOE L.J. (1976)
Etude de l'impédance d'entrée du conduit vocal : couplage source - conduit vocal
Acustica, 35, n° 2

PAILLE J. (1971)
Contribution aux études sur la synthèse paramétrique de la parole - Synthetiseur à formants. Analogue de la source vocale
Thèse de docteur d'état - Université de GRENOBLE

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

SYNTHESE PAR REGLES DE L'INTONATION DE LA PHRASE

Philippe MARTIN
Experimental Phonetics Laboratory
University of Toronto
Institut de Phonétique
Université Libre de Bruxelles

RESUME :

Les procédés à codage prédictif permettent d'obtenir un signal de parole synthétique de très bonne qualité. Pour réaliser la synthèse par règles de phrases, on propose une méthode procédant par concaténation de syntagmes minimaux générés séparément, et assurant la variation des paramètres de la source voisée (principalement les contours mélodiques) de manière à la faire correspondre à la structure syntaxique de la phrase. Cette structure syntaxique est obtenue par un algorithme d'analyse opérant à partir des marques de présupposition que présentent les éléments constitutifs de l'énoncé.

SUMMARY :

A very good synthesized speech signal is obtained by the use of linear predictors methods. In order to synthesize by rule a whole sentence, we propose a method using the concatenation of minimal segments synthesized separately. The parameters of the source corresponding to the prosodic features of the sentence are then generated according to the syntactic structure of the sentence. This syntactic structure is derived by an algorithm from the presupposition marks of the sentence units.

SYNTHESE PAR REGLES DE L'INTONATION DE LA PHRASE

Philippe MARTIN

0. Les méthodes à codage prédictif développées récemment permettent d'obtenir par analyse-synthèse un signal de parole de très bonne qualité. Pour réaliser la synthèse par règles d'une phrase entière, on peut envisager une extension relativement simple de ces techniques en opérant la concaténation d'unités syntaxiques minimales traitées indépendamment. Il se pose alors le problème de la génération des éléments prosodiques de l'énoncé, responsables de la variation des paramètres de contrôle de la partie voisée de la source, tels que la fréquence fondamentale, l'intensité, les pauses, et éventuellement la durée syllabique.

A cet effet, on propose ici un dispositif algorithmique basé sur une théorie de l'intonation exposée ailleurs (Martin, 1975b, 1976) et assumant un parallélisme entre les structures syntaxique et mélodique de la phrase.

Ce dispositif comprend deux parties:

- une analyse syntaxique d'un type particulier, axée sur les rapports de présupposition entretenus par les syntagmes minimaux;
- un ensemble de règles d'attribution de facteurs prosodiques (principalement de contours mélodiques) à chaque syntagme minimal, opérant à partir des résultats de l'analyse syntaxique.

Le procédé de synthèse, dont on ne donnera ici que quelques caractéristiques, évite donc, par le stockage des unités syntaxiques minimales, les problèmes posés par la concaténation de phonèmes ou de syllabes isolées.

1. La méthode repose dans son principe sur le parallélisme des structures mélodique et syntaxique de la phrase. Toutes les neutralisations des traits prosodiques, et plus particulièrement des contours mélodiques qui se produisent souvent dans un énoncé français sont donc supposées résolues selon les indications fournies par l'arrangement syntaxique. Cette hypothèse exclut la présence d'une ambiguïté syntaxique dans l'énoncé (ambiguïté qui serait précisément levée par des marques prosodiques).

En imposant à la prosodie de correspondre à l'arrangement syntaxique, la méthode adopte une solution correcte parmi d'autres prosodies admissibles pour la phrase.

2. Par structure syntaxique, on entendra

ici une classification hiérarchique particulière obtenue par l'analyse des relations de présupposition nouées entre les unités minimales constitutives de l'énoncé.

Entre deux éléments A et B dans l'axe syntagmatique peut exister un des quatre rapports de présupposition suivants:

A ⟶ B : A présuppose B, B ne présuppose pas A, cas de subordination (sélection) à droite. Dans "la grande Marie", l'unité (la grande) présuppose (Marie), alors que l'inverse n'est pas vrai;

A ⟵ B : A ne présuppose pas B, B présuppose A, cas de subordination (sélection) à gauche. Dans "Marie entrait", l'unité (entrait) présuppose (Marie), alors que (Marie) ne présuppose pas (entrait) dans l'énoncé;

A ⟷ B : A et B se présupposent mutuellement, cas d'une prédication (solidarité). Dans "il entrait", le pronom (il) présuppose la présence du verbe (entrait) et réciproquement;

A — B : A ne présuppose pas B et B ne présuppose pas A, cas d'une coordination (combinaison). Dans "ce matin Marie entrait", l'unité (ce matin) ne présuppose ni n'est présupposée par aucune autre unité de l'énoncé.

3. Corrélatives des relations de présupposition entre unités existent des marques qui sont soit externes soit internes au syntagme. Des marques externes sont par exemple les conjonctions de coordination ((et) dans "Pierre et Marie", indiquant la relation A - B) ou les prépositions ((de) dans "le fils de Marie", indiquant la relation A ← B).

Des marques internes sont constituées par les flexions des unités, qui peuvent varier en genre, nombre, personne, cas, etc. Il peut alors y avoir accord entre unités si des marques variables ont la même valeur dans les deux unités concernées (dans "le grand chien", l'article et l'adjectif s'accordent en genre et en nombre), ou rection si la marque de l'unité régissante est invariable alors que celle de l'unité régie, de même valeur, est variable (dans "la grande chaise", l'adjectif est régi par le nom, dont la marque de genre est invariable). L'accord est corrélatif d'une solidarité A - B, alors que la rection indique une sélection de type A ← B ou A → B.

Un élément donné possède une ou plusieurs marques variables s'il existe d'autres formes paradigmatiques de cet élément. Ainsi l'adjectif (grand) possède la marque de genre qui peut prendre les valeurs m ou f puisque les formes (grand) et (grande) existent, alors que les valeurs de la marque de nombre sont neutralisées dans la parole (mais non dans la graphie), puisque les

formes (grand) et (grands) ou (grande) et (grandes) sont respectivement les mêmes quant au nombre. Il y a dans ce cas (toujours sur le plan de la langue parlée) syncrétisme de la marque de nombre. Les marques externes sont utilisées lorsque les unités impliquées ont des marques invariables.

On conviendra de représenter par une majuscule la valeur d'une marque invariable, par une minuscule la valeur d'une marque variable, par des parenthèses les valeurs neutralisées (résolues ou non), par des flèches → ou ← le caractère directionnel éventuel des marques dans leur fonction de présupposition, et enfin par les symboles ——, —→ ou ←— les indications données directement par les marques externes.

L'exemple (le garçon blond) s'analyse alors de la manière suivante:

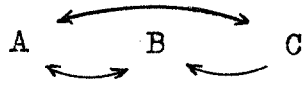
	le	garçon	blond
genre	m→	M	m
nombre	s→	( )	( )
personne	3→	( )	( )
cas	( )	( )	( )
	A	B	C

Il y a syncrétisme de la marque de personne pour le nom et l'adjectif entre les valeurs 2ème et 3ème. D'autre part, les marques de l'article sont toujours présupposantes à droite.

La marque de genre est corrélative d'une double rection A → B et B ← C, alors que la marque de nombre, révélée par un double accord neutralisé, indique les relations de solidarité A ↔ B et A ↔ C. On peut en outre admettre qu'il y a rection entre l'article et le groupe nom + adjectif du fait de la marque de nombre.

Lorsque des relations différentes, dues à des marques distinctes, s'établissent entre deux mêmes unités, c'est évidemment la relation la plus "forte" qui l'emporte, selon l'ordre décroissant: solidarité, sélection, combinaison.

Les relations entre les éléments A, B et C de l'exemple sont donc finalement données par:



4. L'algorithme d'analyse syntaxique doit opérer le regroupement hiérarchique des syntagmes minimaux à partir des marques de présupposition repérées directement lors de la mise en mémoire des différentes formes analysées. L'identification de ces marques à partir de la seule forme des unités, dont la mise en oeuvre constitue une des difficultés principales de l'analyse syntaxique automatique, est donc évitée. En contrepartie,

chaque forme d'un même syntagme doit être analysée et mise en mémoire indépendamment. En l'absence de neutralisation ceci multiplie par exemple le nombre de substantifs à stocker par deux (formes du singulier et du pluriel), et ceux des adjectifs par quatre.

Les n unités d'une phrase entretiennent entre elles n/2.(n-1) relations de présupposition binaires. Si m est le nombre de marques syntaxiques distinctes, il y a m.n/2.(n-1) couples de valeurs à examiner. L'examen d'une marque variable s'arrête lorsqu'une rection avec une marque invariable a été trouvée. Elle continue dans le cas d'un accord avec une autre marque variable.

L'algorithme procède en inspectant les unités de plus en plus éloignées de l'unité considérée dans les deux directions (à moins que l'une des marques ne soit unidirectionnelle).

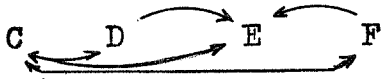
Soit l'exemple suivant, pourvu des valeurs des différentes marques syntaxiques:

ce matin le grand garçon blond mangeait une pomme

genre	m→	M	m→	m	M	m		f→	F
nombre	s→	( )	s→	( )	( )	( )		s→	( )
personne	3→	( )	3→	( )	( )	( )	3	3→	( )
(cas)									
(prép.)									
	A	B	C	D	E	F	G	H	J

A présuppose B par le trait de genre, B présuppose A par le trait de personne, A et B sont donc solidaires. B n'étant en rapport de présupposition avec aucune autre unité que A, le groupe A ↔ B est en rapport de combinaison avec le reste de l'énoncé.

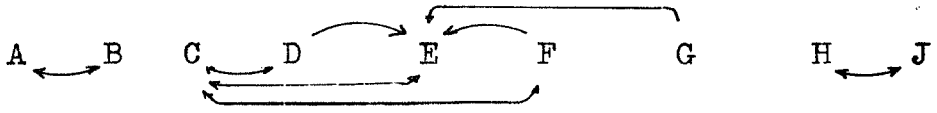
L'article (le) est solidaire des unités D, E et F dans le groupe (le grand garçon blond), et confère aux unités D E F la valeur 3ème personne. Les relations sont données par:



(En fait l'article C est solidaire de D, E et F, ainsi que de touss les groupes plus grands que l'on peut former avec ces unités).

Le verbe (mangeait) sélectionne le groupe nominal (CDEF), et les éléments H et J sont solidaires.

Finalement, les rapports entre les 9 syntagmes minimaux de cet exemple s'expriment par:



On constate que, d'après cette analyse, le groupe H J n'est pas présupposant par rapport au verbe

G. On pourrait alors poser qu'il existe une marque de cas de valeur neutralisée régie par le verbe. En fait, ce n'est qu'en l'absence de toute autre unité dans ce contexte que H J apparaît comme subordonné au verbe.

Relativement aux marques de présupposition considérées, la structure syntaxique de l'exemple s'établit selon le parenthésage:

(A B) ((((C D) E) F) (G (H J)))

Les groupes entre crochets contiennent des unités solidaires, l'article se liant avec l'unité solidaire la plus proche.

Pour se conformer à la théorie prosodique, il faut ensuite rectifier éventuellement le parenthésage de manière à ce que l'énoncé se divise selon les plus grandes unités successives qui ne soient pas solidaires entre elles, ce qui donne:

(A B) (((C D) E) F) (G (H J))

5. On trouvera ailleurs (Martin, 1975a, 1976) la façon de dériver une séquence de contours mélodiques à partir d'un parenthésage. Rappellons qu'il n'y a qu'un seul contour à attribuer par unité minimale qui ne soit pas solidaire d'une autre unité. Ainsi (A B), (C D) et (H J) ne reçoivent qu'un contour, porté par la dernière syllabe prononçée du groupe.

Pour une phrase déclarative, les règles sont:

1) contour final C_o ⟍
2) 1er niveau C_1 ⟋ (contraste de pente avec C_o)
3) 2ème niveau C_2 ⟍ devant C_1 , C_3 ⟋ devant C_o
4) 3ème niveau C_3 ⟋ devant C_2

En les appliquant à l'exemple, on a:

(A B)(((C D) E) F) (G (H J))

C_1 C_3 C_2 C_1 C_3 C_o

La séquence est donc:

ce matin/ le grand garçon blond/ mangeait une pomme.

C_1 C_3 C_2 C_1 C_3 C_o

6. La variation mélodique est, en français, la paramètre le plus important parmi les facteurs

prosodiques, particulièrement en ce qui concerne la perception de l'accent. Les autres paramètres, qui apparaissent en distribution complémentaire avec les contours, demandent à être étudiés de manière plus complète.

La pause après chaque unité peut être rendue proportionnelle au niveau de cette unité dans la structure syntaxique (elle sera nulle entre deux unités solidaires).

Il faut tenir compte d'autre part de facteurs multiplicatifs agissant, au niveau phonétique, sur l'amplitude des contours mélodiques. Pour un même contour, cette amplitude décroit de manière sensiblement linéaire du début à la fin de l'énoncé. Ceci est sans doute lié au rendement du contraste de pente à assurer avec le contour final.

7. La mise en œuvre du dispositif proposé se fait par appel des coefficients de chaque syntagme minimal, soit directement (en donnant l'adresse des coefficients stockés), soit indirectement en donnant sa forme graphique, ce qui nécessite la mise en mémoire d'un dictionnaire. En adoptant, pour fixer les idées, une durée moyenne de 0,3 s pour chaque unité minimale, et en utilisant une fenêtre de 30 ms et 12 coefficients, chaque unité nécessite environ 120 paramètres pour être représentée (en plus des paramètres relatifs aux marques syntaxiques, au nombre d'une dizaine).

Une réalisation cablée travaillant en temps réel et comprenant une mémoire morte de 64000 mots pourrait donc contenir un vocabulaire de près de 500 unités, suffisant pour la génération d'énoncés simples.

8. On a décrit une méthode de synthèse procédant par concaténation de syntagmes minimaux synthétisés indépendamment après analyse en coefficients prédictifs, et déterminant les variations des paramètres de la source voisée responsables des facteurs prosodiques de l'énoncé à partir d'une analyse syntaxique de cet énoncé. Les seules modifications par rapport à l'analyse ne concernant que les paramètres de la source, on doit s'attendre à obtenir de la parole de très bonne qualité.

Il reste alors à intégrer dans le système des règles de génération des paramètres prosodiques moins importants, et également à tenir compte, dans le processus de concaténation, des phénomènes deliaison et d'enchaînement au niveau des phonèmes.

Références

Martin, Ph. (1975a) Intonation et reconnaissance automatique de la structure syntaxique, 6èmes Journées d'Etudes sur la Parole, GALF, Toulouse.

(1975b) Eléments pour une théorie de l'into-

nation, Rapport d'activités de l'Institut de Phonétique, n° 9/1, Bruxelles.

(1976) Une grammaire de l'intonation de la phrase, Rapport d'activités de l'Institut de Phonétique, n° 9/2, Bruxelles.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

TRANSMISSION DE LA PAROLE A 1200 B/S

LE VOCODER CIPHON

Auteur : J. POTAGE

RESUME :

Le CIPHON est un vocoder utilisant la représentation par "lignes de crête" du signal de parole.

Cet article décrit sommairement les principes de l'analyse et de la synthèse par lignes de crête des spectres à court terme du signal de parole, prélevée dans la bande téléphonique. Grâce à cette représentation compacte du signal, une transmission en temps réel à 1.200 b/s devient réalisable.

SUMMARY :

The CIPHON is a vocoder which makes use of the "crest lines" representation of the voice signal.

This paper summarizes the principles of the crest lines analysis and synthesis of the short term spectra of the voice signal which has been extracted from the telephone bandwidth. By way of this dense representation of the signal, a real time 1 200 bits/sec. transmission is feasible.

TRANSMISSION DE LA PAROLE A 1200 b/s

LE VOCODER C I P H O N

Auteurs : J. POTAGE - J.S. BOURGENOT

1 - INTRODUCTION

Le vocoder CIPHON est un dispositif d'analyse, codage et synthèse en temps réel de la parole ayant pour objectifs la transmission numérique à 1200 b/s d'un signal téléphonique dans la bande 300-3400 Hz, avec une bonne qualité et sans dégradation notable de l'intelligibilité.

S'ils synthétisent une parole de qualité moyenne, les vocoders à canaux ne permettent guère de descendre à des débits de transmission inférieurs à 2400 b/s. Par ailleurs une meilleure qualité peut être obtenue avec les vocoders à formants, mais leur mise en oeuvre demeure complexe, surtout en temps réel.

Le CIPHON, en procédant à une analyse-synthèse plutôt du type formant, vise un débit de 1200 b/s avec une qualité comparable à celle du vocoder à formants tout en évitant sa complexité.

	V O C O D E R S			
PRINCIPE	à bande de base	à canaux	à lignes de crête	à formants
Nombre de canaux d'analyse	12 à 20	12 à 20	32 à 64	12 à 20
Décision voisé/non voisé	non	oui	non	oui
Débit b/s	9600	2400	1200	1200
Intelligibilité	Bonne	Bonne	Bonne	Bonne
Agrément	Bon	Moyen	Bon	Bon
Faisabilité Temps réel	oui	oui	oui	non

Fig. 1

La simulation du vocoder CIPHON a été effectuée dans les laboratoires de la Division Télécommunications de THOMSON-CSF, sur un contrat DRME.

Ainsi, à l'aide d'une batterie de filtres de type cochléaire, est-il possible d'obtenir périodiquement une suite $\{Sn\}$ de spectres à court terme du signal de parole (Fig. 3)

2-2-3 -Extraction des lignes de crête

La suite $\{Sn\}$ de spectres fournie par la batterie de filtres développe dans l'espace fréquence - amplitude - temps une surface dont les lignes de crête -obtenues en joignant de proche en proche les maxima relatifs des spectres- représentent l'évolution des pôles de la fonction de transfert du conduit vocal au cours du temps.

Ces lignes, qui durant les phonèmes stables doivent coïncider avec les formants constituent le "squelette sémantique" de la parole. (Fig. 3)

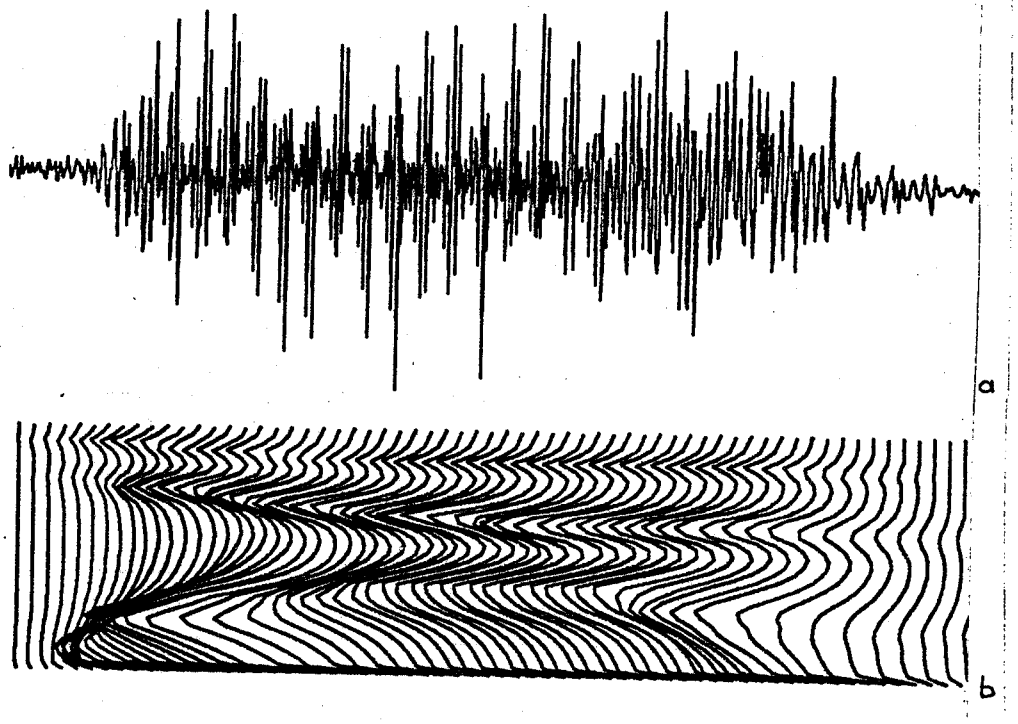


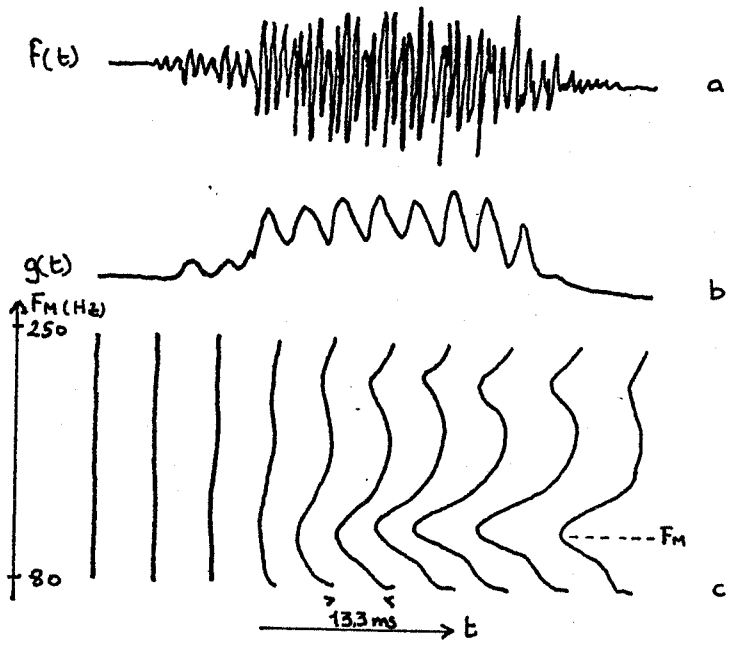
Fig 3

Il est possible de reconstituer l'intelligibilité du message à l'aide de ces lignes de crête comme il est indiqué plus loin. Cependant, une autre information est nécessaire pour reproduire l'intonation : la fréquence de mélodie qui correspond à la fréquence d'excitation des cordes vocales.

2-3 .Analyse de la mélodie

Le signal de sortie d'un canal large bande suivi d'une détection est caractérisé par une modulation d'amplitude en présence de sons voisés. Une analyse à bande étroite de ce signal dans la

bande 80-250 Hz (recouvrant les fréquences de mélodie de la plupart des locuteurs) mettra en évidence un pic correspondant à la fréquence de mélodie que l'on peut ainsi déterminer.



Signal de parole

Sortie redressée d'un filtre large bande

Spectres successifs du signal g(t) obtenus par une batterie de filtres à bande étroite.

Fig 4

2-4 .Synthèse par lignes de crête

2-4-1 -Principe :

La synthèse par lignes de crête consiste à associer un générateur de sinusoïde à chaque ligne de crête et à asservir la fréquence et l'amplitude de ce générateur aux coordonnées de la ligne de crête dans l'espace fréquence - amplitude - temps.

2-4-2 -Synthèse sans mélodie

La parole synthétique est composée d'une suite d'échantillons xn donnés par la formule :

$$x_n = \sum_{i=1}^{L} A_i \sin \varphi_{i,n}$$

avec $\varphi_{i,n} = \varphi_{i,n-1} + 2\pi \frac{F_i}{F_E}$

où
- n = Indice de temps
- Ai = Amplitude de la ligne de crête i
- $\varphi_{i,n}$ = Phase de la ligne de crête i à l'instant n
- Fi = Fréquence de la ligne de crête i
- FE = Fréquence de calcul de l'échantillon xn
- L = Nombre total de ligne de crête.

2-4-3 -Synthèse avec mélodie

Afin de restituer l'intonation, on rend périodique chacune des sinusoïdes et ainsi leur somme en remettant leur phase à zéro avec une fréquence égale à celle de la mélodie Fig. 5a. Le saut de phase de la fonction somme obtenue est d'autre part supprimé par une modulation d'amplitude à 100 % de la fonction somme par une sinusoïde de fréquence égale à celle de la mélodie Fig. 5c.

Chaque échantillon est alors donné par la formule :

$$x_n = \sum_{i=1}^{L} (A_i \sin \varphi_{i,n})(1 + \sin \varphi_{m,n})$$

avec

$$\varphi_{m,n} = \varphi_{m,n-1} + 2\pi \frac{F_M}{F_E}$$

et : F_M = fréquence de mélodie

$\varphi_{m,n}$ = phase instantanée mélodie

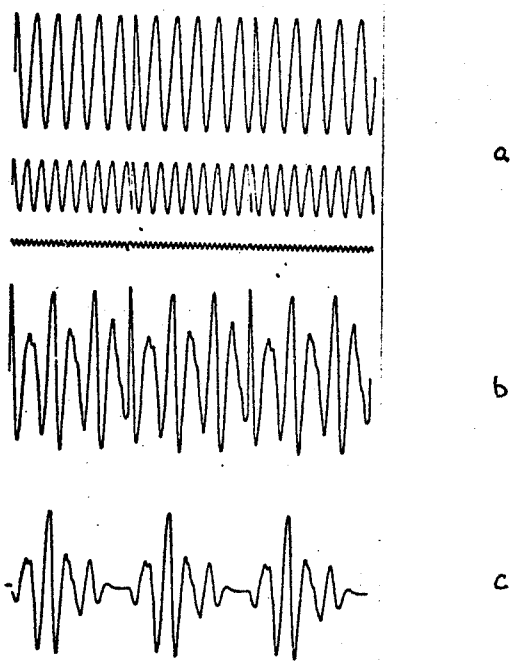


Fig 5

3 - DESCRIPTION DU VOCODER CIPHON

L'organisation générale du vocoder est donnée Fig. 6

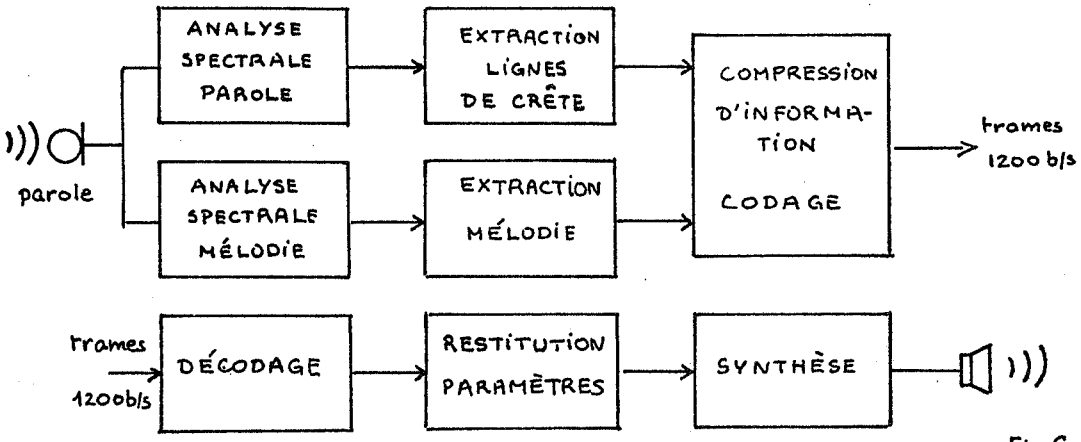


Fig 6

3-1 .Analyse spectrale

L'analyseur spectral du CIPHON est constitué de 2 batteries de filtres numériques permettant de calculer :

- les spectres à court terme du signal de parole (batterie de filtres parole)
- le spectre du signal prélevé dans un canal particulier, afin d'obtenir la mélodie (batteries de filtres mélodie)

3-1-2 -Schéma de principe :

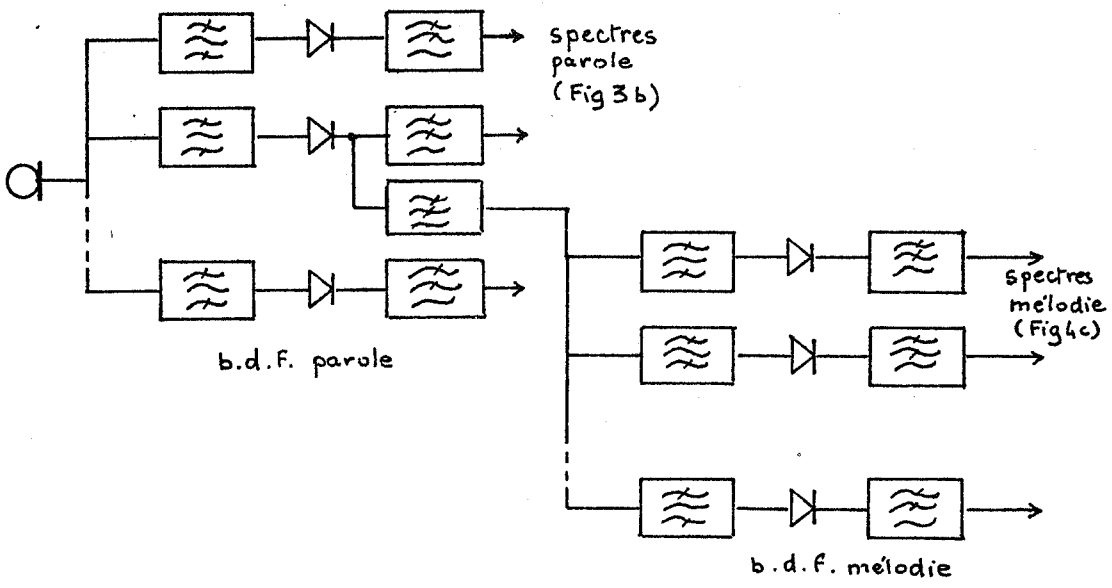


Fig7

3-1-3 -Caractéristiques : Fig. 8

Fig. 8

	BATTERIE DE FILTRES	
	PAROLE	MELODIE
Nombre de canaux	32	32
Filtres passe-bandes numériques	Résonateur type RLC	Résonateur type RLC
Répartition des fréquences centrales	Logarithmique	Logarithmique
Répartition de la Sélectivité	Linéaire	Constante
Filtres passe-bas Fréquence de coupure	50 Hz	200 Hz
Période de calcul d'un spectre	3,33 ms	13,33 ms

Les lois de variation des fréquences centrales et de la sélectivité ont été obtenues en recherchant le meilleur compromis entre une localisation fine des lignes de crête (sans pour autant isoler les harmoniques de la mélodie : filtres à bande étroite donc à grande constante de temps) et une bonne réponse aux transitoires (filtres à bande large, c'est-à-dire à faible constante de temps.

3-2 .Extraction des paramètres

3-2-1 -Extraction des lignes de crête

L'extraction des lignes de crête se fait en trois temps :

- détection des maxima relatifs des spectres à court-terme
- choix des 5 plus grands maxima détectés
- recherche des lignes de crête parmi les maxima retenus

La détection des maxima est basée sur les passages à zéro de la dérivée première des spectres en fonction de la fréquence.

La recherche des lignes de crête consiste à repérer la naissance d'une ligne de crête (INITIALISATION) et ensuite à pister celle-ci parmi les maxima détectés (PISTAGE). Trois lignes de crête sont retenues par l'algorithme.

3-2-2 -Extraction de la fréquence de mélodie

Le repérage de l'abscisse du maximum maximorum des spectres "mélodie" fournit la fréquence de mélodie.

3-3-3 -Débit binaire nécessaire à la transmission des paramètres

En consacrant 5 bits aux fréquences (de lignes de crête, de mélodie) et 4 bits aux amplitudes (codées sur une dynamique de 45 dB par pas de 3dB) et en transmettant les coordonnées des lignes de crête toutes les 3,33 ms et la mélodie toutes les I3,33 ms le débit de transmission est donné Fig. 9 en fonction du nombre n de lignes de crête transmises.

n	Débit b/s
2	5775
3	8475

Fig 9

Afin de descendre à 1200 b/s un codage adaptatif des paramètres est nécessaire.

3-3 .Codage adaptatif des paramètres à 1200 b/s

3-3-1 -Propriétés spécifiques des lignes de crête : (Fig 10 a)

L'évolution des lignes de crête au cours du message parlé est caractérisée par deux types de variations :

- des variations rapides en fréquence et en amplitude lors des attaques des phonèmes
- des variations lentes, voire nulles sur les parties stables des phonèmes ; dans ce dernier cas, les amplitudes sont d'autre part très correlées entre elles.

3-3-2 -Codage adaptatif à 1200 b/s : (Fig 10 b)

Mettant à profit les propriétés énoncées plus haut, un codage à préfixes permet de transmettre en mode différentiel les variations lentes des paramètres durant les périodes stationnaires (on transmet la variation par rapport à l'instant précédent des paramètres à coder).

Durant les transitoires, un mode absolu approché permet d'approximer les paramètres dans les limites tolérables par l'oreille.

Le code génère une trame de 32 bits toutes les 26,66 ms et le retard cumulé par l'analyse, le codage et la synthèse reste inférieur à 100 ms.

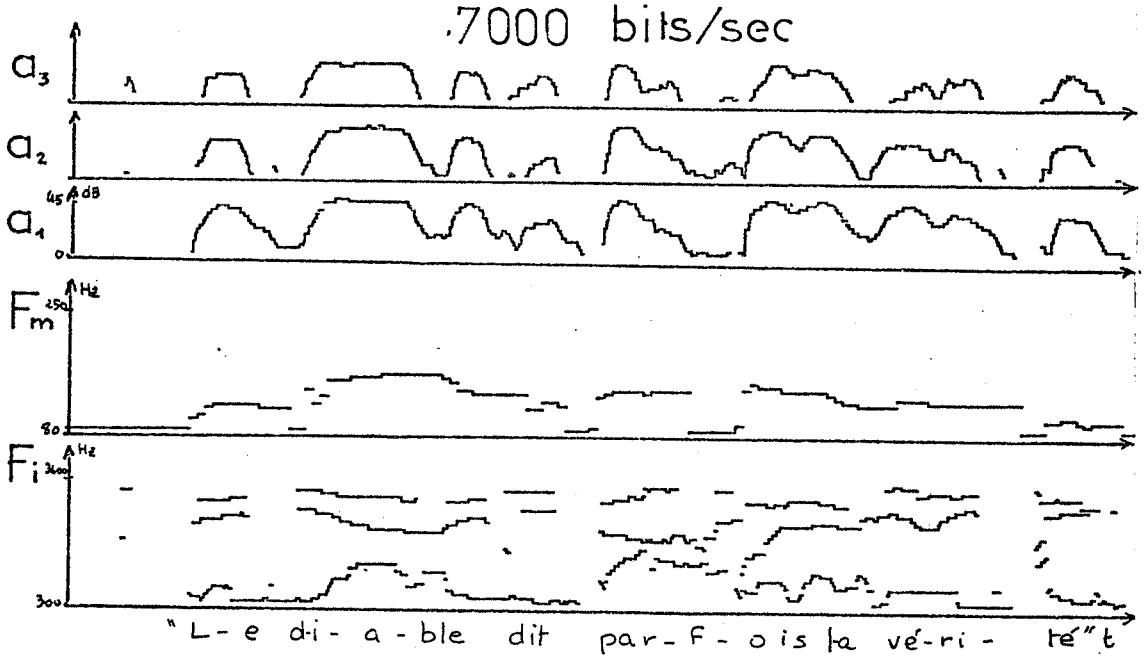


Fig10 a : Coordonnées des lignes de crête et fréquence de mélodie après analyse spectrale

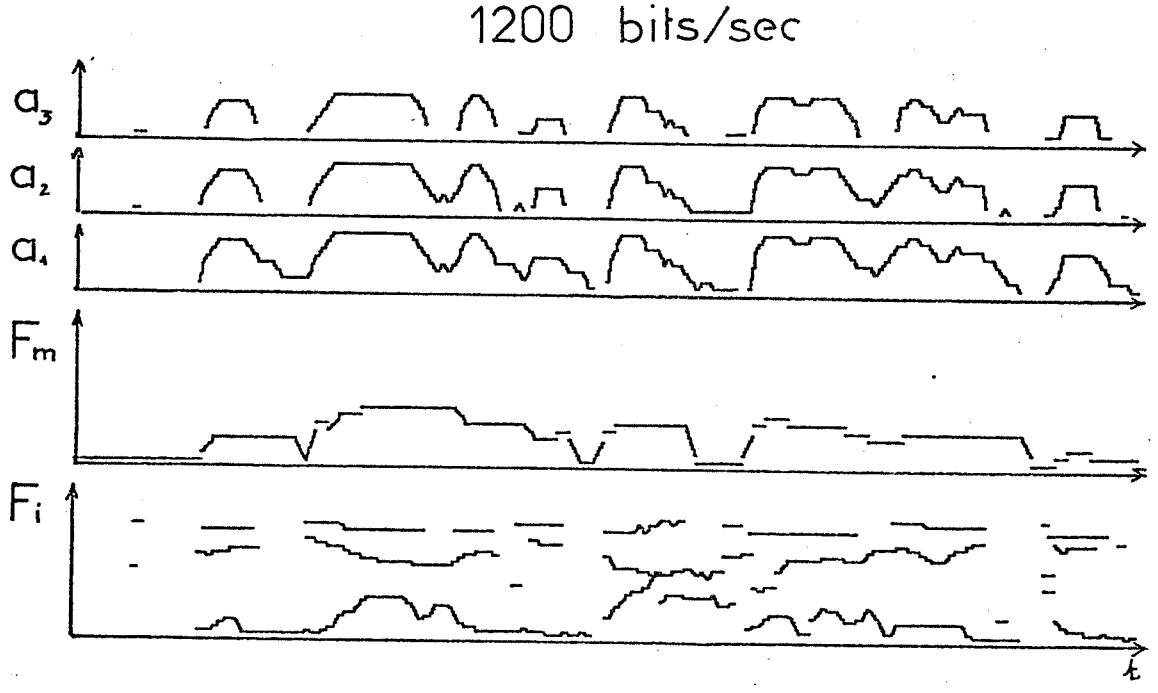


Fig10 b : Coordonnées des lignes de crête et fréquence de mélodie après codage à 1200 b/s et décodage

4 - CONCLUSION

L'analyse et la synthèse par lignes de crête confèrent à la parole synthétique obtenue un agrément certain ; sur le plan de l'intelligibilité, des mesures par tests de rime ont donné 95 % pour une analyse - synthèse sans compression à 1200 b/s. Une baisse d'intelligibilité d'environ 4 % a été décelée après adjonction de la compression. L'intérêt de la représentation par lignes de crête de la parole ne se limite pas à la transmission de celle-ci. La finesse de l'analyseur spectral du CIPHON permet d'envisager celui-ci comme outil de reconnaissance automatique, soit de mots dans un vocabulaire limité (commande vocale), soit de locuteurs (signature vocale).

D'autre part, les lignes de crête constituent un excellent moyen de stockage économique de messages (unités de réponse vocale).

BIBLIOGRAPHIE

1 : J.L. Flanagan : Focal Points is Speech Communication Research (IEEE Trans. Comm. Technol. Vol 19 n° 6 Déc. 1971)

2 : P. Alinat : Reconnaissance de phonèmes au moyen d'une cochlée artificielle (Thèse de Docteur-Ingénieur, 1973, Université de Nice).

3 : C. Dechaux : Design of Data Compression Vocoder (IEEE Intercon. 1972)

4 : P. Deman : La compression d'information à la Division Télécommunications - Revue Technique THOMSON-CSF Vol 7, n° 4, Déc. 1975

5 : P. Deman : La compression d'information téléphonique Colloque GRETSI, Nice - Juin 1975

6 : J.S. Bourgenot et C. Dechaux : Codage de la parole à faible débit : le vocoder CIPHON - Revue Technique THOMSON-CSF Vol 7, n° 4, Déc. 1975

7 : J. Potage : Extraction et codage à 1200 b/s des paramètres de synthèse par "lignes de crête" de la parole Revue Technique THOMSON-CSF Vol 7,n° 4, Déc. 1975

8 : J.F. Bellec et J. Pinel : Détermination des paramètres caractéristiques de la parole et codage à 600 b/s. Revue Technique THOMSON-CSF Vol 7, n° 4, Déc. 1975.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

APPLICATION A LA REPONSE VOCALE AU CONTROLE DE TACHES COMPLEXES : L'AIDE A LA CONSTRUCTION DIRECTE DE PROGRAMMES

par G. RENARD, D. TEIL, J.S. LIENARD, J. SAPALY

RESUME :

SUMMARY :

APPLICATION DE LA REPONSE VOCALE AU CONTROLE DE TACHES COMPLEXES : L'AIDE A LA CONSTRUCTION DIRECTE DE PROGRAMMES

par G. RENARD, D. TEIL, J.S. LIENARD, J. SAPALY

Les applications de la réponse vocale

Les applications qui viennent le plus immédiatement à l'esprit quand il s'agit de réponse vocale sont sans doute celles qui concernent l'interrogation de fichiers documentaires. Ce ne sont cependant pas les seules et il n'est même pas certain que ce soit les plus importantes.

Dans une première analyse, et à un moment où il semble encore prématuré de parler de la mise en oeuvre de la réponse vocale, on peut dire qu'il existe deux grands domaines d'applications suivant que :

- ou bien la parole s'avère indispensable à la transmission de l'information
- ou bien l'ouïe devient un sens privilégié de la communication.

C'est dans la première série d'applications que doit être rangée l'interrogation de fichiers documentaires. Il est bien certain en effet que si l'on veut étendre cette application au niveau de l'usager, le téléphone est le seul intermédiaire valable à ce jour et l'interface nécessaire entre la mémoire et la ligne téléphonique est une unité de réponse vocale.

Il faut cependant préciser les conditions de mise en oeuvre d'une telle application. Actuellement il est possible de dialoguer à distance avec un ordinateur dont la mémoire contient l'information recherchée, en connectant au combiné d'appel un système comprenant d'une part un clavier pour la question, d'autre part un écran de visualisation pour la réponse (et le contrôle de la question). Ce système coûte cher (entre 8 et 10.000 Francs) et sa diffusion semble limitée aux applications professionnelles. Si par contre on dispose d'un répondeur vocal comme interface entre l'ordinateur et la ligne téléphonique, le clavier suffit au niveau du combiné d'appel et la chute de prix est dans le rapport 10. Il est alors tout à fait pensable de mettre en oeuvre de tels systèmes dès maintenant.

Cependant il est facile d'imaginer que cette application ne s'étendra vraiment de manière importante que si le clavier peut lui-même être supprimé et si le combiné banalisé suffit. Ceci sera possible dès l'instant où une reconnaissance de parole, même élémentaire, pourra être mise en oeuvre au niveau de l'ordinateur. La reconnaissance d'une centaine de mots devrait suffire pour qu'un dialogue efficace puisse s'établir entre l'utilisateur et l'ordinateur.

L'autre série d'applications de l'unité de réponse vocale, celle où l'ouïe est utilisée comme sens privilégié de la communication, comporte de très nombreux types dont la mise en oeuvre parait plus immédiate que celle de l'interrogation de fichiers. Un premier exemple est la transmission des messages météorologiques à bord des avions en vol. Actuellement, les informations en provenance du monde entier arrivent par télex dans les divers organismes locaux, où des techniciens les

trient pour rédiger des messages régionaux qui sont enregistrés sur bande magnétique et envoyés par radio aux avions en vol. Or le tri peut très bien être confié à l'ordinateur, qui met en forme le message régional et le transmet par radio caractère par caractère. Ces caractères sont ensuite transformés en message parlé à bord de l'avion même par un répondeur vocal.

Un second type d'applications est l'aide à l'utilisateur, qui peut revêtir des formes diverses, mais qui concerne toujours l'emploi du sens auditif chez celui dont les autres sens sont déjà occupés. Ainsi dans le fonctionnement de certains systèmes complexes, diverses grandeurs sont analysées et surveillées et des alarmes sont déclenchées en cas de dépassements anormaux. Mais l'interprétation globale de plusieurs mesures peut également donner lieu à une alarme, et cette interprétation ne peut parfois être faite en temps réel qu'avec l'aide de l'ordinateur, qui doit alors éditer un message d'un niveau de compréhension élevé, que ne peut donner une lampe-témoin ou un cadran. La plupart du temps les yeux de l'opérateur sont occupés et le message vocal est la meilleure façon de l'avertir. Dans le cas du pilote de l'avion de chasse moderne, le message vocal semble même être le seul moyen possible de communication de ce genre.

Comme autre exemple important d'aide à l'utilisateur, il y a l'aide au contrôle de tâches complexes : non seulement l'opérateur recevra des alarmes, mais il aura confirmation de toutes ses actions au fur et à mesure qu'il les entreprendra, il pourra même avoir des refus d'exécution dont il sera immédiatement averti avec explication de la décision.

Il est évident que toutes ces applications nécessitent un répondeur vocal qui fonctionne en temps réel, ce que sait faire une unité basée sur l'emploi des diphonèmes.

Au L.I.M.S.I. (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur), laboratoire du C.N.R.S., nous avons mis en oeuvre une expérience utilisant ce type d'application : il s'agit de la construction directe en mémoire d'ordinateur d'un programme instruction par instruction au moyen de touches fonction. Un répondeur vocal indique au fur et à mesure ce qui est entré en mémoire ou édite éventuellement un message d'erreur. Nous allons donner quelques précisions sur cette expérience.

La construction directe de programmes

Actuellement la manière la plus courante d'entrer un programme en ordinateur consiste à perforer des cartes au moyen d'une multiperforatrice à clavier fonctionnant en "off line" et faire lire les cartes par l'ordinateur à l'aide d'un lecteur de cartes. Il est possible également, dans le cas de grosses installations utilisant le teleprocessing, d'écrire directement le programme en zone mémoire en mode conversationnel.

Ces divers procédés ont tous en commun le fait d'imposer l'emploi d'un clavier de touches de frappe, du type clavier de machine à écrire, ce qui n'est pas sans inconvénient. Il est nécessaire en effet de taper tous les caractères de chaque instruction, y compris les séparateurs tels que virgules et parenthèses, et de tenir compte des espacements : cela prend du temps et provoque un risque important d'erreurs syntaxiques sans détection possible avant compilation. On pourrait évidemment supprimer ces inconvénients en remplaçant les touches "caractères" du clavier par des touches "fonction" qui perforeraient

d'un coup des parties d'instructions, mais il faudrait alors autant de claviers qu'il existe de langages de programmation. Les multiperforatrices, d'autre part, procèdent à des opérations mécaniques sur les cartes (alimentation, éjection) à l'aide de commandes manuelles qui deviennent fastidieuses par leur répétition et qui accroissent le temps global de perforation. Il y a enfin, commun à tous les claviers, le désagréable inconvénient du bruit.

En vue de répondre à ce problème, le L.I.M.S.I. a rassemblé un certain matériel autour d'un ordinateur IBM 7 pour entrer directement un programme en mémoire à l'aide de touches "fonction" de type graphique. Ce matériel, qui constitue en quelque sorte un Terminal de construction directe de programmes, comprend : un écran de visualisation Tektronix, une tablette d'entrées graphiques STRAND et son convertisseur A/D, une unité de réponse vocale ICOPHONE, et un transcontrôleur de connexion au Système 7.

Le STRAND est une tablette de verre dont la surface a été recouverte d'un dépôt mince conducteur qui n'altère pratiquement pas sa transparence. Des électrodes sont placées sur les bords et alimentées de telle façon qu'une sonde conductrice placée en un point quelconque de la tablette prenne successivement et de manière répétée des tensions proportionnelles respectivement à l'abscisse et à l'ordonnée du point. Comme la commutation d'alimentation en x et y est suffisamment rapide, il est possible de relever, de convertir et de mémoriser les coordonnées d'une suite de points correspondant à un échantillonnage donné de toute ligne tracée de manière continue par la sonde sur la tablette.

Dans l'application présente de construction de programmes, le STRAND n'est pas utilisé en entrée de courbes, mais en grille de touches "fonction". On vient disposer sous la tablette de verre une feuille sur laquelle est reproduite une grille de cases carrées, chaque case étant affectée d'une fonction. En touchant avec la sonde un point intérieur d'une case, on enregistre le couple de coordonnées du centre de la case, qui constitue l'adresse d'une fonction située dans un fichier en mémoire d'ordinateur. Un programme analyse cette fonction et exécute l'ordre correspondant.

Il y a deux types de fonctions : celles de construction d'instruction et celles de traitement et d'entrées-sorties. Les premières remplissent la majorité des cases et comprennent :

- la désignation du langage choisi
- les caractères alphanumériques et les caractères spéciaux
- les caractères arithmétiques et les séparateurs
- les mots-clef de déclaration (dimension, type, etc)
- les mots-clef d'exécution (branchement, comparaison, boucle, etc)
- les mots-clef de contrôle (compilateur, fichier, etc)
- les fonctions de bibliothèque usuelles (sinus, logarithme, etc)
- la validation d'instruction.

Parmi les fonctions de traitement et d'entrées-sorties, il y a :

- la visualisation d'une séquence mémorisée
- son impression
- sa perforation sur cartes
- la lecture d'un paquet de cartes (en vue de modification)
- l'insertion ou l'effacement d'une instruction ou d'une partie d'instruction
- la compilation d'un programme
- son exécution.

L'intérêt de l'entrée par tablette graphique réside non seulement dans le fait qu'on peut construire tout ou partie d'instruction en ne touchant qu'une seule case avec la sonde et qu'on évite ainsi la plupart des fautes d'orthographe, mais aussi dans le fait qu'il est facile à l'utilisateur de changer de composition de grille et de grandeur de cases, en fonction de ses besoins ou de ses goûts. D'autre part un certain nombre de cases non affectées sont toujours disponibles pour permettre au programmeur de créer des identificateurs et de pouvoir les introduire d'un seul coup et non caractère par caractère dans son programme.

L'intérêt de la réponse vocale

Examinons quel est le déroulement de la procédure, tout d'abord dans le cas où il n'y a pas d'unité de réponse vocale. L'utilisateur qui veut construire un programme n'a pas besoin de l'avoir écrit préalablement puisqu'il peut faire d'emblée cette opération sur le Terminal. Il lui suffit d'avoir fait un bon organigramme.

Il dispose sous la tablette la grille correspondant au langage qu'il veut utiliser, et avec le crayon-sonde commence à construire son programme en touchant les cases appropriées de la grille, la première étant celle qui indique à l'ordinateur de quel langage il s'agit, donc quel est le programme de construction qu'il faut appeler. A chaque fois qu'il touche une case, il voit s'inscrire sur l'écran le contenu de celle-ci et s'assure ainsi qu'il a touché la bonne case. S'il commet une erreur syntaxique au niveau de l'instruction, la case qui vient d'être touchée n'est pas validée, et au lieu de son contenu c'est un message d'erreur qui apparaît sur l'écran. Il peut ainsi toucher la bonne case et poursuivre la construction de son instruction. A la fin de celle-ci, il touche la case "validation d'instruction" (ce qui correspond à une nouvelle image-carte en Fortran et à un point virgule en PL1).

Les erreurs syntaxiques ne sont pas les seules qui peuvent être faites dans la construction d'un programme, mais ici ce sont les seules qui déclenchent le refus de validation de la touche concernée. C'est pour cela qu'ont été prévues des cases de correction (effacement, rajout, insertion), qui permettent de corriger toutes les erreurs non syntaxiques et même certaines erreurs syntaxiques dont la détection n'est pas immédiate (par exemple une insuffisance ou un excédent de parenthèses).

Si l'on observe bien le déroulement de la procédure, on constate vite un fait : malgré la rapidité qu'apporte le système par rapport aux processus classiques, une perte de temps manifeste apparaît qui vient de ce que l'opérateur porte son regard alternativement sur la tablette pour toucher la case et sur l'écran pour constater qu'il a touché la bonne case ou qu'au contraire il a commis une erreur syntaxique, et ceci dans un va-et-vient incessant. C'est la constatation de ce fait qui nous a incité, au L.I.M.S.I., à introduire dans le système l'une des unités de réponse vocale Icophones que nous avions réalisées précédemment.

Le changement a été alors considérable à tous égards : non seulement sur le plan de la vitesse d'exécution (on estime avoir obtenu au moins le facteur 2), mais également sur le plan de la clarté puisque les messages d'erreur ne sont plus édités qu'à l'Icophone et ne viennent plus encombrer l'écran.

Trois qualités de l'Icophone sont ici mises en évidence parce qu'elles s'avèrent nécessaires : son intelligibilité (même si c'est au détriment de l'esthétique), sa réponse en temps réel (le temps entre l'instant où la case est touchée et celui où le message commence n'est pas décelable par l'oreille), et son vocabulaire illimité (n'importe quel langage de programmation doit pouvoir être mis en oeuvre, n'importe quel identificateur doit pouvoir être créé à tout moment, n'importe quel commentaire ou message doit pouvoir être dit suivant les besoins et les goûts de chaque utilisateur).

Les applications du Terminal

Les applications d'un tel Terminal de construction de programmes sont de deux sortes : la programmation dans les Centres de Calcul ou les Centres reliés à une grosse unité de traitement en Teleprocessing, et l'apprentissage de la programmation. L'une et l'autre doivent mettre en oeuvre plusieurs terminaux fonctionnant en simultanéité apparente, ou plus exactement un terminal muni de plusieurs stations. Chaque terminal serait donc constitué d'un organe central de type miniordinateur avec ses périphériques classiques, et d'un certain nombre de stations comprenant chacune une tablette d'entrées graphiques, un écran de visualisation et un synthétiseur de parole muni d'un casque d'écoute. L'unité de réponse vocale serait donc en fait une unité multisynthèse dont le processeur unique serait intégré au miniordinateur central et qui aurait un synthétiseur par station, la simultanéité apparente d'édition de différents messages étant possible pour plus d'une trentaine d'entre eux.

Dans l'application d'apprentissage de la programmation, les messages d'erreurs dits par l'Icophone peuvent évidemment être plus ou moins explicites suivant le niveau de l'élève : au début de l'apprentissage les messages seront clairs, du type : "Erreur ; vous ne pouvez mettre qu'une virgule, ou une parenthèse ou fin d'instruction ; revoyez le paragraphe N de votre cours". Un peu plus tard, les messages seront plus laconiques et signaleront seulement qu'il y a une erreur, avec l'éventualité d'un message plus complet au bout de trois erreurs successives au même endroit.

Conclusion

La mise en oeuvre et l'emploi de la réponse vocale dans le Terminal de construction de programmes a mis en évidence de manière très nette l'intérêt que peut présenter l'utilisation de l'ouïe dans l'aide à l'exécution de tâches complexes. Un champ très vaste d'applications s'offre de ce fait à la synthèse de la parole, encore insoupçonné il y a peu de temps.

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 au 21 mai 1976

GLOTTOMETRIE EN TEMPS REEL-APPLICATIONS

par

P. ANDRE - M. FILLEAU - F. HAMELIN

(Laboratoire d'Automatique de Besançon)

RESUME : Les auteurs décrivent une chaîne d'instrumentation réalisée pour mesurer et visualiser en temps réel la fréquence instantanée de vibration des cordes vocales;ils présentent ensuite quelques applications à l'analyse et à la synthèse du signal vocal.

SUMMARY : The authors describe an instrumentation system performing real time measurement and display of the vocal cords vibration instantaneous frequency;then they present some applications to the analysis and synthesis of the vocal signal.

GLOTTOMETRIE EN TEMPS REEL - APPLICATIONS

par

P. ANDRE - M. FILLEAU - F. HAMELIN

(Laboratoire d'Automatique de Besançon)

1) GENERALITES

On connaît bien aujourd'hui l'importance du facteur "hauteur" dans la parole ; de nombreuses études (mélodie, micromélodie, etc...) nécessitant l'enregistrement fidèle de son évolution temporelle, nous nous sommes efforcés de développer une instrumentation appropriée.

A priori le fondamental de la voix peut être atteint soit indirectement à travers le signal de parole capté par un micro, soit directement, par étude du mouvement des cordes vocales. Dans le premier cas les nombreuses fréquences"parasites" introduites par les résonances du canal vocal et les transitoires engendrés par les consonnes sonores rendent cette extraction malaisée, peu fiable, voire impossible en temps réel (méthodes analogiques ou numériques).

Ces difficultés disparaissent lors de l'analyse directe du signal glottique ; différentes méthodes sont envisageables :

- les méthodes radiographiques ou optiques [1], [2] (radiocinématographie ou cinématographie rapides, laryngostroboscopie). Elles sont en général coûteuses, délicates d'emploi et, nécessitant un dépouillement, interdisent toute exploitation en temps réel;

- la méthode électro-glottographique [3] que nous avons retenue consiste à détecter les variations d'impédance du larynx et, contrairement aux précédentes, fournit instantanément et en continu un signal représentant le mouvement des cordes vocales, ceci sans risque ni gêne pour le locuteur.

Concernant la mesure de fréquence de ce signal les méthodes usuelles d'analyse spectrale ou par auto-corrélation nécessitent un matériel évolué donc onéreux, et sont par ailleurs peu aptes à rendre compte des variations brutales de hauteur. Nous avons opté pour une mesure directe de la durée de chaque cycle glottique, mesure qui fournit en temps réel la valeur de la fréquence instantanée du fondamental.

La chaîne glottométrique que nous présentons se compose donc de deux éléments : l'électro-glottographe et l'analyseur de mélodie.

2) ELECTRO-GLOTTOGRAPHE

2.1. Principe

Choisi en raison de sa simplicité de mise en oeuvre, l'électroglottographe fournit un signal image des cycles glottiques

selon le principe de fonctionnement suivant, schématisé en figure 1.

Deux électrodes métalliques "encadrent" le cartilage thyroïde au niveau de la glotte et sont alimentées par un générateur de courant haute fréquence. Ce courant traverse le larynx et crée entre les électrodes une tension H.F. proportionnelle à l'impédance de ce dernier. Celle-ci variant au cours de l'émission de sons voisés, la tension inter-électrode se trouve modulée en amplitude par le mouvement des cordes vocales. Le glottogramme est le signal basse fréquence obtenu après démodulation et filtrage.

2.2. Variations de l'impédance du larynx

Les études expériemntales de Philipson et Fricke d'une part, de Duvelleroy [4] d'autre part ont permis de définir un modèle qui représente assez bien le comportement électrique du larynx lorsqu'il est traversé par un courant alternatif de fréquence comprise entre quelques dizaines et quelques centaines de kilohertz (Fig.2).

Dans ce schéma les éléments susceptibles de varier au cours de la phonation sont la capacité C, dont l'effet est presque annulé par la faible valeur de R, et la résistance glottique série r qui diminue lors de l'ouverture de la glotte.

Ces variations ont deux causes essentielles :

- les déformations lentes (f < 10 Hz) et de grande amplitude des tissus entourant le larynx (mouvements d'ensemble de celui-ci) ;

- la modification de la configuration interne du larynx due au mouvement des cordes vocales et beaucoup plus rapide (f > 50 Hz), qui engendre une variation relative d'impédance de l'ordre de 0,1 %.

La différence de vitesse entre les deux types de variations permet de les dissocier assez facilement à l'aide de filtres.

2.3. Réalisation

L'électroglottographe dérivé de l'appareil original de Fabre [3], est composé de trois circuits élémentaires (Fig.1).

2.3.1. Le générateur-modulateur

Un oscillateur fournit une tension à une fréquence voisine de 400 KHz. Celle-ci est convertie en courant au moyen d'une résistance R_1 de valeur très supérieure au module de l'impédance Z du larynx. Ce montage, voisin du montage potentiométrique, a été préféré au montage en pont malgré la meilleure sensibilité de ce dernier ; il ne requiert en effet qu'un seul règlage, celui de R_1, pour s'adapter à la forte dispersion de la valeur de Z selon les individus . Au contraire le

montage en pont impose,pour un locuteur donné, de fréquents réajustements dus à la relative instabilité de Z, ceux-ci étant par surcroît longs et délicats en raison des deux paramètres, une résistance et une capacité qui doivent être règlés simultanément. Le pont doit de plus être légèrement déséquilibré au repos (ce qui réduit sa sensibilité) pour éviter le défaut dit de doublement de fréquence.

2.3.2. Le démodulateur

Ce dispositif met en oeuvre un élément non linéaire à seuil variable. Celui-ci suit les variations lentes mais importantes (quelques %) de l'amplitude de la porteuse que nous voulons éliminer. On peut ainsi extraire les variations rapides mais de faible valeur correspondant au glottogramme.

2.3.3. Le banc de filtres

Il est constitué d'une cascade de deux filtres. Le premier a une bande qui s'étend de 16 Hz à 1,6 KHz et fournit un signal respectant la forme du signal glottique sans toutefois permettre une mesure précise de fréquence. Le second à bande plus étroite (60 Hz - 600 Hz) fournit à l'analyseur de mélodie un signal "propre" c'est-à-dire débarassé du bruit et des résidus à très basse fréquence.

2.4. Utilisation

Après avoir enduit la peau de pâte conductrice on met en place le collier porte-électrodes. On règle ensuite la résistance R_1 ("Règlage H.F.") de sorte qu'en l'absence de phonation l'aiguille d'un indicateur de niveau se trouve dans une zone repérée.

L'appareil est alors prêt à fonctionner et fournit :

- le signal de seuil représentant les mouvements lents du larynx et du cou ;

- le glottogramme proprement dit (après le premier filtre) pour visualisation directe (oscilloscope, enregistreur rapide ...) (Fig. 3a) ;

- le signal filtré destiné à attaquer l'analyseur de mélodie (Fig.3b).

3) ANALYSEUR DE MELODIE

3.1. Spécifications

L'obtention en temps réel du graphe exact des variations de hauteur de la voix en fonction du temps impose à l'analyseur trois caractéristiques fondamentales :

- aptitude à la restitution des variations brusques et de grande amplitude de la hauteur (mesure de la durée de chaque cycle glottique) ;

- précision de la mesure, ici 2 %, ce qui correspond pratiquement à la sensibilité d'une oreille moyenne ;

- fiabilité,ce qui nous a conduit à utiliser un traitement digital et à l'emploi de circuits intégrés.

La plage de mesure couvre une bande de fréquence admise pour la parole normale de 60 à 600 Hz.

En raison du caractère "musical" de la mélodie, nous nous sommes imposé de fournir les résultats selon une échelle logarithmique, la mesure se présentant comme un "intervalle" de fréquence exprimé en octaves tons et quarts de ton ; cette disposition permet de comparer aisément les variations relatives de hauteur chez des locuteurs très différents (hommes, femmes, enfants, ...).

La sortie est disponible à la fois sans forme analogique, pour visualisation immédiate, et numérique pour traitement différé

3.2. Principe

Le principe de l'analyseur [5] - [6] consiste à déterminer la fréquence F du signal glottique par rapport à une fréquence de référence F_o (600 Hz).

La plage de mesure d'une décade est quantifiée en 80 intervalles élémentaires d'un quart de ton délimités par les fréquences F_o, F_1 ... F_k ... F_{80}, chacun de ceux-ci correspondant à une variation relative constante entre deux fréquences discrètes consécutives, égale à $a = 2^{1/24}$.

On peut écrire :

$$F_k = a^{-k} F_o \quad ,$$

d'où :

$$k = - \mathrm{Log}_a \frac{F_k}{F_o} \quad ;$$

on dit alors que F_k se situe à k quarts de ton en dessous de F_o.

En passant aux périodes :

$$T_k = a^k T_o \quad .$$

La mesure consiste à déterminer le rang k de la durée T_k immédiatement inférieure à la période T du signal glottique (Fig.4). Elle est réalisée au moyen du dispositif représenté (Fig.5).

Les différentes durées possibles T_p sont approchées par des valeurs T_p^* choisies parmi les dates fournies par une horloge primaire. Chaque fois que le temps écoulé depuis le début de la période considérée du signal passe par l'une des valeurs T_p^* une impulsion I_p est émise par un générateur approprié. Ces impulsions I_p sont totalisées dans un compteur dont le contenu en fin de période représente la mesure.

3.3. Réalisation

La précision imposée conduit à choisir une horloge primaire de fréquence voisine de 128 F_o.

Nous prendrons donc pour valeur approchée de T_p : $T_p^* = (128\ a^p)^* \frac{T_o}{128}$ où $(128\ a^p)^*$ représente la meilleure approximation entière de $128\ a^p$.

Dans ces conditions, l'impulsion I_p apparaîtra lors de la coïncidence entre les valeurs de $(128\ a^p)^*$ stockées dans une table séquentielle S et le contenu d'un compteur C dénombrant les impulsions d'horloge primaire émises depuis le début de la période T.

En remarquant que $T_{p+24} = 2T_p$ on est conduit à réaliser un générateur ne couvrant qu'une octave, le passage d'une octave à la suivante s'effectuant en doublant la période du signal d'horloge au moyen d'un diviseur de fréquence programmé.

Une nouvelle simplification apparaît lorsqu'on constate que les accroissements successifs (arrondis à la valeur entière la plus proche) Δ_p de $128\ a^p$ ne prennent dans toute l'octave que 4 valeurs distinctes (4, 5, 6, 7), chacune n'apparaissant jamais plus de 7 fois. On peut donc n'inscrire dans S que ces quatre chiffres codés sur 3 bits (celui de fort poids étant toujours à 1) ; les états internes correspondant à des valeurs constantes de Δ_p seront différenciés à l'aide de trois variables secondaires.

Ce mode de travail incrémental de S implique un fonctionnement différent du compteur C qui, après une première évolution jusqu'à 128 représentant la période T_o, doit être remis à zéro puis compter jusqu'à égalité entre son contenu et celui de S. Chaque coïncidence entraîne l'émission d'une impulsion I_p, la remise à zéro de C et l'incrémentation d'un pas de S pour disposer de la nouvelle valeur de Δ_p.

On trouve dans le schéma général de l'analyseur donné en figure 6 :

- la table séquentielle S

- le compteur C

- le comparateur qui détecte la coïncidence entre les contenus de S et C et délivre l'impulsion I_p

- le diviseur programmable commandé par le signal de fin de cycle de S (changement d'octave)

- le compteur d'intervalles élémentaires scindé en trois parties pour les quarts de ton, les tons et les octaves

- un circuit de mise en forme qui délivre un signal carré synchrone du glottogramme injecté à l'entrée ; les fronts montants de ce signal matérialisent les limites des cycles glottiques et initialisent l'ensemble de mesure.

Aucun règlage n'est nécessaire avant utilisation.

4) APPLICATIONS

Cette chaîne glottométrique trouve des applications dans divers domaines :

4.1. Etude de la mélodie de la parole concernant :

- l'intonation significative : affirmation interrogation, etc... [7] ; la figure 7 présente un exemple de schéma intonatif relevé au moyen de notre appareillage ;

- l'intonation particulière (accents locaux, nationaux ...) ;

- les langues à tons pour lesquelles des variations rapides quantifiées de hauteur ont une valeur distinctive ;

- la micromélodie, des phonèmes vocaliques.

4.2. Synthèse de la parole

La suite des valeurs de la fréquence, préalablement enregistrée dans la mémoire d'un ordinateur peut être utilisée pour commander la partie synthétiseur d'un vocoder, afin d'apprécier notamment le "naturel" de la parole ainsi obtenue, par comparaison avec l'utilisation d'une source glottique de fréquence fixe. En plus de cette expérience en cours envisageons également dans le même but, l'usage du glottogramme comme signal d'excitation d'un modèle analogique du canal vocal [8] .

4.3. Fonctionnement inverse : synthétiseur de mélodie

On peut transformer l'analyseur en générateur programmable par comparaison de la sortie numérique à la valeur désirée de la fréquence (founie, par exemple, par un ordinateur) ; la sortie est alors constituée par le signal de coïncidence qui se substitue au signal glottique pour déclencher l'analyseur.

Nous avons réalisé un tel générateur [9] avec une résolution du huitième de ton et une précision de l'ordre du savart ; il peut être utilisé :

- comme source glottique programmable pour divers types de synthétiseurs de parole ;

- comme référence pour l'accord d'instruments de musique ;

- comme synthétiseur de mélodies musicales (la plage de fréquence étant alors étendue de 16 Hz à 16 KHz).

5) CONCLUSION

Les deux éléments de cette chaîne sont d'emploi simple, de bonne fiabilité, et d'un prix de revient modéré.

De ce fait et compte tenu de leur champ d'applications, nous pensons qu'ils devraient devenir d'un emploi courant dans les laboratoires s'intéressant à la Communication parlée.

B I B L I O G R A P H I E

[1] FARNSWORTH (D.W)
High speed motion pictures of the human vocal cords
Bull Laboratory Records, t.18, p.203-208 (1940)

[2] FROKJAER - JENSEN (B)
A photoelectric glottograph
A.R.I.P.U.C. (Copenhague), n° 2, p.5-19, (1967)

[3] FABRE (P.)
Un procédé percutané d'inscription de l'accollement glottique au cours de la phonation : glottographie en H.F., premiers résultats
Bulletin de l'Académie Nationale de Médecine, p.66-69(1957)

[4] DUVELLEROY (M.)
La glottographie en module et en phase, moyen d'étude des schémas électriques équivalents au larynx
Thèse de médecine - Paris (1961)

[5] FILLEAU (M.) - LHOTE (F.)
Mesure logarithmique de fréquence instantanée par un dispositif digital
Revue d'Acoustique, n° 27, p.231-234 (1973)

[6] FILLEAU (M.)
Dispositif d'analyse en temps réel de la mélodie de la parole
Thèse de Docteur-ingénieur - Université de Besançon (1975)

[7] LHOTE (E.) - FILLEAU (M.)
Reconnaissance de patrons intonatifs
Journées d'étude sur la parole du G.A.L.F.- Toulouse -
(Mai 1975)

[8] CHEVILLARD (A.)
Contribution à l'étude analogique de l'appareil phonatoire
Thèse de Dr Ing. - Université de Besançon (1973)

[9] ANDRE (P.) - LHOTE (F.)- FILLEAU (M.)
Générateur digital de fréquences à affichage logarithmique
Congrès Franco-allemand de Chronométrie - Strasbourg(mai 1975)

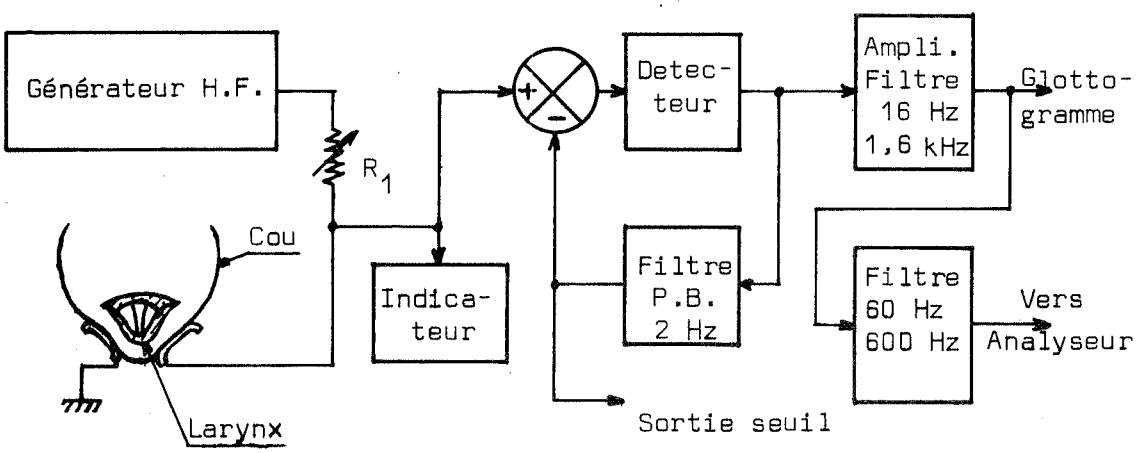


Fig.1 : SCHEMA DE PRINCIPE DU GLOTTOGRAPHE

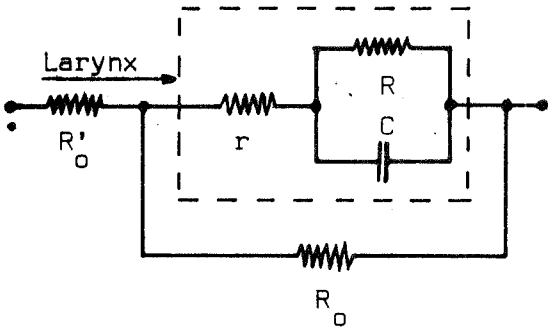
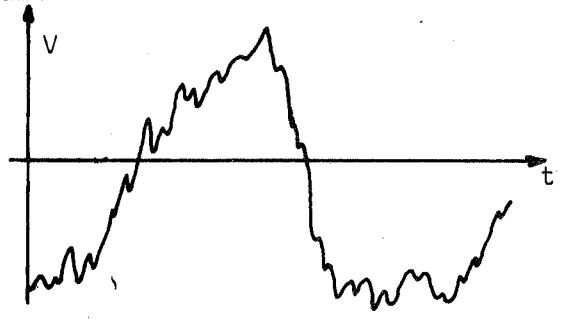
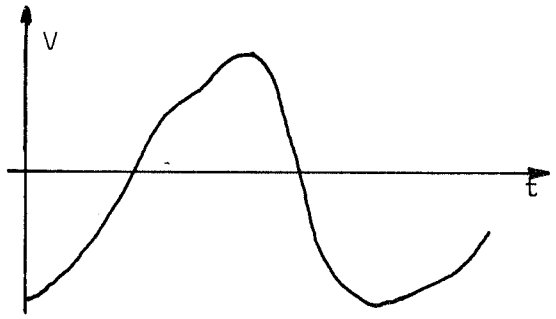


Fig.2:IMPEDANCE INTER-ELECTRODES



a - Après le 1e Filtre



b - Après le 2e Filtre

Fig.3: FORME DES SIGNAUX DE SORTIE

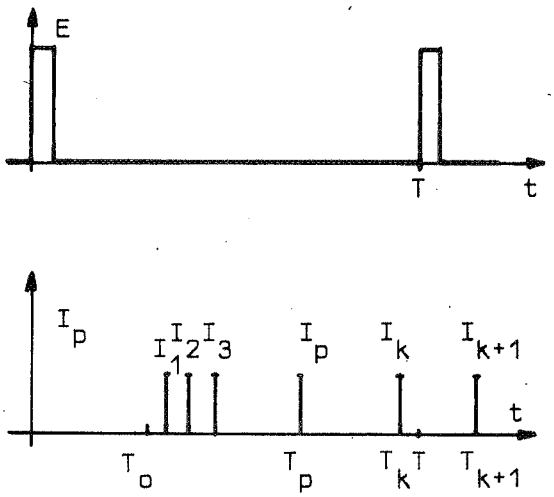


Fig.4: PRINCIPE DE LA MESURE DE FREQUENCE

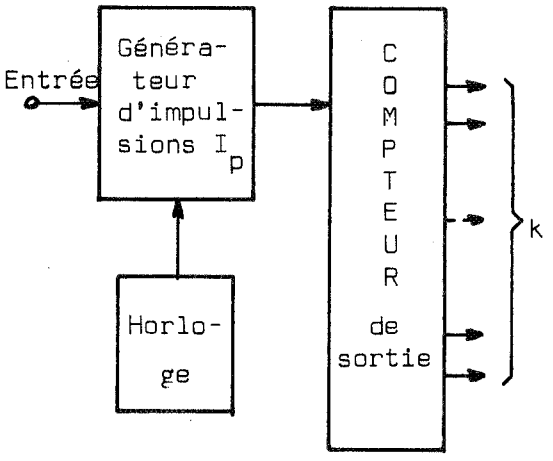


Fig.5: PRINCIPE DE L'ANALYSEUR

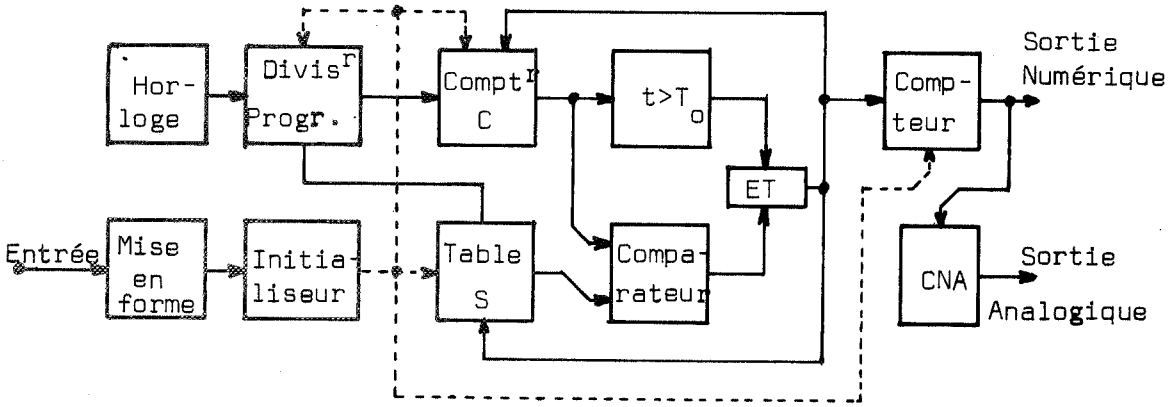
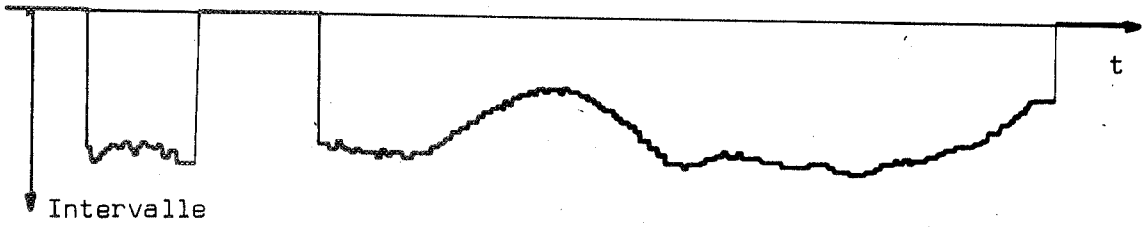
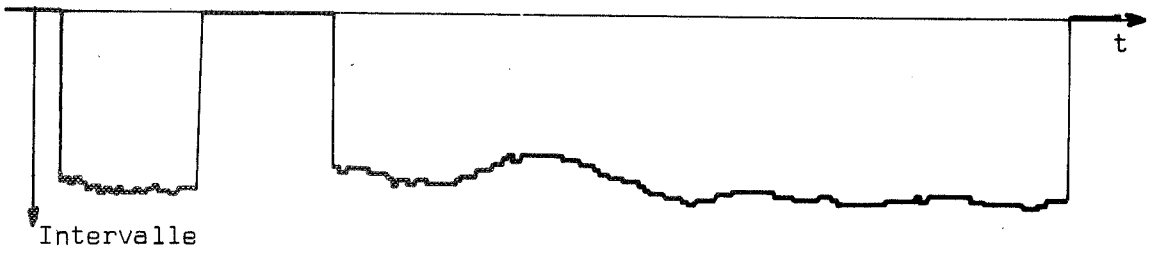


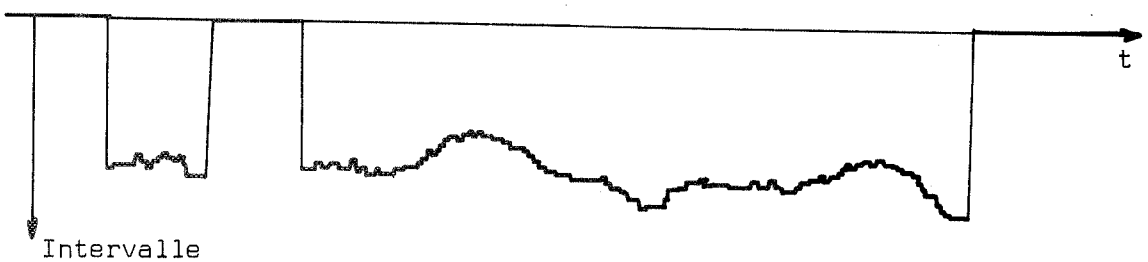
Fig.6: SCHEMA SYNOPTIQUE DE L'ANALYSEUR



a - Forme interrogative



b - Forme affirmative



c - Affirmation péremptoire

Fig.7: EXEMPLES DE PATRONS INTONATIFS SUR "IL FAIT BEAU AUJOURD'HUI"

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 au 21 mai 1976

ANALYSE DU SIGNAL VOCAL

UTILISATION DES EXTREMA DU SIGNAL ET DE LEURS AMPLITUDES

DETECTION DU FONDAMENTAL ET RECHERCHE DES FORMANTS

Marc BAUDRY et Benoît DUPEYRAT

Services d'Electronique de Saclay

RESUME :

Le signal vocal est étudié dans sa représentation amplitude -temps en codant les intervalles entrè extrema et les amplitudes de ces extrema.

Ces paramètres permettent :

- de détecter le début de chaque cycle de voisement après segmentation des phonèmes,
- d'analyser la composition spectrale d'un cycle de voisement.

SUMMARY :

The voice signal is studied in its amplitude-time representation by coding the intervals between extrema and the amplitudes of these extrema.

With these parameters, we can :

- detect the beginning of each pitch cycle after segmentation of of the phonemes,
- analyse the spectral composition of a pitch cycle.

ANALYSE DU SIGNAL VOCAL - UTILISATION DES EXTREMA DU SIGNAL ET DE LEURS AMPLITUDES - DETECTION DU FONDAMENTAL ET RECHERCHE DES FORMANTS

Marc BAUDRY et Benoit DUPEYRAT des Services d'Electronique de Saclay

I. LE SYSTEME D'ANALYSE DU SYSTEME VOCAL

Le centre de ce système est constitué par un miniordinateur MULTI-20-06 d'INTERTECHNIQUE (32K octets). Nous disposons d'un périphérique qui stocke en mémoire centrale le résultat du codage de la parole dans sa représentation amplitude-temps. L'entrée microphone est utilisée directement dans l'ambiance bruitée de la salle machine. La parole codée peut-être stockée sur disque magnétique, elle peut également être affichée sur un écran de visualisation. De plus, les programmes d'analyse sont dotés de sorties graphiques qui permettent une interprétation visuelle rapide des résultats. Nous pouvons ainsi analyser une variété de signaux de parole aussi grande que nécessaire.

Cet ensemble a permis la réalisation d'un premier système de reconnaissance de la parole [1][2] qui montre la possibilité d'analyser, en temps réel, le signal vocal à partir de sa représentation amplitude-temps.

Nous rappelons que nous codons les intervalles entre les extrema du signal ce qui permet de diminuer le nombre d'échantillons. De plus il est fondamental de noter que nous codons également les amplitudes de ces extrema [3] ce qui permet d'avoir des informations sur toutes les composantes fréquentielles du signal (fondamental et formants) et sur leurs énergies relatives.

La fréquence d'échantillonage est de 10Khz et la digitalisation se fait sur 6 bits.

Définissons les paramètres utilisés (figure 1). Soient e_{i-1}, e_i la suite des extrema du signal se produisant aux temps t_{i-1}, t_i...

On appelle s_i le segment joignant les deux extrema e_{i-1} et e_i

On appelle durée l_i le temps $t_i - t_{i-1}$ séparant ces extrema. On appelle amplitude a_i l'amplitude de l'extremum e_i.

L'exploitation plus approfondie du codage des amplitudes dans cette nouvelle étude apporte d'importantes améliorations dans les opérations de segmentation du signal et de préclassification.

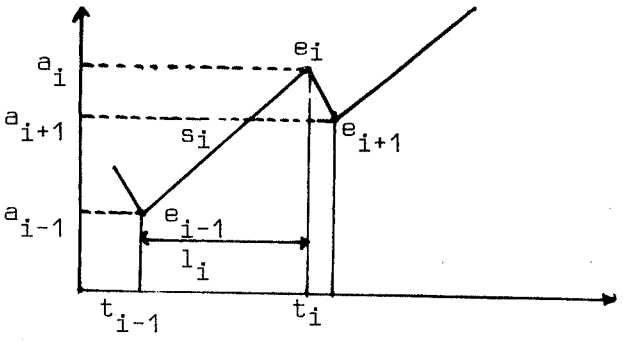


FIGURE 1

Ainsi les programmes en cours de développement permettront :
- d'isoler les zones du signal vocal,
- de déterminer le caractère voisé ou non du signal vocal,
- de détecter précisement le début de chaque cycle de voisement,
- de segmenter un mot en phonèmes,
- d'analyser la composition spectrale d'un cycle de voisement.

II. ANALYSE DU SIGNAL

1. Détection des zones de signal vocal

- Nous connaissons l'amplitude moyenne du bruit de fond a_b. Lorsqu'un extremum e_i a une amplitude supérieure au bruit de fond soit $a_i < a_b$ il peut s'agir d'un bruit parasite ou du début d'un signal vocal. Nous décidons qu'il s'agit d'un signal vocal si pendant une durée de 50ms il n'y a pas une suite de e_i tels que $a_i < a_b$ et $\Sigma l_i > 10$ms.

Cette fenêtre de 50 ms est une limite inférieure de la durée d'un signal vocal isolé.

.../...

- La fin d'un son est détectée lorsqu'une suite d'extrema e_i est telle que $a_i<a_b$ et $\Sigma l_i>500$ ms.

- De même les zones de silence des plosives non voisées (p,t,k) sont détectées par la présense d'une suite de e_i tels que $a_i<a_b$ et 30 ms$<\Sigma l_i<$500 ms.

- Un silence plus court est un affaiblissement du signal vocal lors d'une transition, par exemple s - voyelle.

Les amplitudes a_i peuvent varier de -32 à +31 mais en l'absence de signal vocal elles varient de -1 à +1.

La valeur moyenne du signal est définie par $\frac{\Sigma |a_i| \, l_i}{\Sigma l_i}$

Pour une voyelle cela donne des valeurs de l'ordre de 20 à 30. Prenons une valeur de 1 pour le bruit de fond nous avons un rapport signal sur bruit de l'ordre de 30 dB.

2. Détection de la présence du fondamental ou non

- Pour avoir une idée des composantes fréquentielles du signal vocal nous construisons l'histogramme H1 des longueurs l_i en portant dans chaque classe C_n une valeur proportionnelle à Σl_n. C'est-à-dire que la classe C_n est proportionnelle au temps passé avec des longueurs l_n. Remarquons de plus que la surface de l'histogramme est égale à la durée de la fenêtre d'analyse.

L'histogramme H1 utilisant les longueurs l_i est donc construit à partir des distances entre passages par zéro de la dérivée du signal.

Nous allons de la même façon construire l'histogramme H2 des distances entre passages par zéro du signal direct.

- Les seuls sons non-voisés sont les fricatives s, ʃ, f et les plosives p, t, k.

Le cas des plosives non voisées est particulier, il est simplement déterminé par une absence de signal vocal (voir §.1).

Les fricatives ont des caractéristiques spéctrales bien particulières : elles présentent un signal de bruit dans des bandes de fréquences variant entre 2000 et 10000 Hz suivant le cas. Ceci se traduit par des pics importants dans les classes 1, 2 et 3 de H1 (figure 2b)

- Ce critère est suffisant pour éliminer la plupart des sons voisés, examinons les différents cas possibles :

CAS DES VOYELLES ET DES LIQUIDES l, r.

* Rappelons tout d'abord que les voyelles sont produites par un signal d'excitation qui engendre un signal de résonance que l'on peut, en première approximation, assimiler à une composition de sinusoïdes amorties. De ce fait, la composante donnant le deuxième formant peut disparaître au cours du cycle de voisement et ne plus apporter de contribution à H1. (Le troisième formant apparaît très peu car son amplitude est faible par rapport à nos 64 niveaux de digitalisation).

* Deux cas se présentent suivant la fréquence du deuxième formant :

- Si elle est inférieure à 2000 Hz, H1 ne présente pas de pics dans les classes 1, 2, 3 (fig. 3b)(Cas des sons a, ã, ɔ, ɛ̃, ɔ̃, u).

- Si elle est supérieure à 2000 Hz, le résultat dépend de l'amplitude de cette composante du signal. (Cas des sons i, e, ɛ, y). Si cette composante disparaît au cours du cycle, H1 laissera apparaître des fréquences inférieures à 2000 Hz et nous pourrons dire que le son est voisé. Mais quelque fois elle est suffisamment importante pour que H1 ressemble à celui d'une fricative non voisée (fig. 4b)

CAS DES PLOSIVES VOISEES, b, d, g ET DES NASALES m, n.

Le fondamental étant prédominant, il n'y a pas de confusions possibles.

.../...

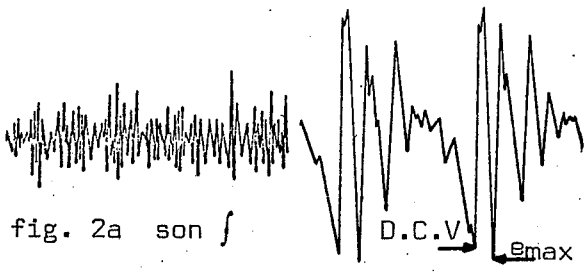


fig. 2a son ʃ

fig. 3a son a

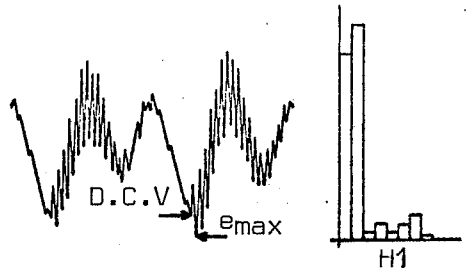


fig. 4a son i

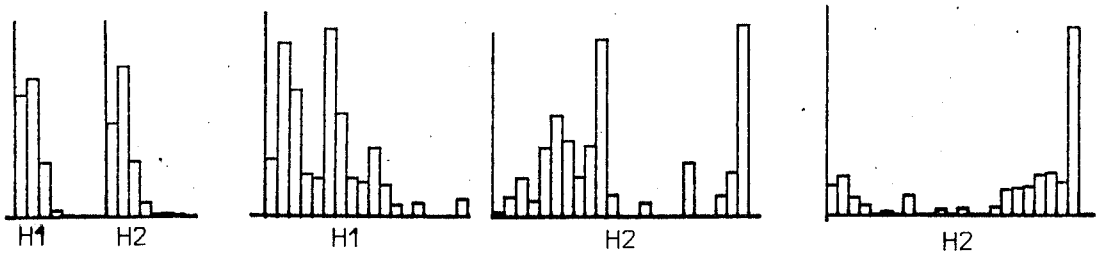


fig. 2b son ʃ

fig. 3b son a

fig. 4b son i

CAS DES FRICATIVES VOISEES ʒ, z, v.

L'amplitude du bruit de friction diminue du ʒ au z, et du z au v. Aussi le fondamental apporte souvent une contribution dans H1 pour z et v et permet de décider du caractère voisé. Mais, surtout pour ʒ, H1 est souvent semblable à celui des fricatives non voisées.

Traitement des confusions

Pour les sons voisés mal classés nous allons utiliser le fait que le deuxième formant (et le troisième) a une amplitude inférieure à celle du fondamental et du premier formant. Ceci va se traduire sur l'histogramme H1.

Pour les fricatives non voisées H2 est très semblable à H1 (fig. 2) mais pour les sons voisés il montre la présence d'une basse fréquence fig. 3 et 4).

Conclusion

Remarquons que H2 seul nous permet de séparer les sons voisés mais il nécessite le calcul des passages par zéro du signal. C'est pourquoi nous continuons à faire un premier choix d'après H1 car nous disposons directement des informations l_i pour le construire et il règle la majorité des cas.

Nous avons dans tous les cas un moyen sûr et rapide de détecter le caractère voisé ou non d'un signal vocal.

3. Détection de chaque cycle de voisement

- Le cycle de voisement d'un signal vocal voisé fait apparaître trois grandes caractéristiques (fig. 3a et 4a) :
- un maximum d'amplitude dû au maximum de résonance de chaque composante du signal,
- un amortissement de ces composantes pendant le cycle,
- une certaine périodicité dans la répétition du cycle.

Notons de plus que tous les sons voisés donnent des amplitudes a_i de même signe pour le maximum de résonance.

- Nous cherchons un extremum e_i qui définisse le Début du Cycle de Voisement (D.C.V.) avec suffisamment de régularité.

Nous choisissons un point qui se situe visuellement au début du maximum de la zone de résonance. Le D.C.V. ainsi choisi n'est pas tou-

.../...

jours le maximum absolu d'amplitude e_{max} par suite des amplitudes, des fréquences et des phases respectives des différentes composantes du cycle (rappelons entre autre que le déphasage des composantes augmente avec leur fréquence) (fig. 3a, 4a et 5)

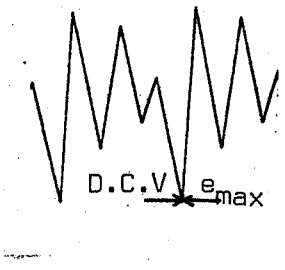


fig. 5 son ɔ̃

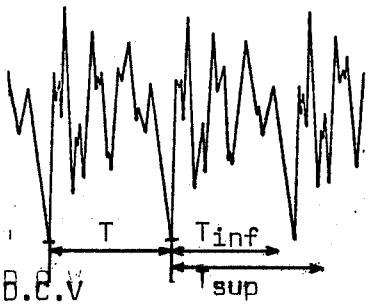


fig. 6 son e

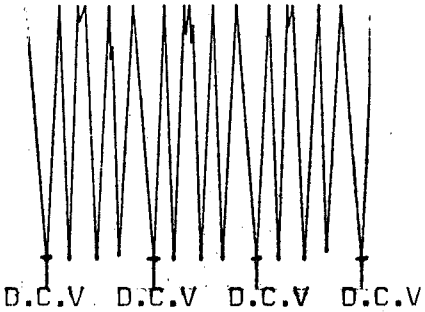


fig. 7 son a

➨ Dans un premier temps nous supposons connue la position d'un D.C.V. ainsi que la valeur de la période fondamentale T.

Cette période varie assez lentement d'un cycle au suivant : moins de 20 %. Aussi à partir du dernier D.C.V. trouvé nous cherchons le D.C.V. suivant dans une fenêtre incluant les variations possibles de T (fig. 6). Nous avons pris T_{inf} = 0.875 T et T_{sup} = 1.25 T.

Le D.C.V. étant voisin de e_{max} nous commençons par chercher e_{max} dans cette fenêtre, puis nous tenons compte des formants présents dans le cycle pour trouver le D.C.V.

Trois cas principaux se présentent :

a/ Cas des sons u, ɔ̃ et parfois des sons a, ã, ə, o (fig. 5). Il y a un seul e_i dans la fenêtre ou bien les autres e_i sont d'amplitude beaucoup plus faible que e_{max}. Nous prenons e_{max} comme D.C.V.

b/ Cas des sons i, e, ɛ, y, ɛ̃ (fig. 4a). Il y a une succession de e_i d'amplitudes voisines dûe à la présence d'un deuxième formant important et de fréquence élevée.

Tout en restant entre les deux passages par zéro du signal direct qui encadrent e_{max}, nous recherchons l'extremum e_i le plus à gauche tel que $a_i > \frac{3}{4} a_{max}$ et nous le prenons comme D.C.V.

c/ Cas des sons a, ã, ə, o (fig. 3a). Ces phonèmes ont un premier formant situé entre 400 et 1000 Hz et un deuxième formant assez faible. Il arrive souvent que la deuxième oscillation du cycle soit d'amplitude plus forte que la première. Mais la valeur t_{max} (due à la fréquence du premier formant) est toujours inférieure à 1,2 ms alors que le temps de montée du D.C.V. est toujours plus grand : typiquement 1,2 à 2,5 ms. Nous pouvons (sous certaines conditions) éliminer le premier e_{max} trouvé et revenir en arrière jusqu'à un nouvel e_{max} satisfaisant et choisir le D.C.V. comme d'habitude.

➨ Il se pose un problème particulier par suite du codage des amplitudes sur 64 niveaux. Les consonnes ont en moyenne une amplitude de 20 % de celle des voyelles. Nous voulons avoir le minimum de contrainte au niveau de l'intensité d'élocution de la parole aussi plutôt que d'avoir une amplitude trop faible pour les consonnes nous préférons accepter une saturation en amplitude des voyelles.

De ce fait, nous pouvons avoir des amplitudes égales pendant une partie plus ou moins longue du cycle de voisement : le caractère amorti du signal est masqué en partie (fig. 7).

.../...

Il peut même arriver qu'une amplitude saturée du cycle précédent apparaîsse dans la fenêtre de recherche du nouveau D.C.V. Il faut donc attendre la fin de la saturation du cycle en cours avant de chercher le nouveau D.C.V.

Une saturation importante de ce type n'est pas forcément préjudiciable à la bonne reconnaissance du mot global.

Bien entendu, au delà d'une certaine saturation le programme demande au locuteur de bien vouloir parler moins fort !

Initialisation du processus

Actuellement nous supposons que la fréquence fondamentale se situe entre 100 et 200 Hz. Nous cherchons les premiers maxima absolu dont les distances les uns aux autres satisfont à l'hypothèse sur la fréquence Nous obtenons ainsi une valeur approchée de la période fondamentale.

A l'avenir nous utiliserons le programme de recherche des formants (voir §.V) en le faisant fonctionner sur une fenêtre suffisamment grande pour que la plus basse fréquence trouvée soit celle du fondamental.

Une fois connue, une valeur approchée de la période T (quelque soit la méthode utilisée) il faut initialiser la recherche des D.C.V. dans chaque phonème voisé.

Pour cela nous prenons comme premier D.C.V. l'extremum e_{max} de la fenêtre de longueur T commençant à la fin du silence ou à la première zone identifiée comme voisée à la suite d'un son non voisé.

Conclusion

Le programme ainsi réalisé permet de suivre toutes les transitions entre phonèmes voisés quels que soient le locuteur et sa rapidité d'élocution.

La détection des D.C.V. est en accord avec celle faite visuellement sur le signal (fig. 8 et 9).

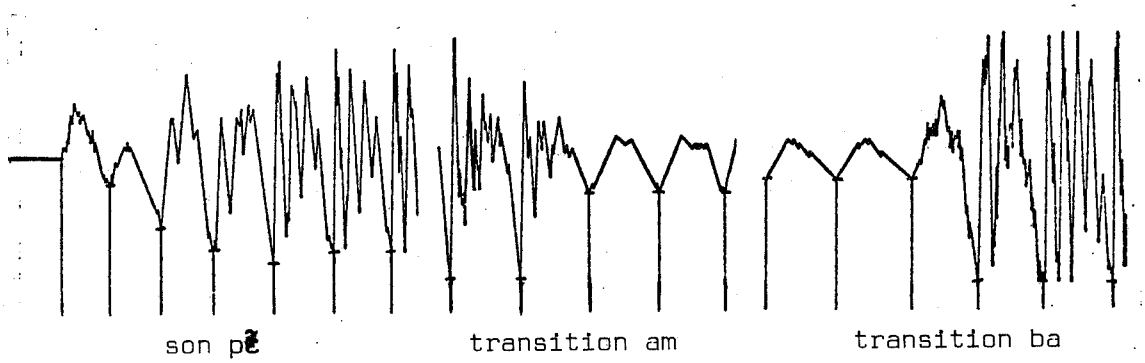


fig. 8

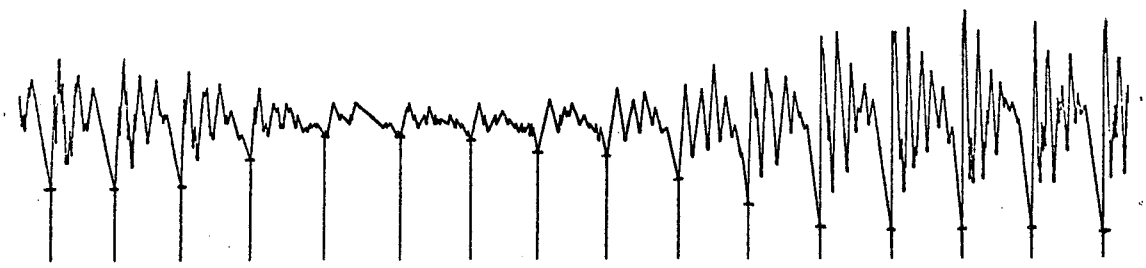


fig. 9 : transition ra de ira

.../...

4. Segmentation des phonèmes

- Les segmentations début de son et fin de son ont été vues au §.1. Lorsque le silence trouvé peut correspondre à celui d'une plosive non voisée, l'analyse des transitions vers les phonèmes voisins à l'aide du programme de recherche des formants devra séparer les trois classes p, t, k. Notons que l'explosion du p, qui est la plus marquée, se caractérise de plus par un démarrage du signal de signe contraire de celui d'un D.C.V.

- Il faut d'abord savoir si le son à analyser est voisé ou non. Cette décision est prise aussi bien après un silence (début de mot ou occlusion) que dans une transition voisée-non voisée ou l'inverse. Nous utilisons les critères définis au §.2. La fenêtre d'analyse pour H1 et H2 est de une à deux périodes T suivant les cas. Nous localisons donc ces transitions avec précision.

De plus l'examen de H1 et H2 nous permet de séparer les trois classes s, ʃ et f.

- Dans le cas d'une succession de sons voisés, l'examen de la succession des cycles de voisement permet d'une part de localiser avec précision les transitions et d'autre part de localiser les zones quasi-stables. Ainsi l'analyse fine du signal sur la recherche des formants n'est faite que là où elle est nécessaire.

Pour effectuer ces segmentations nous utilisons les variations d'amplitude des D.C.V., les variations du fondamental et des formants dont nous avons une idée rapide à l'aide de H1 et H2.

La pré-classification des phonèmes voisés est en cours et donnera au moins les classes suivantes :

- voyelles (dont la séparation en quelques classes sures est possible),
- plosives voisées : b, d, g,
- nasales : m, n,
- les liquides l, r seront seront probablement pré-classées comme voyelles ou nasales.

5. Analyse de la composition spectrale d'un cycle de voisement

5-1. Cette recherche est fondée sur l'observation de la courbe obtenue en joignant les extrema du signal initial (fig. 10a, 11a et 12a). En effet, l'observation visuelle de cette courbe permet de reconnaître les cycles de voisement et de séparer les composantes haute et basse fréquences du signal. Elle permet aussi d'avoir une information sur l'énergie de chaque bande de fréquence.

Nous cherchons donc à réaliser un programme qui fasse automatiquement cette séparation et qui nous donne des informations correspondant à la fréquence et à l'énergie.

5-2. Le programme est composé de deux parties.

La première partie réalise un filtrage de la fréquence la plus haute du signal en extrayant sa valeur et un coefficient énergétique associé.

La deuxième partie consiste à reconnaître les nouveaux extrema du signal filtré.

Détaillons chacune de ces parties.

5-2.1 - Filtrage

5-2.1a) Nous construisons l'histogramme des l_i : Cet histogramme permet de calculer un seuil de durée L qui sera nécessaire au filtrage proprement dit. Ce seuil L est égal à l'abscisse du 1er minimum suivant le 1er maximum de l'histogramme (fig. 10b, 11b, 12b).

.../...

5-2.1b) Nous pouvons calculer maintenant les paramètres du filtrages : D_L et A_L.

D_L est la valeur moyenne des durées inférieures ou égales à L. C'est une approximation de la demi-période de la fréquence à filtrer.

A_L est égal à la moyenne des valeurs absolues des différences d'amplitudes recontrées dans les segments de durées inférieures ou égales à L.

Si l'on pose : $\delta_i = 0$ si $l_i > L$
$= 1$ si $l_i \leq L$

$$N_L = \Sigma\, \delta_i$$

$$D_L = \frac{\Sigma\, l_i \times \delta_i}{N_L}$$

$$A_L = \frac{\Sigma\, |a_i - a_{i-1}| \times \delta_i}{N_L}$$

$F_L = \frac{1}{2D_L}$ est la fréquence à filtrer, A_L son coefficient énergétique associé.

5-2.1c) Tous les extrema e_i tels que $l_i \leq L$ sont déplacés au milieu du segment s_i. A la suite de ce traitement, tous les nouveaux e_i ne sont plus des extrema. Les erreurs d'arrondi au moment du calcul du milieu laissent subsister des variations dues à la fréquence filtrée. (fig. 10c, 11c, 12c)

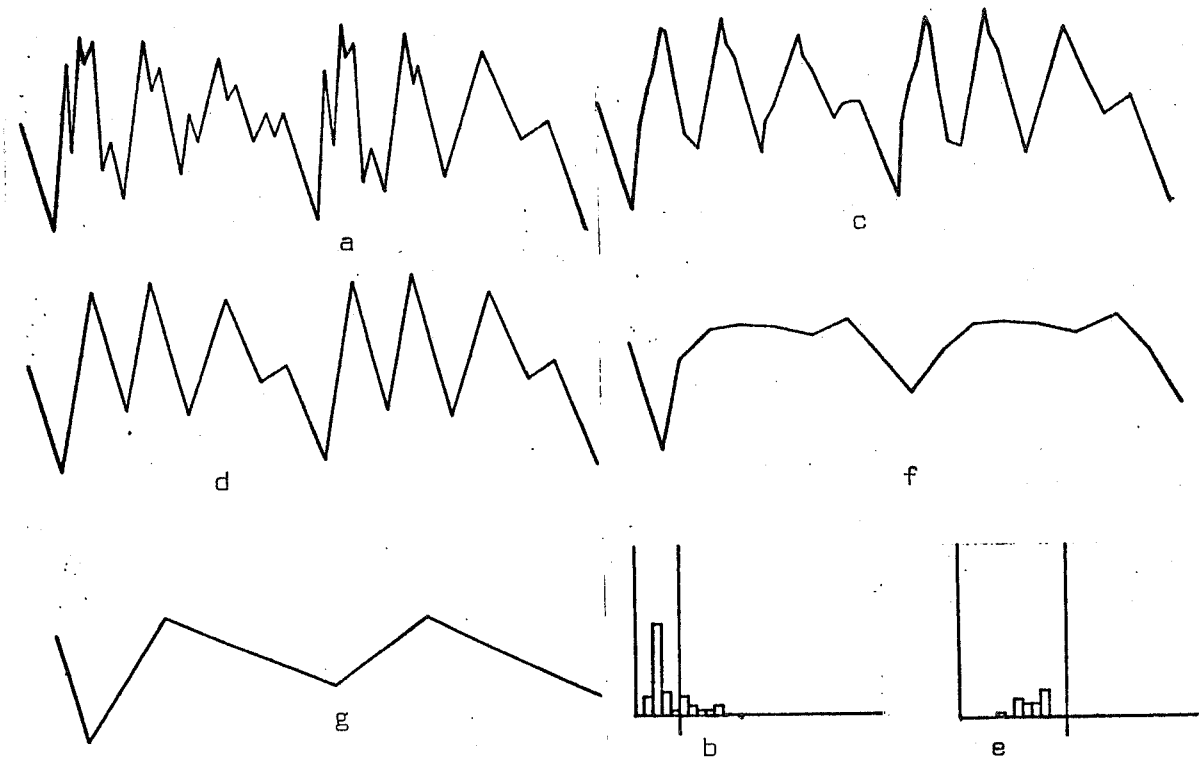


fig. 10

.../...

fig. 11

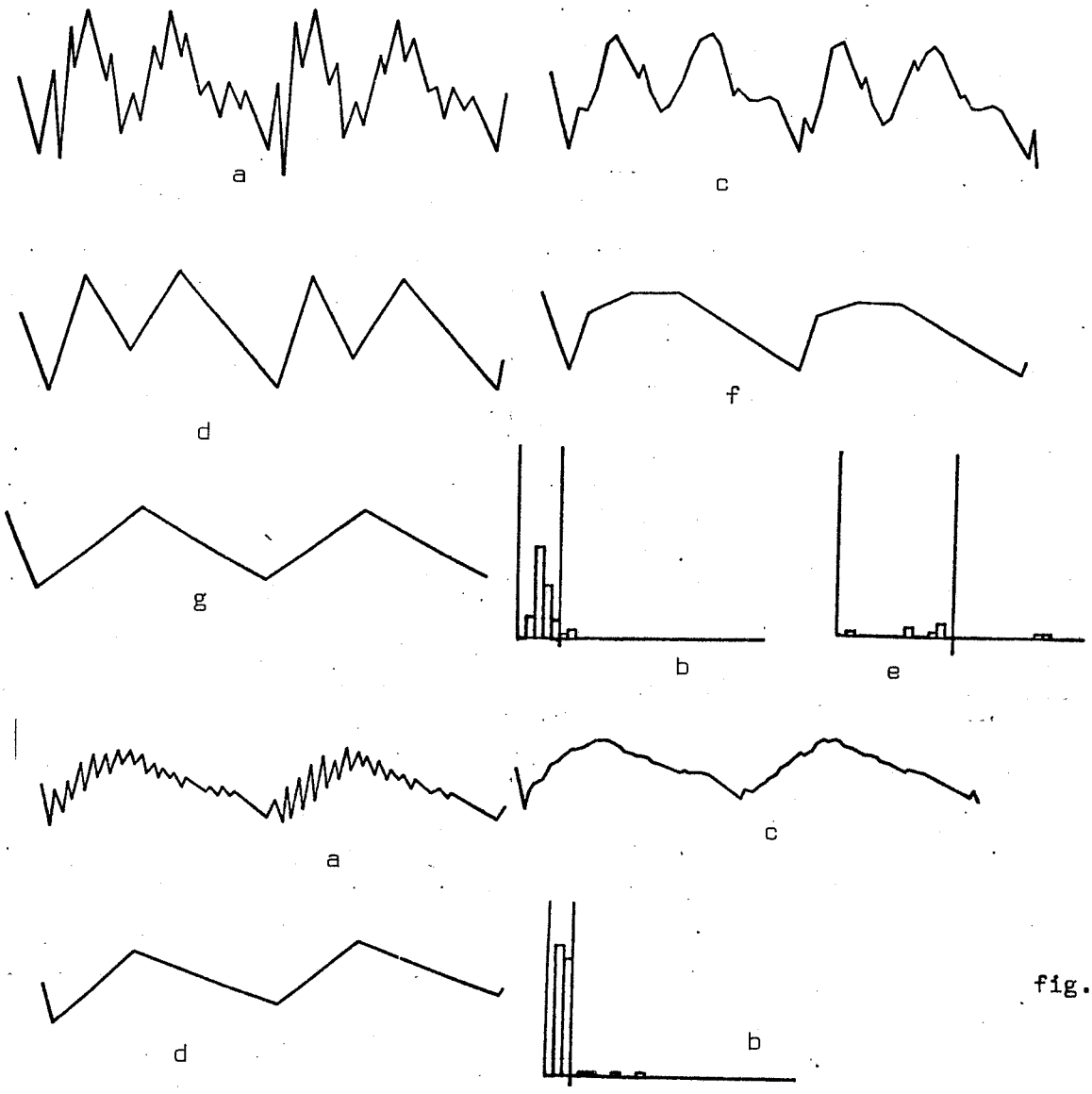


fig. 12

5-2.2 - Reconstruction d'une suite d'extrema avec les résultats du filtrage.

Le programme recalcule les extrema de la séquence filtrée.

Deux types d'extrema se présentent :

1/ Les e_i tels que $l_i \leq L$

2/ Les e_i tels que $l_i > L$

Nous nous interesserons ensuite au nombre d'extrema consécutifs n de type 1.

Ces extrema sont précédés par un extremum de type 2 appelé l_{2d} et suivis d'un extremum de type 2 appelé l_{2F}.

Soit l_1 la somme des l_i de type 1 consécutifs.

Si n est pair, $l_{2d} \leftarrow l_{2d} + l_{1/2}$ et $l_{2F} \leftarrow + l_{1/2}$.

.../...

FIGURE 13 →

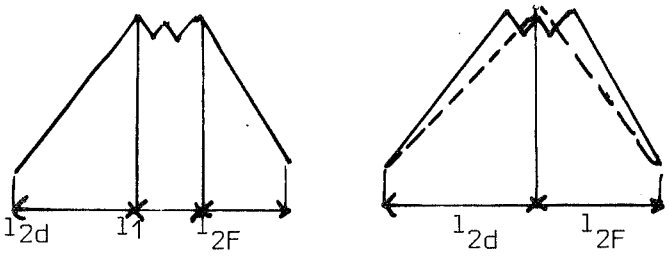
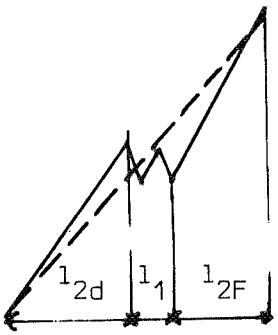


FIGURE 14 ↘

Si n est impair $l_{2d} \leftarrow l_{2d} + l_1 + l_{2F}$ et l_{2F} est supprimée.

On obtient donc une suite de segments tels que $l_i > L$. Les bords de la fenêtre d'analyse sont traités de façon particulière. (fig. 10d, 11d, 12d).

5-3. On itère les processus 5.1 et 5.2 jusqu'à ce que $L \geq P/2$. P étant la période du fondamental. (fig. 10efg, 11efg).



Les résultats obtenus peuvent être représentés en utilisant un sonogramme. (fig. 15)

Remarquons que nous obtenons directement les positions des fréquences formantiques.

La seule hypothèse faite sur le son étudié pendant la fenêtre d'analyse est qu'il soit suffisamment stable.

Cette hypothèse est nécessaire pour que l'histogramme nous permettre de calculer le seuil permettant le filtrage.

Ce programme utilisé sur quelques cycle dans la partie centrale des voyelles nous donne la fréquence et l'énergie des formants présents.

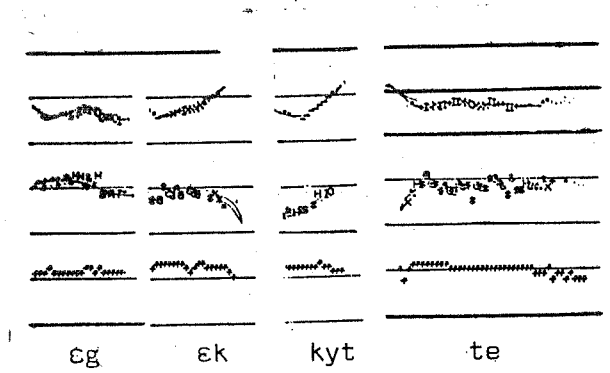


FIGURE 15

Si l'on exécute successivement sur les cycles d'une transition on peut suivre l'évolution en amplitude et en fréquence.

III. CONCLUSION

Le deuxième système de reconnaissance de la parole que nous réalisons, permettra de suivre l'évolution du signal vocal avec beaucoup plus de précision grâce à la détection du cycle de voisement et à l'extraction des fréquences formantiques.

Ces deux programmes, en cours de réalisation, devront être intègrés dans le système global avec les autres programmes en cours d'étude principalement la reconnaissance des phonèmes avec les nouveaux paramètres et la recherche en dictionnaire, compte tenu des erreurs possibles, pour la reconnaissance de mots isolés.

Notons que nous disposons d'une information sur la mélodie qui pourra être utilisée ultérieurement.

.../...

[1] M. BAUDRY, B. DUPEYRAT et C. FRANK
Reconnaissance Automatique de la Parole - Etude de la Segmentation
(3éme Journée du G.A.L.F. - LANNION du 31.05 au 2.06 1972)

[2] M. BAUDRY et B. DUPEYRAT
Temporal Analysis of the Voice Signal - Automatic Recognition of the spoken words.
(IInd International Joint Conference on Pattern Recognition - COPENHAGUEN du 13 au 15 Août 1974)

[3] B. DUPEYRAT
Reconnaissance Automatique de la Parole (Méthode des Passages par Zéro) Reconnaissance Automatique de Voyelles Isolées
(Thèse de Docter-3éme Cycle - Soutenue le 10 Juin 1975 à l'Université de Paris-VI-ORSAY)

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

UN MODELE D'OREILLE APPLIQUE A L'ANALYSE DE LA PAROLE

CAELEN J. PERENNOU G.

C. E. R. F. I. A. Université Paul Sabatier TOULOUSE

RESUME :

Dans le cadre du projet A. R. I. A. (Analyse et Reconnaissance des Informations Acoustiques), cette étude est orientée vers le traitement et l'analyse de la parole. Le modèle mathématique d'oreille décrit, est un système autoadaptatif à plusieurs niveaux de filtrage : l'oreille moyenne, filtre variable du second ordre, la cochlée, banc de filtres couplés, les neurones, système de filtres à sortie impulsionnelle. Ces impulsions sont comptées dans un histogramme de fréquence. Les paramètres extraits du "spectre d'énergie" et de l'histogramme, permettent d'analyser la parole dans les termes acoustiques habituels: structure formantique, plosive, fricatice etc ...

SUMMARY :

In the scope of the plan A. R. I. A. (Analyse et Reconnaissance des Informations Acoustiques), this research is oriented to the treatment and analysis of full speech. The mathematical model of ear that we describe, is selfadaptatif system with several levels of filtering : the middle ear works as a variable filter, the cochlea as a series of coupled filters, the neurons as a system of filters spread along the cochlea which give impulsions which are counted and recorded in histogram. At the output of the whole system, the parameters extracted from the energy spectrum and from frequency histogram enable the analysis of speech according to the wanted acoustical parameters : formantic, fricative, stop-consonant structure ...

UN MODELE D'OREILLE APPLIQUE A L'ANALYSE DE LA PAROLE

CAELEN J. (1) - PERENNOU G.(2)

C.E.R.F.I.A. TOULOUSE (3)

1. INTRODUCTION

A défaut de fournir des conclusions sur les mécanismes de la perception auditive, un modèle d'oreille peut fournir des vues intéressantes sur les méthodes d'analyse et de détection des paramètres d'un signal acoustique.

Nous rapppelons ci-après les éléments essentiels du modèle déjà décrit dans (13) (21) et soulignons l'intérêt du triple filtrage autoadaptatif de l'oreille ainsi que des différents niveaux d'analyse.

Cette étude s'insère dans le projet ARIA mené au Laboratoire C.E.R.F.I.A.(3) La recherche des méthodes pour isoler les formants et l'adaptation du modèle au signal fait l'objet de la collaboration. L'incidence sur le modèle et le traitement informatique est effectuée par Mr. CAELEN.

2. DESCRIPTION DU MODELE

1.1. Oreille externe et oreille moyenne

Peuvent se modéliser par un filtre du second ordre du moins pour des fréquences inférieures à 5000 hz (1) (2) (3) (4). L'action des muscles tympanique et stapédien modifient les caractéristiques de ce filtre : sa fréquence centrale se déplace vers les aigus avec l'intensité du signal d'entrée, sa bande passante diminue légèrement (fig.1). Le modèle choisi est un filtre numérique variable du second ordre.

1.2. la cochlée

Comprend deux parties (13) :

1. Equation de propagation d'une onde de compression dans la périlymphe ;
2. Equation de vibration de la membrane basilaire.

(1) Assistant à l'I.U.T.
(2) Professeur à l'U.P. - I.U.T.
(3) Cybernétique des Entreprises, Reconnaissance des Formes, Intelligence Artificielle - 118, route de Narbonne - 31077 TOULOUSE CEDEX -

$$\frac{1}{C^2(X)} \frac{\partial^2 Y}{\partial t^2} - \varepsilon(X,Y) \frac{\partial^2 Y}{\partial X^2} + \lambda(X,Y) \frac{\partial Y}{\partial t} + \mu(X,Y)Y = g(X) f\left(t - \frac{X}{V}\right)$$

$Y(X,t)$ déplacement de la membrane au point X

Conditions initiales $Y(X,0) = \frac{\partial t}{\partial t}(X,0) = 0$

Conditions aux limites $Y(0,t) = Y(\ell,t) = 0$ ℓ longueur de la M.B.

$g(X)\, f\left(t - \frac{X}{V}\right)$ solution de 1.

$C(X,Y)$	célérité de l'onde
$\varepsilon(X,Y)$	tension longitudinale (couplage)
$\lambda(X,Y)$	amortissement
$\mu(X,Y)$	tension transversale

N.B. Ces coefficients sont des équivalents mécaniques des coefficients de l'ensemble du canal cochléaire.

Dans ces conditions le comportement de la membrane basilaire est celui des filtres variables couplés (fig. 2).

1.3. Filtrage neuronique

Les cellules internes et externes fixées à l'organe de Corti, recueillent les potentiels modulés par les déplacements de la membrane basilaire. Ces potentiels excitent un premier réseau de neurones qui se comportent comme des filtres (7)(8)(9)(10)(11)(12)(15) et délivrent en sortie des impulsions modulées en fréquence et en intensité. Le fréquence centrale de ces filtres est en correspondance avec la fréquence de résonance de la membrane basilaire. Il semble que la bande passante de ces filtres soit étroite (16) (fig. 3).

1.4. La logique de détection

1.4.1. Le spectre d'énergie :

$$W(X, t_2 - t_1) = \int_{t_1}^{t_2} Y^2(X,t)\, dt \qquad t_2 - t_1 = cte$$

1.4.2. L'inhibition latérale des neurones

Si X_0 est tel que $\frac{\partial W}{\partial X} = 0$ et $\frac{\partial^2 W}{\partial X^2} < 0$

et Z(X,t) le signal de sortie du neurone correspondant au point X de la membrane basilaire alors :

$$\forall X \in V(X_0) \quad \begin{cases} Z(X,t) = 0 \text{ si } W(X,t_2-t_1) < W(X_0,t_2-t_1).\ \alpha\ . \\ Z(X,t) = Z(X,t) \text{ sinon} \end{cases}$$

avec $0 < \alpha < 1$

$V(X_0)$ un voisinage de X_0.

1.4.3. Les passages par zéro des signaux z(x,t):

A chaque changement de signe de z(x,t) (pour $z(x,t) \neq 0$) le neurone N_x correspondant émet une impulsion δ_ε^x. On mémorise, par comptage dans un histogramme H(f), les intervalles θ entre deux impulsions successives $f = \frac{1}{\theta}$.

On calcule également pour chaque x, la fréquence moyenne et la variance des passages par zéro ainsi obtenus : $\underline{f}^x$, v^x.

1.5. La logique de commande

1.5.1. Le filtre variable de l'oreille moyenne :

Soient f_{om}^1 et f_{om}^2 les fréquences minimale et maximale que peut atteindre le filtre, et $u_w \in [0,1]$ un coefficient dépendant de l'énergie du signal d'entrée par la relation $u_w = k \log \frac{w_e}{\underline{w}_e}$ $w_e \geq \underline{w}_e$

où w_e est l'énergie du signal d'entrée mesuré sur un intervalle de temps fixe et $\underline{w}_e$ l'énergie minimum perçue.

On définit alors la fréquence centrale du filtre par :

$$f_{om} = f_{om}^1 + u_w\ (f_{om}^2 - f_{om}^1)$$

et l'amortissement par :

$$\alpha_{om} = \alpha_{om}^1\ u_w + \alpha_{om}^2$$

1.5.2. Les filtres couplés de la membrane basilaire :

On définit de même les fréquences centrales $f^x(t)$ en tenant compte d'un facteur de confiance $u_c^x = \frac{1}{1+V^x}$ $u_c^x \in [0,1]$ et de la valeur précédente $f^x(t-\delta t)$ où δt est la période de réajustement des coefficients.

$$f^x(t) = f^x(0) + u_w (f^x(t-\delta t) + u_c^x (\underline{f}^x - f^x(t-\delta t)) - f^x(0))$$

La répartition des filtres de départ $f^x(0)$ est logarithmique (15) (18)(19) (fig.4).

3. ANALYSE DE LA PAROLE

3.1. La segmentation consonne-voyelle

Elle se fait sur le signal d'entrée de la cochlée (potentiel microphonique) plus précisemment sur la courbe d'énergie calculée sur une durée fixe de 15 ms. Les maxima correspondent aux zones vocaliques et les maxima aux consonnes ou au bruit. Des conditions supplémentaires de durée permettent d'éliminer les accidents de la courbe. (fig.5).

3.2. Les traits acoustiques extraits du spectre

Après lissage grossier de la courbe w(x) on détecte les principaux maxima en x c'est-à-dire $\{x_0 : \forall x \in V(x_0) \quad w(x_0) > w(x)\}$ où $V(x_0)$ est un voisinage de x_0 dont l'ordre de grandeur est du tiers de la mebrane basilaire.

Dans ces conditions on peut définir les traits acoustiques suivants :

(1) $\exists ! \; x_0 \in V(\ell)$: c'est-à-dire il existe un maximum d'énergie dans les fréquences basses (ℓ extrêmité apicale de la cochlée).

(2) $\frac{dx_0}{dt} >> 1$: zone d'instabilité du spectre.

(3) $\exists ! \; x_0 \in V(0)$: c'est-à-dire il existe un maximum d'énergie dans les hautes fréquences.

(4) card $\{x_0\} \geq 3$: par plusieurs maxima réparties le long de la membrane, ce trait caractérise la structure vocalique.

Ces traits permettent alors de classer les phonèmes en grandes catégories.

Traits / Phonèmes	(1)	(2)	(3)	(4)
Voyelles				x
semi-voyelles liquides		x		x
nasales	x			
plosives		x		
fricatives	x si voisé		x	

3.3. Détection des formants

Il apparait sur la membrane basilaire des sons différentiels dont le comportement n'est ni linéaire avec la fréquence ni linéaire avec l'amplitude des composantes (6)(8)(20). L'expression de ces distorsions n'est donc pas prévisible, encore moins pour la parole dont la structure est complexe. Pour avoir une détection sûre des formants il faut donc multiplier les critères et les vérifications.

(1) $\{f_1 : \max H(f) \text{ en } f_1\}$ on relève les maxima sur l'histogramme des passages par zéro.

(2) $\{f_2 : f_2 = \underline{f}_x \text{ si } v^x \text{ est minimum en } x\}$ on relève les fréquences moyennes des points x pour lesquels la variance est minimum.

(3) $\{f_3 : f_3 = f_x \text{ si } w(x) \text{ est maximum en } x\}$ on relève la fréquence des filtres pour lesquels l'énergie de sortie est maximum.

Les formants {f} vérifient l'ensemble de ces trois critères c'est-à-dire : $\{f\} = \{f_1\} \cap \{f_2\} \cap \{f_3\}$.

La hauteur de crête dans H(f) peut servir de paramètre supplémentaire puisque fonction des amplitudes des formants (fig.7).

4. CONCLUSION

On peut d'ores et déjà remarquer l'amélioration qu'apporte le filtrage à bande étroite des neurones sur le filtrage à bande large de la membrane basilaire. Le seul filtrage même à bande étroite de cette dernière (apparition du spectre harmonique de la glotte) ne permet pas une localisation suffisamment précise des formants. Le rôle des filtres variables est lui aussi prépondérant dans la précision des résultats, la densité des filtres autour de certains pôles d'attraction limite les influences de régions voisines, le couplage les groupe lorsque l'intensité augmente et l'amortissement croissant limite le temps de réponse et des transitoires (intérêt dans les phénomènes brefs). Le jeu de l'oreille moyenne dans ce système apporte lui aussi une adaptation en fonction de l'énergie et une sensibilité plus grande dans les moyennes et hautes fréquences (région des 2ème et 3ème formants).

Il est permis de penser que les performances de l'oreille humaine sont en rapport avec son pouvoir d'adaptation aux signaux les plus complexes. Il semble également qu'une bonne qualité de perception ne soit possible qu'après plusieurs niveaux de filtrage.

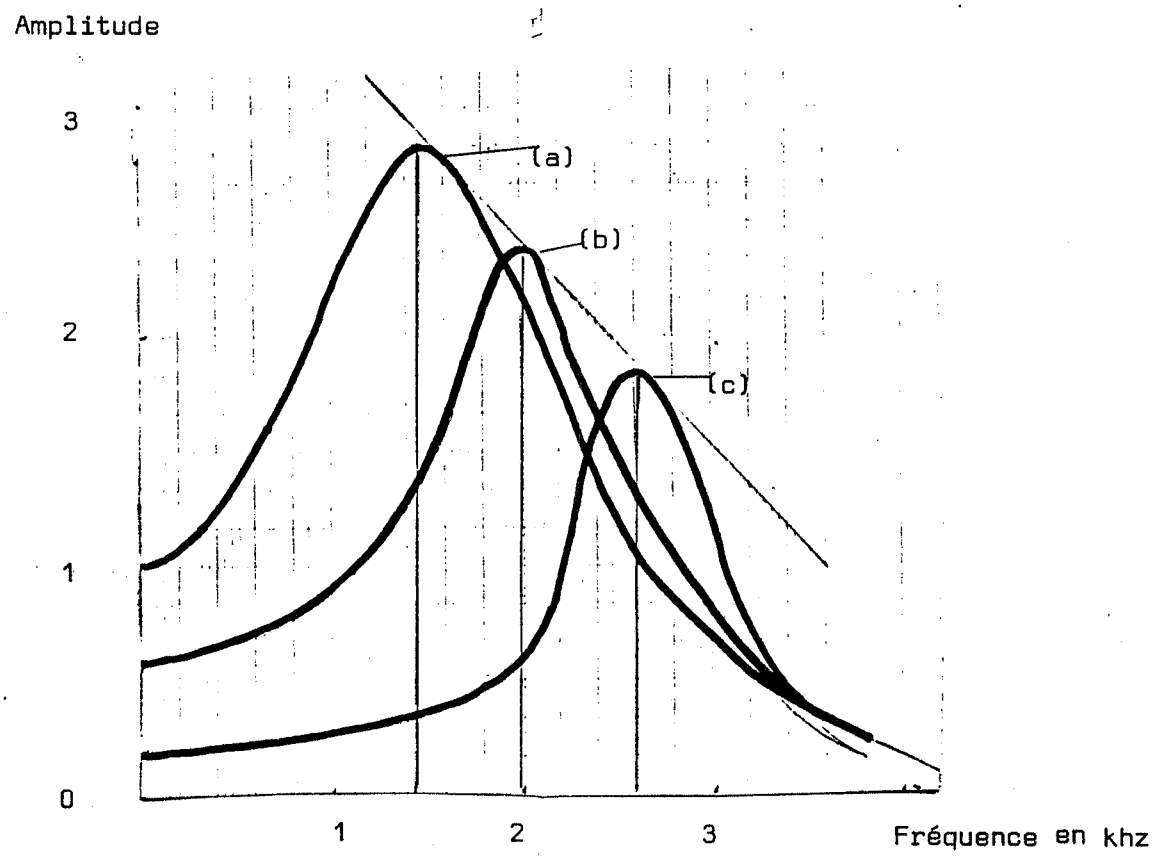


Fig.1 - Courbes de résonance du filtre "oreilles externe et moyenne"
(a) et (c) courbes limites à faible et forte énergies.

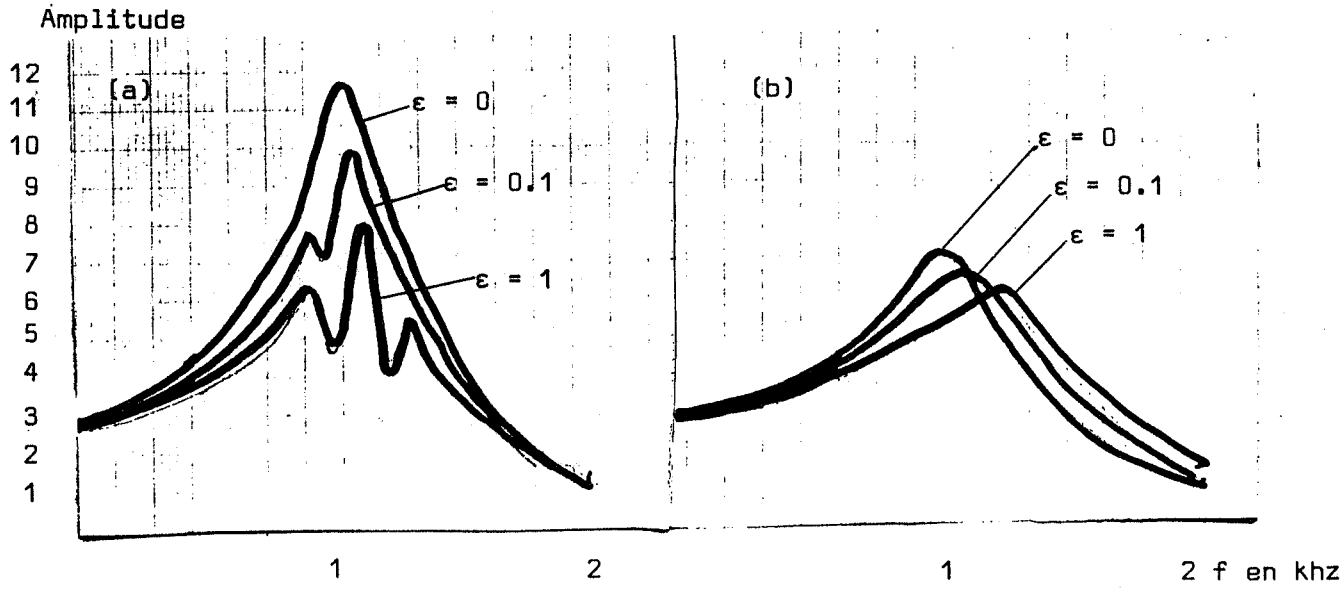


Fig.2 - Courbes de résonance d'un point de la membrane basilaire
(a) $\lambda = 5$
(b) $\lambda = 50$

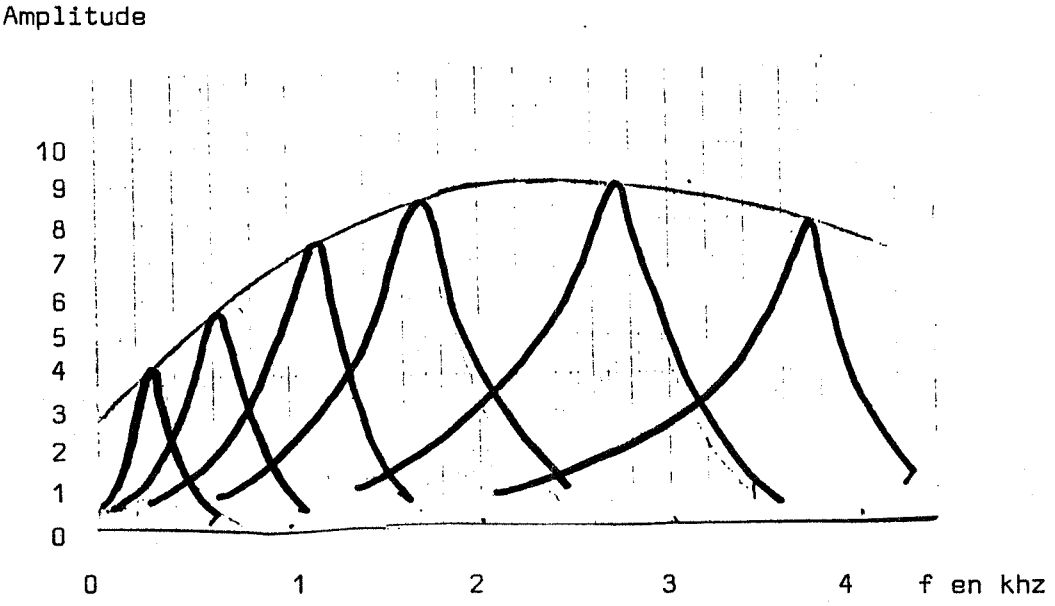


Fig.3 - Courbes de résonance de quelques neurones. L'enveloppe des crêtes est fonction de la courbe de sensibilité de l'oreille. La bande passante et la répartition en fréquence de ces filtres sont prises conformément aux données de E. ZWICKER (18)(19).

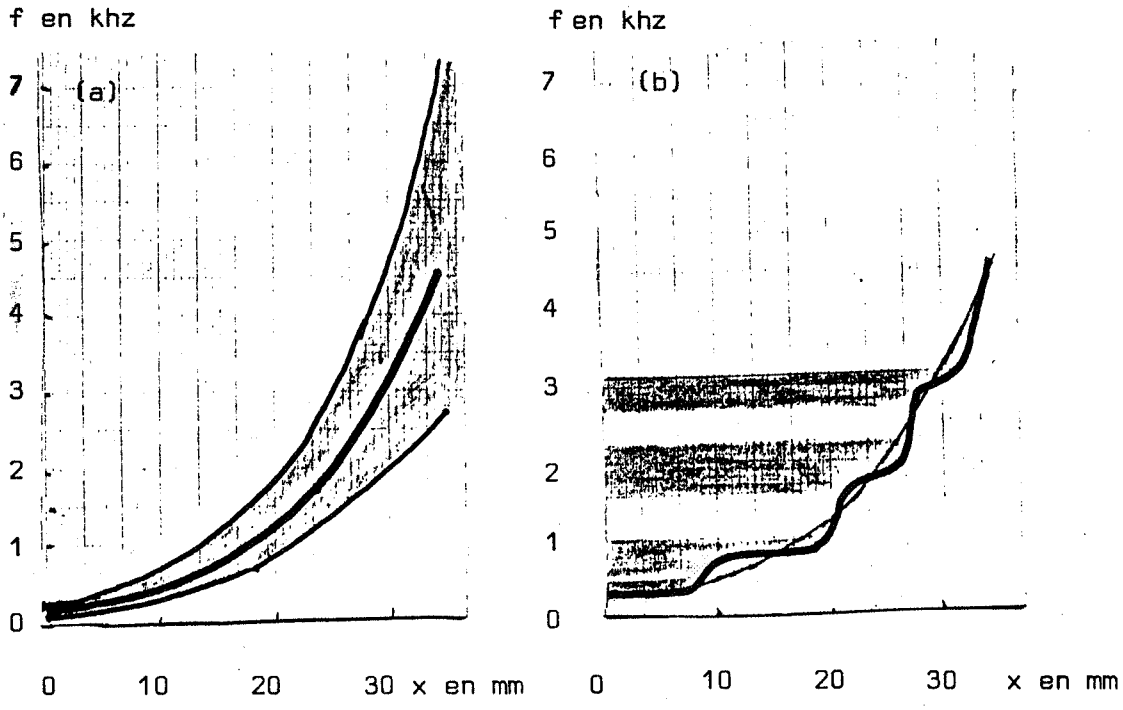


Fig.4 - (a) Répartition des filtres et bande passante en grisé

(b) Répartition ajustée $f_x(t)$ pour le "i" français. En grisé les zones de plus grande concentration des filtres autour des formants.

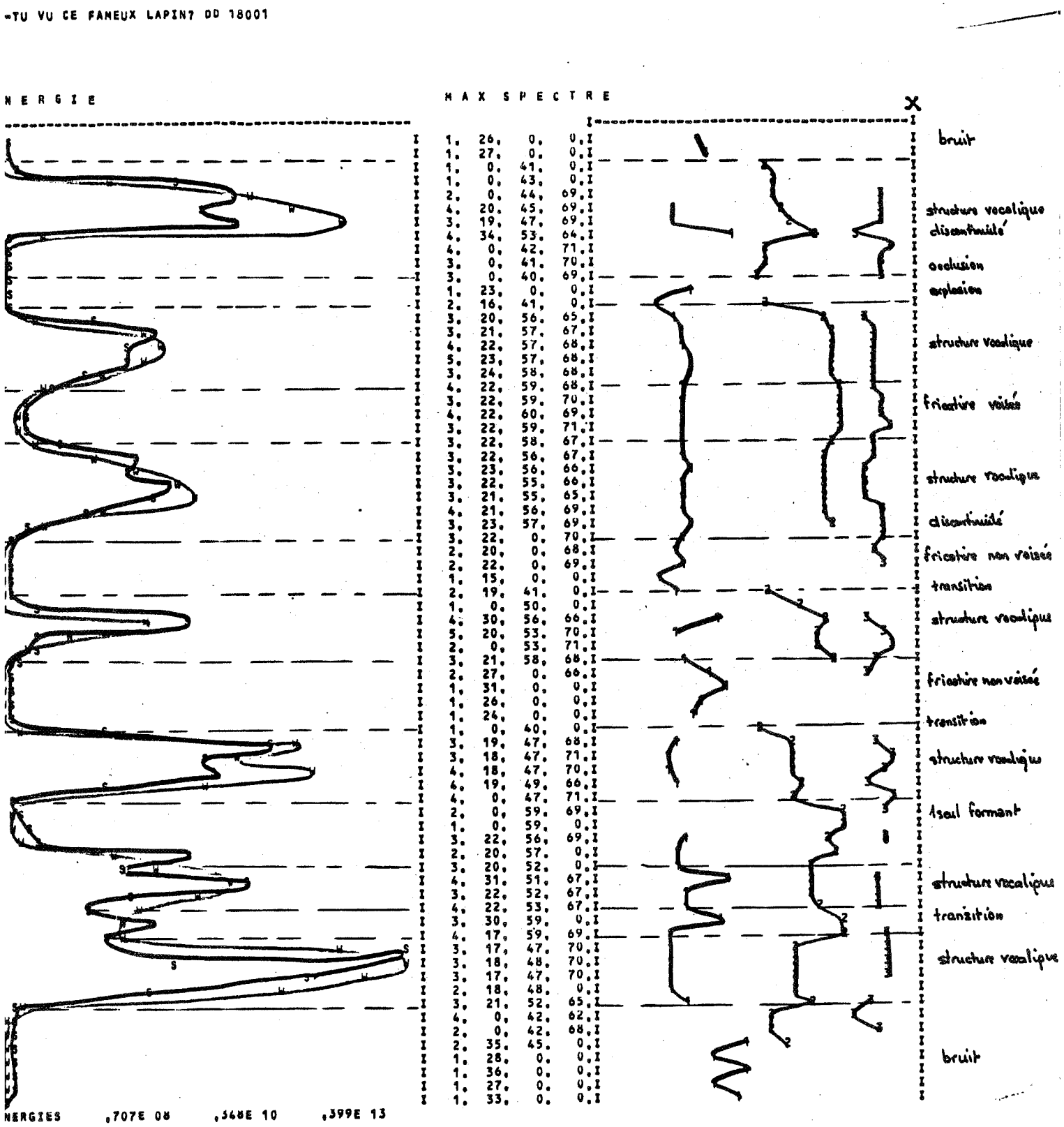


Fig 5. Courbes d'énergie du signal et de la membrane basilaire
Evolution des maxima du spectre d'énergie dans la plan (x,t)

```
7500.  1     1.
6667.  2     3.
6000.  3    14.
5455.  4     6.
5000.  5     9.
4615.  6    10.
4286.  7    11.
4000.  8    20.
3750.  9    27.
3529. 10    28.
3333. 11    51. I
3158. 12   115. IIII
3000. 13   201. IIIIIII
2857. 14   326. IIIIIIIIIII
2727. 15   365. IIIIIIIIIIII
2609. 16   275. IIIIIIIII
2500. 17   178. IIIIII
2400. 18   178. IIIIII
2308. 19   106. III
2222. 20    58. II
2143. 21    64. II
2069. 22    33. I
2000. 23    30. I
1935. 24    32. I
1875. 25    37. I
1818. 26     8.
1765. 27    13.
1714. 28    18.
1667. 29     9.
1622. 30     5.
1579. 31     0.
1538. 32    25.
1500. 33     5.
1429. 34    32. I
1364. 35    68. II
1304. 36   165. IIIII
1250. 37   295. IIIIIIIIII
1200. 38   735. IIIIIIIIIIIIIIIIIIIIIIII
1154. 39   991. IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
1111. 40   431. IIIIIIIIIIIIII
1071. 41   178. IIIIII
1034. 42    37. I
1000. 43    15.
 968. 44    16.
 938. 45     8.
 909. 46    17.
 882. 47    34. I
 857. 48   133. IIII
 833. 49   750. IIIIIIIIIIIIIIIIIIIIIIIII
 811. 50  1368. IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 789. 51   567. IIIIIIIIIIIIIIIII
 769. 52    79. II
 750. 53    41. I
 714. 54    11.
 682. 55     0.
 652. 56     0.
 625. 57     0.
 600. 58     0.
 577. 59     0.
 556. 60     0.
 536. 61     0.
 517. 62     0.
 500. 63     0.
 484. 64     0.
 469. 65     0.
 455. 66     0.
 441. 67     0.
 429. 68     0.
 417. 69     0.
 405. 70     0.
 395. 71     0.
 385. 72     0.
 375. 73     0.
 214. 98     0.
 208. 99     0.
 203.100     0.
 197.101     0.
 192.102     0.
 188.103     0.
 183.104     0.
 179.105     0.
 174.106     0.
 170.107     0.
 167.108     0.
 163.109     0.
 160.110     0.
 156.111     0.
 153.112     0.
 150.113   308. IIIIIIIIII
 144.114   847. IIIIIIIIIIIIIIIIIIIIIIIIIIII
 139.115   328. IIIIIIIIII
 134.116    56. I
 129.117     0.
 125.118    60. II
 121.119     0.
 117.120     0.
 114.121     0.
 110.122     0.
 107.123     0.
 104.124     0.
 101.125     0.
```

```
7500.  1     3.
6667.  2     7.
6000.  3     9.
5455.  4    11.
5000.  5    16. I
4615.  6    13.
4286.  7    16. I
4000.  8    10.
3750.  9    15.
3529. 10    33. II
3333. 11    39. II
3158. 12    24. I
3000. 13    39. II
2857. 14    48. III
2727. 15    65. IIII
2609. 16    53. III
2500. 17    12.
2400. 18    29. I
2308. 19    37. II
2222. 20    38. II
2143. 21    50. III
2069. 22    26. I
2000. 23    31. II
1935. 24    16. I
1875. 25    28. I
1818. 26    17. I
1765. 27    17. I
1714. 28    49. III
1667. 29    18. I
1622. 30     9.
1579. 31    24. I
1538. 32    39. II
1500. 33    51. III
1429. 34   113. IIIIIII
1364. 35   147. IIIIIIIII
1304. 36   171. IIIIIIIIIII
1250. 37   392. IIIIIIIIIIIIIIIIIIIIIIIII
1200. 38   407. IIIIIIIIIIIIIIIIIIIIIIIIII
1154. 39   351. IIIIIIIIIIIIIIIIIIIIIII
1111. 40   309. IIIIIIIIIIIIIIIIIII
1071. 41   199. IIIIIIIIIIIII
1034. 42   221. IIIIIIIIIIIII
1000. 43   183. IIIIIIIIIII
 968. 44   133. IIIIIIIII
 938. 45    90. IIIIII
 909. 46   310. IIIIIIIIIIIIIIIIIII
 882. 47   415. IIIIIIIIIIIIIIIIIIIIIIIII
 857. 48   710. IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 833. 49   474. IIIIIIIIIIIIIIIIIIIIIIIIIIIII
 811. 50   384. IIIIIIIIIIIIIIIIIIIIIII
 789. 51   365. IIIIIIIIIIIIIIIIIIIII
 769. 52   247. IIIIIIIIIIIIII
 750. 53   522. IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 714. 54   236. IIIIIIIIIIIII
 682. 55   202. IIIIIIIIIII
 652. 56   141. IIIIIIIII
 625. 57    73. IIII
 600. 58    64. IIII
 577. 59    92. IIIIII
 556. 60    28. I
 536. 61    14.
 517. 62    44. II
 500. 63    31. II
 484. 64    16. I
 469. 65    16. I
 455. 66     0.
 441. 67    69. IIII
 429. 68     0.
 417. 69    18. I
 405. 70     0.
 395. 71     0.
 385. 72     0.
 375. 73    61. IIII
 214. 98     0.
 208. 99     0.
 203.100     0.
 197.101     0.
 192.102     0.
 188.103     0.
 183.104     0.
 179.105     0.
 174.106     0.
 170.107     0.
 167.108     0.
 163.109    94. IIIIII
 160.110     0.
 156.111    49. III
 153.112     0.
 150.113    52. III
 144.114   107. IIIIIII
 139.115   273. IIIIIIIIIIIIIIII
 134.116    57. III
 129.117   117. IIIIIII
 125.118     0.
 121.119    62. IIII
 117.120     0.
 114.121   134. IIIIIIIII
 110.122     0.
 107.123    72. IIII
 104.124   219. IIIIIIIIIIIII
 101.125     0.
```

Fig 6. Histogrammes du a à gauche, pris à la sortie des neurones, à droite pris sur la membrane basilaire.

C?.	8	17.	
5C.	9	86.	III
29.	10	1046.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
33.	11	787.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
58.	12	68.	III
00.	13	13.	
57.	14	5.	
27.	15	0.	
C9.	16	0.	
CC.	17	6.	
C?.	18	10.	
C8.	19	63.	II
22.	20	183.	IIIIIIII
43.	21	777.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
69.	22	260.	IIIIIIIIIIII
00.	23	46.	II
35.	24	4.	
75.	25	4.	
18.	26	8.	
65.	27	0.	
14.	28	4.	
67.	29	0.	
22.	30	0.	
79.	31	0.	
38.	32	0.	
C0.	33	0.	
29.	34	0.	
64.	35	0.	
C4.	36	0.	
50.	37	0.	
00.	38	0.	
54.	39	0.	
11.	40	0.	
71.	41	0.	
34.	42	0.	
00.	43	0.	
68.	44	0.	
38.	45	0.	
09.	46	0.	
82.	47	0.	
57.	48	0.	
33.	49	0.	
11.	50	0.	
89.	51	0.	
769.	52	0.	
750.	53	0.	
714.	54	0.	
682.	55	0.	
652.	56	0.	
625.	57	0.	
600.	58	0.	
577.	59	0.	
556.	60	0.	
536.	61	0.	
517.	62	0.	
500.	63	0.	
484.	64	0.	
469.	65	0.	
455.	66	0.	
441.	67	0.	
429.	68	0.	
417.	69	0.	
405.	70	19.	
395.	71	19.	
385.	72	20.	
375.	73	41.	I
366.	74	21.	
357.	75	128.	IIIII
349.	76	109.	IIII
341.	77	89.	IIII
333.	78	68.	III
326.	79	70.	III
319.	80	143.	IIIIII
313.	81	218.	IIIIIIIIII
306.	82	99.	IIII
300.	83	25.	I
294.	84	103.	IIII
288.	85	52.	II
283.	86	0.	
278.	87	0.	
273.	88	0.	
268.	89	0.	
263.	90	0.	
259.	91	0.	
254.	92	0.	
250.	93	0.	
242.	94	0.	
234.	95	0.	
227.	96	0.	
221.	97	0.	
214.	98	0.	
208.	99	0.	
203.	100	0.	
197.	101	0.	
192.	102	0.	
188.	103	0.	
183.	104	0.	
179.	105	85.	III
174.	106	306.	IIIIIIIIIIIIII
170.	107	445.	IIIIIIIIIIIIIIIIIIII
167.	108	91.	IIII
163.	109	93.	IIII
160.	110	0.	
156.	111	0.	

4000.	8	121.	IIIIIIII
3750.	9	221.	IIIIIIIIIIIIII.II
3529.	10	502.	IIIIIIIIIIIIIIIIIIIIIII.IIIIIIIIIIIIIII
3333.	11	661.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
3158.	12	364.	IIIIIIIIIIIIIIIIIIIIIIIIIIII
3000.	13	181.	IIIIIIIIIIIIII
2857.	14	86.	IIIIII
2727.	15	87.	IIIIII
2609.	16	65.	IIII
2500.	17	76.	IIIII
2400.	18	83.	IIIIII
2308.	19	96.	IIIIII
2222.	20	131.	IIIIIIIII
2143.	21	89.	IIIIII
2069.	22	81.	IIIII
2000.	23	65.	IIII
1935.	24	63.	IIII
1875.	25	45.	III
1818.	26	29.	II
1765.	27	30.	II
1714.	28	31.	II
1667.	29	18.	I
1622.	30	9.	
1579.	31	24.	I
1538.	32	0.	
1500.	33	26.	I
1429.	34	16.	I
1364.	35	28.	II
1304.	36	12.	
1250.	37	12.	
1200.	38	13.	
1154.	39	13.	
1111.	40	0.	
1071.	41	0.	
1034.	42	0.	
1000.	43	0.	
968.	44	0.	
938.	45	0.	
909.	46	0.	
882.	47	0.	
857.	48	0.	
833.	49	9.	
811.	50	0.	
789.	51	0.	
769.	52	0.	
750.	53	0.	
714.	54	0.	
682.	55	0.	
652.	56	12.	
625.	57	12.	
600.	58	0.	
577.	59	0.	
556.	60	0.	
536.	61	14.	I
517.	62	44.	III
500.	63	30.	II
484.	64	16.	I
469.	65	32.	II
455.	66	84.	IIIIII
441.	67	0.	
429.	68	18.	I
417.	69	0.	
405.	70	19.	I
395.	71	19.	I
385.	72	0.	
375.	73	41.	II
366.	74	0.	
357.	75	64.	IIII
349.	76	22.	I
341.	77	22.	I
333.	78	45.	III
326.	79	47.	III
319.	80	24.	I
313.	81	24.	I
306.	82	25.	I
300.	83	25.	I
294.	84	26.	I
288.	85	105.	IIIIIII
283.	86	54.	III
278.	87	0.	
273.	88	55.	IIII
268.	89	113.	IIIIIIII
263.	90	29.	II
259.	91	0.	
254.	92	60.	IIII
250.	93	0.	
242.	94	0.	
234.	95	32.	II
227.	96	0.	
221.	97	0.	
214.	98	0.	
208.	99	0.	
203.	100	0.	
197.	101	0.	
192.	102	40.	II
188.	103	0.	
183.	104	0.	
179.	105	298.	IIIIIIIIIIIIII.IIIIIII
174.	106	130.	IIIIIIIII
170.	107	311.	IIIIIIIIIIIIIIIIIIIIIII
167.	108	46.	III
163.	109	47.	III
160.	110	0.	
156.	111	0.	

Fig 7. Histogrammes du i à gauche pris à la sortie des neurones, à droite, pris sur la membrane basilaire

[illegible]	[illegible]	[illegible]	
2500.	17	6.	
2400.	18	3.	
2308.	19	278.	IIIIIIIIIIII
2222.	20	853.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2143.	21	159.	IIIIIII
2069.	22	89.	IIII
2000.	23	524.	IIIIIIIIIIIIIIIIIIIIIIII
1935.	24	1027.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
1875.	25	555.	IIIIIIIIIIIIIIIIIIIIIIIII
1818.	26	134.	IIIIII
1765.	27	39.	I
1714.	28	18.	
1667.	29	14.	
1622.	30	9.	
1579.	31	0.	
1538.	32	0.	
1500.	33	10.	
1429.	34	0.	
1364.	35	0.	
1304.	36	0.	
1250.	37	0.	
1200.	38	0.	
1154.	39	0.	
1111.	40	0.	
1071.	41	0.	
1034.	42	0.	
1000.	43	0.	
968.	44	0.	
938.	45	0.	
909.	46	0.	
882.	47	0.	
857.	48	0.	
833.	49	0.	
811.	50	0.	
789.	51	0.	
769.	52	0.	
750.	53	0.	
714.	54	0.	
682.	55	0.	
652.	56	0.	
625.	57	0.	
600.	58	0.	
577.	59	0.	
556.	60	0.	
536.	61	0.	
517.	62	30.	I
500.	63	0.	
484.	64	0.	
469.	65	0.	
455.	66	0.	
441.	67	0.	
429.	68	0.	
417.	69	18.	
405.	70	38.	I
395.	71	19.	
385.	72	59.	II
375.	73	20.	
366.	74	42.	I
357.	75	21.	I
349.	76	44.	II
341.	77	155.	IIIIIII
333.	78	137.	IIIIII
326.	79	186.	IIIIIIII
319.	80	190.	IIIIIIII
313.	81	170.	IIIIIII
306.	82	124.	IIIII
300.	83	277.	IIIIIIIIIIII
294.	84	205.	IIIIIIIII
288.	85	53.	II
283.	86	161.	IIIIIII
278.	87	54.	II
273.	88	111.	IIIII
268.	89	170.	IIIIIII
263.	90	29.	I
259.	91	29.	I
254.	92	0.	
250.	93	61.	II
242.	94	63.	II
234.	95	0.	
227.	96	0.	
221.	97	0.	
214.	98	0.	
208.	99	36.	I
203.	100	75.	III
197.	101	192.	IIIIIIII
192.	102	356.	IIIIIIIIIIIIIIII
188.	103	607.	IIIIIIIIIIIIIIIIIIIIIIIIIIII
183.	104	827.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
179.	105	804.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
174.	106	43.	II
170.	107	177.	IIIIIIII
167.	108	45.	II
163.	109	46.	II
160.	110	94.	IIII
156.	111	0.	
153.	112	50.	II
150.	113	51.	II
144.	114	107.	IIII
139.	115	221.	IIIIIIIIII
134.	116	56.	II
129.	117	0.	
125.	118	0.	
121.	119	0.	
117.	120	0.	

[illegible]	[illegible]	147.	IIIII
2500.	17	186.	IIIIII
2400.	18	210.	IIIIIII
2308.	19	302.	IIIIIIIIII
2222.	20	358.	IIIIIIIIIIII
2143.	21	310.	IIIIIIIIII
2069.	22	343.	IIIIIIIIIII
2000.	23	381.	IIIIIIIIIIIII
1935.	24	295.	IIIIIIIIII
1875.	25	150.	IIIII
1818.	26	117.	IIII
1765.	27	134.	IIII
1714.	28	102.	III
1667.	29	77.	II
1622.	30	52.	I
1579.	31	43.	I
1538.	32	39.	I
1500.	33	26.	
1429.	34	48.	I
1364.	35	45.	I
1304.	36	23.	
1250.	37	24.	
1200.	38	25.	
1154.	39	7.	
1111.	40	14.	
1071.	41	7.	
1034.	42	7.	
1000.	43	0.	
968.	44	8.	
938.	45	0.	
909.	46	8.	
882.	47	0.	
857.	48	0.	
833.	49	0.	
811.	50	0.	
789.	51	0.	
769.	52	10.	
750.	53	0.	
714.	54	0.	
682.	55	0.	
652.	56	0.	
625.	57	12.	
600.	58	0.	
577.	59	0.	
556.	60	41.	I
536.	61	28.	
517.	62	15.	
500.	63	15.	
484.	64	16.	
469.	65	16.	
455.	66	50.	I
441.	67	0.	
429.	68	18.	
417.	69	37.	I
405.	70	38.	I
395.	71	19.	
385.	72	59.	II
375.	73	61.	II
366.	74	0.	
357.	75	21.	
349.	76	109.	III
341.	77	22.	
333.	78	68.	II
326.	79	47.	I
319.	80	95.	III
313.	81	72.	II
306.	82	74.	II
300.	83	0.	
294.	84	26.	
288.	85	26.	
283.	86	27.	
278.	87	54.	I
273.	88	84.	II
268.	89	28.	
263.	90	115.	IIII
259.	91	117.	IIII
254.	92	59.	II
250.	93	30.	I
242.	94	31.	I
234.	95	32.	I
227.	96	33.	I
221.	97	103.	III
214.	98	70.	II
208.	99	73.	II
203.	100	225.	IIIIIII
197.	101	193.	IIIIII
192.	102	198.	IIIIII
188.	103	1379.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
183.	104	1204.	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
179.	105	297.	IIIIIIIIII
174.	106	130.	IIII
170.	107	178.	IIIIII
167.	108	92.	III
163.	109	47.	I
160.	110	47.	I
156.	111	48.	I
153.	112	49.	I
150.	113	52.	I
144.	114	160.	IIIII
139.	115	166.	IIIII
134.	116	0.	
129.	117	0.	
125.	118	0.	
121.	119	0.	
[illegible]	[illegible]	0.	

Fig 8. Histogrammes du y à gauche pris à la sortie des neurones, à droite, pris sur la membrane basilaire.

BIBLIOGRAPHIE

(1) Transformation of sound pressure level from the free field to the eardrum in the horizontal plane.

E.A.G. SHAW JASA vol. 56 n° 6

(2) Basilar membrane and middle ear vibration in guinea pig measured by capacitive probe.

J.P. WIBSON, J.R. JONHSTONE JASA vol. 57 n° 3

(3) Determination of the transfert function of the external ear by an response impulse measurement.

V. MELLERT, K. SIEBRANE, S. MEHRGAARDT JASA vol. 56 n° 6

(4) Tympanis muscle effects on middle ear transfert caracteristic.

A.L. WUTTAL JASA vol. 56 n° 4

(5) Adaptation cochléaire à l'intensité sonore.

F. ROBERT "Acta Medica Belgica" (1960)

(6) On the nonmonotonic behavior of cubis distorsion products in the numan ear.

R. WEBERT, V. MELLERT JASA vol. 57 n° 1

(7) Response patterns of cochlear nucleus neurons ti excerpts from sustained vowels.

T.J. MOORE, J.L. CASHIN JASA vol. 56 n° 5

(8) Neural coding and psychophysical discrimination data.
R. DUNCAN LUCE, DAVID M. GREEN JASA vol. 56 n° 5

(9) Experiments in hearling.

V. BEKEZY

(10) Neural mechanism of the auditory and vestibular system.

RASMUSSEN (Sprinfield Illinois)

(11) Organisation nerveuse et sensorielle de l'homme.

J. TUSQUES Maloine S.A. Editeurs

(12) Les fondements cybernétiques de l'activité nerveuse.

BALCEANU NICOLAU (expansion scientifique française)

(13) Un modèle mathématique de cochlée et son application à l'analyse du signal vocal.

J. CAELEN, Thèse Doct. Ingénieur Toulouse (1974)
Rapport Interne (février 1974)

(14) A propos de marqueurs lexicosyntaxiques.

D. DOURS, R. FACCA, G. MAURAND, G. PERENNOU
6è JEP TOULOUSE (1975)

(15) Etude des phonèmes de la langue française au moyen d'une cochlée artificielle. Application à la reconnaissance de la parole.

A. ALINAT Revue tech. Thomson CSF vol. 7 n° 1 (1975)

(16) Zones fréquentielles et reconnaissance des voyelle du Français.

A. LANDERCY, R. RENARD Revue de Phonétique APpliquée Mons 1975

(17) Perception des voyelles françaises filtrées.

A. LANDERCY

(18) Critical band width in loudness summation.

E. ZWICKER, G. FLOTTORP, S. STEVENS JASA vol. 29 n° 5

(19) Subdivisions of the audible frequency range.

E. ZWICKER JASA vol. 33 n° 2

(20) Localisation of non-linearisties in the cochlea

M.A. VERGEVER, J.J. KALKER Journal of engineering mathematics vol. 9 n° 1 (1975)

(21) J. CAELEN - Un modèle mathématique de cochlée
5èmes Journées d'Etude sur la Parole - ORSAY - 1974 -

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

ANALYSE COMPARATIVE DU SIGNAL DE LA PAROLE

G. CARAYANNIS

RESUME : On définit par analyse comparative du signal, l'analyse d'un segment lorsqu'il est tenu compte de certaines informations concernant le segment voisin. Deux méthodes d'analyse comparative sont suggérées. Sur les régions stables des phonèmes, les résultats fournis par ces analyses tendent vers les résultats d'une analyse normale. Par contre, sur les régions de transitions, les paramètres obtenus peuvent être utilisés pour la segmentation de la parole.

SUMMARY :The comparative analysis of the signal is defined as the analysis of a short-time segment where information concerning the neighbourhood segment is included. Two methods of comparative analysis are presented here. On the stable regions of phonemes, results given by these analysis methods are equivalent to a classic analysis. On the transitions regions however the parameters obtained can be used for speech segmentation.

ANALYSE COMPARATIVE DU SIGNAL DE LA PAROLE

G. CARAYANNIS

1. INTRODUCTION.

Le but d'une analyse comparative du signal de la Parole peut être double. a) Extraire un certain nombre de paramètres qui caractérisent un segment par rapport à un segment voisin.

b) Se servir de ces paramètres pour pouvoir segmenter le signal. Les paramètres extraits ainsi ne représentent pas seulement certains aspects du signal, mais ils caractérisent en plus la transition. Cette caractérisation de la transition (d'une manière abstraite) est envisagée ici de deux façons différentes :

1) Par la fonction d'intercorrélation de deux signaux voisins.

2) En introduisant un critère discriminant.

Dans les deux cas un modèle linéaire est utilisé. En ce qui concerne la première méthode, on utilise les relations relativement simples, qui existent entre la fonction d'autocorrélation et la fonction d'intercorrélation de deux signaux moyennant un modèle linéaire. On démontre qu'une estimation de ce modèle peut être obtenue à partir de la résolution d'un système d'équations linéaires. La similitude de ce système avec les équations d'autocorrélation de la prédiction linéaire est notée. Les propriétés des matrices de Toeplitz peuvent être utilisées dans le cadre de cette étude pour obtenir la solution rapidement. La deuxième méthode étudiée agit un peu différemment. On cherche un modèle linéaire qui soit aussi près que posvible du premier signal et aussi loin que possible du second. On arrive ainsi à se servir des propriétés du second signal pour normaliser les paramètres du premier. Ce type de normalisation donne un caractère discriminant à l'algorithme.

Les méthodes utilisées jusqu'à présent pour la segmentation de la parole agissent de deux façons :

a) Soit elles procèdent sur la forme du signal lui-même à l'aide de critères simples et arrivent à des performances comparables à celles de l'oeil humain (/1/, /2/).

b) Soit elles procèdent sur les paramètres qui représentent les signaux (après analyse) (/3/).

Les méthodes introduites ici diffèrent des précédentes, étant donné que le traitement porte non pas sur un seul segment mais sur le degré de

similitude de deux segments voisins exprimé par leurs fonctions d'auto et inter-corrélation.

2. AUTOCORRELATION ET INTERCORRELATION

Une interprétation parmi les plus simples de l'intercorrélation de deux processus stochastiques : (s (t), s' (t)) est celle qui peut être obtenue en considérant s (t) comme l'entrée d'un système linéaire dont la sortie est s' (t) :

$$s'(t) = \int_0^\infty h(u)\, s(t-u)\, du + e(t) \qquad (1)$$

Le modèle discret équivalent est :

$$s'_n = \sum_{r=0}^{\infty} h_r\, s_{n-r} + e_n \qquad (2)$$

Supposons les deux processus à moyenne nulle et soit r_k, la fonction d'autocorrélation du premier r'_k celle du second et c_k leur fonction d'inter-corrélation, k = 0,.....,p .On aura les relations suivantes :

$$r_k = r_{-k} = E\,(s_n\, s_{n-k}) \qquad (3)$$

$$r'_k = r'_{-k} = E\,(s'_n\, s'_{n-k}) \qquad (4)$$

$$c_k = c_{-k} = E\,(s_n\, s'_{n-k}) \qquad (5)$$

A partir de (2) et (5) on obtient :

$$c_k = \sum_{r=0}^{\infty} h_r\, r_{k-r} \qquad k = 0, \pm 1, \ldots\ldots \qquad (6)$$

$$r'_k = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} h_r h_s\, r_{k+r-s} + r^e_k \qquad (7)$$

Pour r = fini = p les relations (6) et (7) s'écrivent sous forme matricielle :

$$\underline{c} = R\,\underline{h} \qquad (6a)$$

$$r'_k = \underline{h}^T R^{(k)}\, \underline{h} \qquad (7a)$$

où

$$\underline{c}^T = \left[c_1, c_2, \ldots\ldots, c_p \right] \qquad (8)$$

$$\underline{h}^T = \left[h_1, h_2, \ldots\ldots, h_p \right] \qquad (9)$$

R est la matrice d'autocorrélation du signal s_n ayant la structure de Toeplitz.

$$R = \begin{bmatrix} r_0 & r_1 & \cdots & r_{p-1} \\ r_1 & r_0 & \cdots & \\ \vdots & \vdots & \ddots & \\ r_{p-1} & \cdots & \cdots & r_0 \end{bmatrix} \qquad R^{(k)} = \begin{bmatrix} r_k & r_{k+1} & \cdots & r_{k+p-1} \\ \vdots & & & \vdots \\ r_{k+p-1} & \cdots & \cdots & r_k \end{bmatrix} \qquad (10)$$

Ainsi le modèle régressif linéaire peut être déterminé à partir du système linéaire (6a). Ce système ressemble beaucoup à celui utilisé par la prédiction linéaire pour déterminer un modèle autorégressif :

$$\underline{r} = R\,\underline{a} \qquad (11)$$

où $\underline{r}^T = [r_1, \ldots, r_p]$ et $\underline{a}^T = [a_1, \ldots, a_p]$ (12)

L'équation (11) peut être vue comme la limite vers laquelle tend (6a) quand $s_n \rightarrow s'_n$

En effet : $\lim_{s_n \rightarrow s'_n} \{\underline{c}\} = \underline{r}$ ainsi $\lim_{s_n \rightarrow s'_n} \{\underline{h}\} = \underline{a}$

Cette propriété sera utilisée pour la segmentation comme on le verra par la suite.

3. MISE EN OEUVRE DU PREMIER SEGMENTATEUR.

La résolution de l'équation (6a) est très facile dans la pratique vu la structure de la matrice R. Le principe de la segmentation repose sur la résolution simultanée de deux équations suivantes :

$$\underline{c} = R\,\underline{h} \qquad \hat{\underline{h}} = R^{-1}\,\underline{c}$$

$$\underline{r} = R\,\underline{a} \qquad \hat{\underline{a}} = R^{-1}\,\underline{r}$$

Si R est une matrice de Toeplitz on peut démontrer que :

$$R^{-1} = W^T \hat{D}\, W \qquad (13)$$

où

$$W = \begin{bmatrix} 1 & & & \\ a_{11} & 1 & \mathbf{0} & \\ a_{12} & a_{22} & 1 & \\ \vdots & \vdots & & \\ a_{1p} & a_{2p} \cdots & & 1 \end{bmatrix}$$

et $\hat{D} = \text{diag.}\ (1/\alpha_0\ ,\ 1/\alpha_1\ ,\ \ldots\ldots,\ 1/\alpha_p)$

où a_{ij} sont les coefficients de prédiction linéaire pour des systèmes d'ordre croissant et α_j les erreurs de prédiction correspondantes (/4/).

Ainsi la résolution de deux équations (6a) et (11) peut se faire simultanément en utilisant l'algorithme de Levinson.

Le critère utilisé pour la segmentation est le suivant :

$$\gamma = (\underline{h} - \underline{a})^T\ (\underline{h} - \underline{a})$$

La méthode peut s'intégrer facilement dans un programme d'analyse par prédiction linéaire, le seul calcul supplémentaire important étant celui de p termes d'intercorrélation.

Un pas de glissement de la fenêtre temporelle égal à 10 ms a été adopté dans la pratique.

Comparée à une procédure basée sur le filtre de Kalman multi-dimensionnel (/3/), la méthode décrite ici présente les avantages suivants :

1) Elle est beaucoup plus rapide

2) Elle fournit des bons résultats de segmentation même sur les transitions rapides étant donné que le problème du délai d'intégration des connaissances inhérant au filtre de Kalman n'existe plus.

4. INTRODUCTION D'UN CRITERE DISCRIMINANT.

Il serait intéressant d'essayer de modéliser un segment du signal de la parole en tenant compte de l'existence du segment voisin. Ainsi un modèle qui est aussi près que possible du premier signal et aussi loin que possible du second pourrait être recherché. Le schéma fonctionnel d'une telle modélisation est donné sur la figure 1.

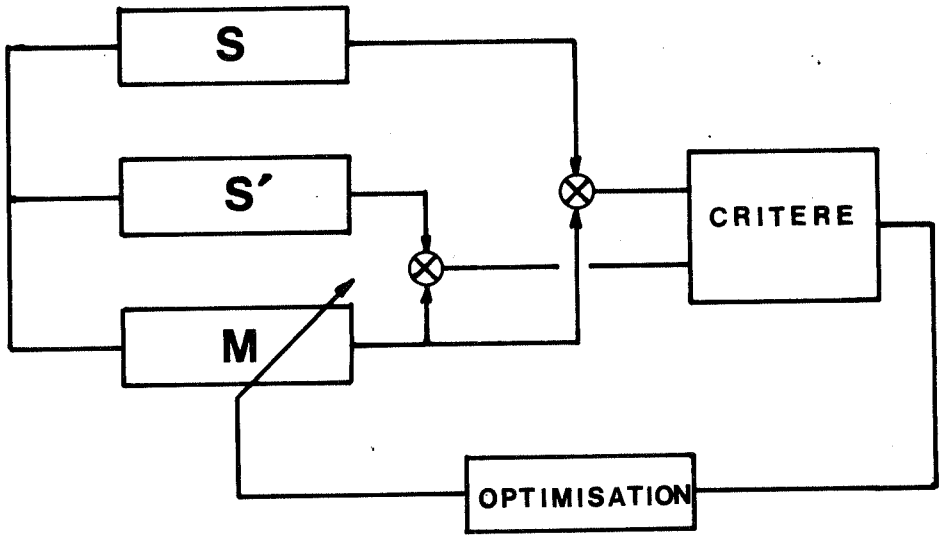


Figure 1.

On sait que l'erreur de représentation du système S par le modèle M est donné par : (/4/)

$$\sigma^2 = \underline{a}^T R \underline{a} \qquad (14)$$

où R est la matrice de corrélation du premier signal.

D'autre part l'erreur de représentation du système S' par le même modèle est donné par :

$$\sigma'^2 = \underline{a}^T R' \underline{a} \qquad (15)$$

On peut alors exiger que la première quantité devienne minimale sous la condition que la seconde soit constante. Ce problème de minimisation sous contraintes peut être résolu en faisant appel au Lagrangien

$$L = \underline{a}^T R \underline{a} - (\underline{a}^T R' \underline{a} - 1) \qquad (16)$$

Pour l'optimum les équations sont les suivantes :

$$R \hat{\underline{a}} = \lambda_{min} R' \hat{\underline{a}} \qquad (17)$$

Ce système correspond à un problème à valeurs propres généralisé. Il peut s'écrire autrement :

$$R'^{-1} R \hat{\underline{a}} = \lambda_{min} \hat{\underline{a}} \qquad (18)$$

Sous cette forme il peut être vu comme un problème à valeurs propres simple; la matrice à diagonaliser étant $R'^{-1} R$.

On aura d'autre part les équations suivantes :

$$A^T R A = \Lambda \qquad (19)$$

$$A^T R' A = I \qquad (20)$$

$$\Lambda = \text{diag} (\lambda_1, \ldots, \lambda_p) \qquad (21)$$

où A est la matrice principale du faisceau des formes $\underline{x}^T R \underline{x} - \lambda \underline{x}^T R' \underline{x}$ ayant comme colonnes les vecteurs propres correspondants (/6/, /7/).

On sait, d'après la théorie des faisceaux des formes que la transformation :

$x = A\xi$ réduit simultanément les formes

$\underline{x}^T R \underline{x}$ et $\underline{x}^T R' \underline{x}$ aux sommes de carrés :

$$\sum_{k-1}^{p} \lambda_k \xi_k^2 \quad \text{et} \quad \sum_{k=1}^{p} \xi_k^2$$

et que : $\lambda_{min} \leqq \dfrac{\underline{a}^T R \underline{a}}{\underline{a}^T R' \underline{a}} \leqq \lambda_{max}$ (rapport de Rayleigh borné par les valeurs propres)

5. MISE EN OEUVRE DU SECOND SEGMENTATEUR.

Le segmentateur fonctionne en testant soit la matrice $R'^{-1} R$ soit λ_{min}. Nous avons les relations suivantes :

$$\lim_{s_n \to s'_n} (R'^{-1} R) = I \qquad \text{et} \qquad \lim_{s_n \to s'_n} (\lambda_{min}) = 1$$

l'inverse de la matrice R' est facile à calculer vu sa structure de Toeplitz (voir § 3).

Si l'on procède à une analyse de la parole par prédiction linéaire, les matrices R et R' ainsi que l'inverse R'^{-1} sont de toute façon calculées. La seule opération supplémentaire à effectuer est une multiplication matricielle. On cherche actuellement à modifier la méthode pour :

1) Tenir compte de l'intercorrélation de deux signaux
2) Arriver à une solution raisonnable sur les régions stationnaires.

Ceci est possible si à la place de (18) on résoud une équation du type

$$(R'^{-1} C) R \underline{a} = \lambda \underline{a} \qquad (22)$$

où C est la matrice d'intercorrélation de deux signaux.

Si l'on pose $R'^{-1} C = H^{-1}$ la contrainte utilisée est du type $\underline{a}^T H \underline{a} = 1$. La signification physique de cette relation reste à trouver. La matrice H est probablement liée au modèle (6a) (relation intercorrélation-autocorrélation).

6. CONCLUSION.

Deux méthodes d'analyse comparative du signal ont été présentées. Elles peuvent être mises en oeuvre en introduisant quelques instructions supplémentaires dans un programme d'analyse par prédiction linéaire. Ainsi en analysant le signal on peut arriver à le segmenter au même temps, en utilisant l'information sur les différences de deux classes relatives aux deux signaux voisins.

REFERENCES

/1/ D.R. REDDY : "Segmentation of speech sounds".
J. Acoust. Soc. Am., 40, 1966, 307-312.

/2/ D.R. REDDY, P.J. VICENS : "A procedure for segmentation of connected speech".
J. Audio Eng. Soc, 1968.

/3/ G. CARAYANNIS : "Modélisation des Transitions phonémiques. Application à la segmentation de la parole".
4èmes Journées du Groupe de la Communication Parlée du GALF - Bruxelles 1973.

/4/ G. CARAYANNIS : Modélisation des signaux pour Transformation Karhumen-Loëve. Application à l'Analyse de la Parole".
Rapport d'Activités de l'Institut de Phonétique de Bruxelles, 9/2, 1975, 97-108.

/5/ K.P. LI, G.W. HUGHES, T.B. SNOW : "Segment classification in continuous speech".
IEEE Trans. on Audio, 20, 1972, 142-150.

/6/ F.R. GANTMACHER : "Théorie des matrices", Tome 1, Dunod, Paris 1966.

/7/ F. REZA : "Linear spaces in Engineering"
Ginn and C°, 1971.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

MODELISATION DU SIGNAL DE PAROLE PAR DIAGONALISATION DE SA MATRICE D'AUTOCORRELATION.

G. CARAYANNIS et C. GUEGUEN.

RESUME : Le présent travail introduit un nouveau type de "prédiction linéaire" basée sur l'expansion Karhumen-Loëve de la matrice de corrélation des échantillons de la parole. Ce résultat est obtenu par une nouvelle normalisation des paramètres. On peut démontrer que grâce à certaines propriétés très importantes des matrices de Toeplitz, les pôles du modèle autorégressif se trouvent sur la circonférence du cercle unitaire. En conséquence, seules les fréquences des formants sont calculées et le résultat peut être interprété comme une transformée de Fourier spéciale.

L'application à l'analyse de la parole est développée en comparant avec la méthode classique de prédiction linéaire et la méthode de Cepstre.

SUMMARY : This paper introduces a new type of "linear prediction", method based on the Karhumen-Loëve expansion of the correlation matrix of the speech samples. This result is obtained via a new normalization of the parameters.

It is shown that, due to some important properties of Toeplitz matrices, the poles of the autoregressive model lie on the unit circle. Consequently, only the formant frequencies are computed and the result can be interpreted as a special discrete Fourier transform.

Application to speech analysis is developed with a comparison to the usual linear prediction and cepstrum methods.

MODELISATION DU SIGNAL DE PAROLE PAR DIAGONALISATION DE SA MATRICE D'AUTOCORRELATION.

G. CARAYANNIS et C. GUEGUEN.

INTRODUCTION.

La modélisation du signal utilisant la prédiction linéaire s'est révélée très efficace pour l'analyse de la parole. La valeur s_n de l'échantillon n est prédite par une combinaison linéaire d'un certain nombre de valeurs précédentes:

$$\hat{s}_n = - \sum_{i=1}^{p} a_i s_{n-i} \qquad e_n = s_n - \hat{s}_n \quad (1)$$

Les coefficients de prédiction sont calculés par minimalisation de l'erreur quadratique moyenne σ^2 sous une fenêtre temporelle.

Soit $\underline{a}$ le vecteur des coefficients de prédiction.

($a_0 = 1$, inclus), le critère est facilement exprimé par

$$\sigma^2 = \underline{a}^T R \underline{a}$$

avec $\underline{a}^T = [a_0, a_1, \ldots, a_p]$ (2)

où R est la matrice de corrélation des échantillons de la parole, R peut être considérée comme ayant la structure de Toeplitz avec des éléments r_i, $i = 0, \ldots, p$

$$R = E(\underline{s}_n \underline{s}_n^T)$$

où $\underline{s}_n = [s_n, s_{n-1} \ldots\ldots s_{n-p}]$ (3)

Dans ce cas, les coefficients de prédiction inconnus, sont solution de l'équation de Yule-Walker /3/ :

$$\underline{a}^T R = [\sigma^2, 0 \ldots\ldots 0] \quad (4)$$

Plusieurs méthodes ont été développées pour résoudre cette équation. Les coefficients d'autocorrélation r_i peuvent être estimés sous une fenêtre temporelle, et des algorithmes rapides tenant compte de la structure spéciale de R peuvent être appliqués (/1/, /2/, /3/). Mais le même problème peut être formulé aussi comme un filtre de Kalman dégénéré. La possibilité d'introduire des non-stationnarités ainsi que l'aspect itératif de ce filtre demeurent toujours d'un grand intérêt pour l'analyse du signal (/4/, /5/, /6/).

Toutes ces approches ont déjà donné des résultats importants pour la parole et d'autres signaux (comme le EEG), mais il a toujours été noté que si le calcul des fréquences des formants est correct, il n'en est pas toujours de même pour les amortissements.

Etant donné que pour plusieurs applications ces amortissements

sont dépourvus d'intérêt, il est intéressant d'élaborer sur la même base une méthode spéciale pour calculer seulement les fréquences de formants. Ceci est possible en introduisant une autre normalisation pour les coefficients de prédiction.

2. FORMULATION DE LA METHODE.

L'hypothèse principale du schéma de la prédiction linéaire est que le signal s_n est la réponse d'un modèle autorégressif excité par une séquence de bruit pseudo-blanc suivant la relation :

$$a_0 s_n + a_1 s_{n-1} + \ldots\ldots a_p s_{n-p} = e_n \qquad (5)$$

a_0 est pris égal à 1 pour assurer une solution non-triviale à l'équation (4).

Nous introduisons maintenant une autre normalisation raisonnable du vecteur des paramètres par la contrainte : (/7/)

$$\sum_{i=0}^{p} a_i^2 = 1 \quad \text{ou} \quad \underline{a}^T \underline{a} = 1 \qquad (6)$$

Le nouveau problème à résoudre est celui de l'optimalisation sous contraintes. Il peut être résolu en se servant du Lagrangien :

$$L = \underline{a}^T R \underline{a} - \lambda (\underline{a}^T \underline{a} - 1)$$

on est conduit ainsi aux équations d'optimalité suivantes :

$$R \underline{a} = \lambda \underline{a} \quad \text{et} \quad \text{Min}(\sigma^2) = \lambda \text{ min} \qquad (7)$$

Les coefficients $\underline{a}_i$ sont par conséquent donnés par le vecteur propre $\underline{a}$ qui correspond à la plus petite valeur propre de R.

On se réferera à ce choix particulier comme modélisation linéaire factorielle.

On peut noter que le vecteur $\underline{a}$ peut toujours être ajusté après optimisation pour avoir $a_0 = 1$.

Pour l'erreur e_n appelée jusqu'à présent erreur de prédiction on utilisera un terme plus approprié, celui de l'erreur de représentation.

Cette approche est liée à la méthode bien connue de l'expansion Karhumen-Loëve.

Soit s_n un processus stationnaire discret tel que :

$$E(s_n) = 0 \quad \text{et} \quad E(s_n s_{n-i}) = r_i$$

Une manière de représenter ce processus aléatoire consiste à opérer une transformation orthogonale qui fournit des variables non-corrélées avec variance λ_i.

Si λ_i sont les valeurs propres et $\underline{a}_i$ les vecteurs propres correspondants R peut s'écrire sous la forme :

$$R = \sum_{i=0}^{p} \lambda_i \, a_i \, a_i^T \qquad (8)$$

Pour l'analyse en composantes principales, le vecteur aléatoire $\underline{s}_n$ est représenté d'une manière satisfaisante dans le sous-espace défini par les a_i associés aux valeurs propres les plus importantes de R. Cette propriété peut être exploitée pour visualiser la parole.

Ici, par contre, l'intérêt sera centré sur les plus petites valeurs propres. Elles correspondent à ceux parmi les vecteurs $\underline{a}_i$ qui sont les plus orthogonaux en moyenne à $\underline{s}_n$. Les différents modèles déduits aboutissent à une erreur de représentation petite.

L'information contenue dans le signal étant résumée par R, tout ensemble de paramètres lié à R est intéressant. C'est le cas pour l'ensemble des λ_i arrangés en ordre décroissant qui contiennent des informations pertinentes.

Une autre méthode pour réduire une forme quadratique en une somme de carrés, est celle de triangularisation de Cholensky.

Les termes diagonaux α_i qui sont liés à des modèles d'ordre croissant sont aussi intéressants pour la représentation du signal.

3. PROPRIETES FONDAMENTALES DE LA MODELISATION LINEAIRE FACTORIELLE.

Plusieurs propriétés intéressantes de la méthode sont liées à la structure de la matrice de corrélation (Toeplitz). Soit J un opérateur linéaire qui renverse l'ordre des lignes (respect. colonnes) d'une matrice donnée quand il est appliqué à gauche (resp. droite). La propriété d'invariance de R avec cette transformation :

$$J \, R \, J = R \qquad (9)$$

a comme résultat une structure spéciale pour le vecteur propre $\underline{a}$: (/8/)

$$\underline{a} = \pm J \, \underline{a} \qquad (10)$$

Cette propriété implique que les racines du polynôme A(z)

$$A(z) = \sum_{i=0}^{p} a_i \, z^{-i}$$

(qui sont les pôles du modèle), se situent sur la circonférence du cercle unitaire /7/.

Le spectre déduit par ce modèle sans pertes est par conséquent un ensemble de j contributions à des fréquences discrètes ψ_i. Ainsi le signal original peut être reconstruit comme une somme de p/2 sinusoïdes :

$$s_n = \sum_{i=0}^{p/2} \gamma_i \cos (n \psi_i)$$

La contribution de chaque fréquence peut être calculée à partir du système de Van der Monde :

$$r_j = \sum_{i=0}^{p/2} \frac{\gamma_i^2}{2} \cos (j \psi_i) \qquad (11)$$

L'équation (11) peut être interprétée comme une transformée de Fourier inverse qui est calculée sur un petit nombre des points ψ_i à la place de fréquences équi-réparties comme on le fait usuellement.

La relation de la méthode avec la prédiction linéaire classique repose sur la valeur du déterminant de R. Si $\lambda_{min} = 0$ les deux solutions coïncident. Ce modèle factoriel peut être interprété comme le cas limite de la prédiction linéaire quand la distance modèle-système diminue (ordre plus élevé par exemple).

L'ordre en effet est une caractéristique importante de la modélisation. L'effet blanchisseur sur l'erreur de prédiction est basé sur l'hypothèse que l'ordre est correct.; c'est-à-dire que r_i satisfait l'équation du modèle pour i>p.

L'hypothèse correspondante ici est que r_p est tel que le déterminant de R est égal à zéro. Il y a deux valeurs distinctes de r_p qui assurent cette propriété.

Les deux solutions correspondantes peuvent être combinées de façon à reconstruire la prédiction linéaire usuelle d'ordre p-1 (/9/).

4. APPLICATION A L'ANALYSE DE LA PAROLE.

Comme il a été noté précédemment, la précision de la représentation dépend essentiellement de la plus petite valeur propre de R.

La figure (1) montre les λ_i pour une voyelle française (/a/) sous une fenêtre de Hamming. Les résultats avec pré-emphase (a) et sans

pré-emphase (b) sont comparés. Comme il a été déjà évoqué , la pré-emphase limite la variabilité des λ_i et suivant le tableau 1 la matrice de corrélation est mieux conditionnée.

sans pré-emphase (b)	avec pré-emphase (a)
det R = $0.114.10^{-10}$	det R = $0.123.10^{-3}$
$\lambda_0 = 4.613, \lambda_{12} = 0.006$	$\lambda_0 = 2.2952, \lambda_{12} = 0.083$

Tableau 1

L'effet de la pré-emphase peut aussi être observé au niveau des termes diagonaux α_i de la matrice triangulaire de Cholensky (il s'agit des erreurs de prédiction pour des modèles d'ordre croissant). Les résultats sont illustrés à la figure 2.

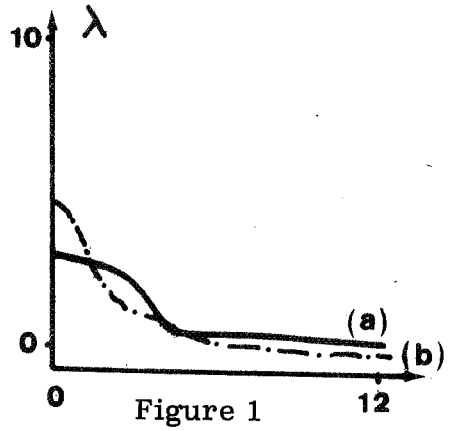


Figure 1

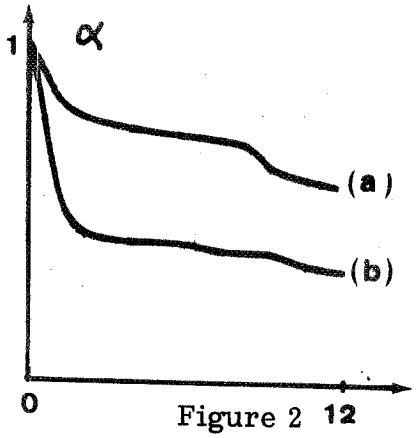


Figure 2

Concernant la méthode de "modélisation factorielle", la pré-emphase n'est pas nécessaire quand l'ordre du modèle est plus petit ou égal à l'ordre réel du système. Pourtant on a remarqué que la pré-emphase exerçait un rôle favorable pour obtenir une bonne précision de détection formantique pour les ordres supérieurs.

Sur la figure 3 on voit une comparaison des raies spectrales obtenues pour les deux valeurs extrêmes de λ, λ_{max} et λ_{min}. pour une voyelle française. Les écarts fréquenciels dûs à des erreurs de représentation très importantes (λ_{max}) sont illustrés. Par contre pour tous les λ_i de valeur proche à la valeur minimale ($\lambda_i \simeq \lambda_{min}$) la position des raies ne varie pas beaucoup.

La figure 4 nous donne une comparaison des différentes méthodes de modélisation : prédiction linéaire en utilisant le filtre de Kalman (courbe LP), enveloppe du spectre déduite après déconvolution par le cepstre (courbe CEP), modélisation factorielle basée sur l'expansion karhumen-Loëve (raies KL). Les raies de KL ne coïncident pas partout avec les résultats de la prédiction linéaire. On a remarqué beaucoup de coïncidences avec la méthode de Cepstre (voir figure 4). D'une manière générale en faisant augmenter l'ordre du système du modèle KL les nouveaux pics que l'on obtient tendent à coïncider avec les pics supplémentaires obtenus en faisant augmenter le nombre de canaux du cepstre.

L'ordre du système a une grande importance pour ce modèle factoriel parce qu'il est lié directement au nombre de pics.

L'influence de l'ordre au niveau des fréquences de formants obtenues est montré sur le tableau 2 et comparée aux valeurs respectives obtenues par prédiction linéaire (pas de pré-emphase).

		8	10	12
F 1	L - P	526	659	663
	K - L	470	794	716
F 2	L - P	1436	1471	1467
	K - L	1228	1510	1464
F 3	L - P	2685	2570	2519
	K - L	2070	2552	2364
F 4	L - P	3740	3703	3648
	K - L	3529	3652	3671

Tableau 2

Figure 3.

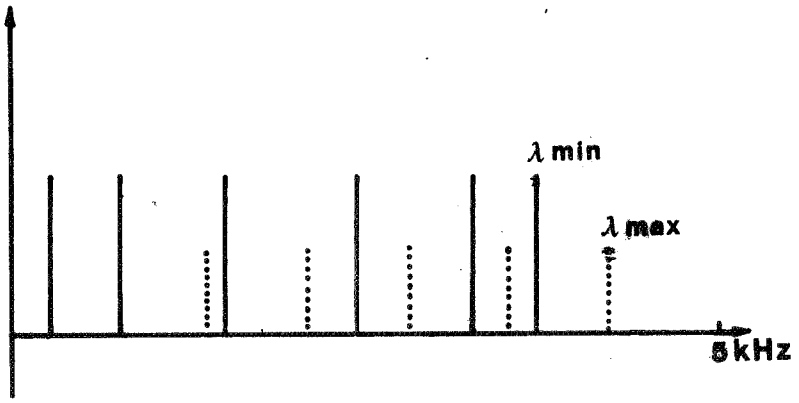
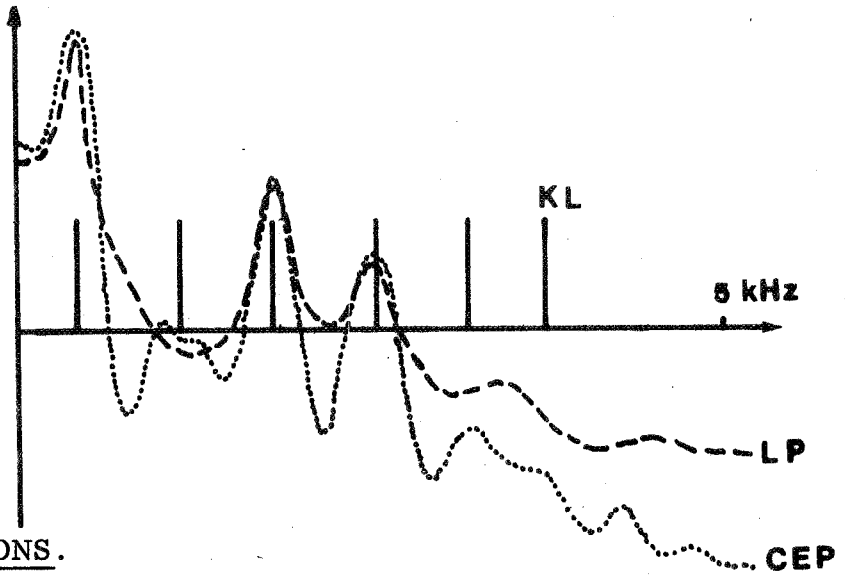


Figure 4.



5. CONCLUSIONS.

Le modèle linéaire factoriel , présenté ici comme technique KL calcule sur un signal donné un petit nombre de fréquences constituantes. Il correspond à un modèle sans pertes. Ayant appliqué ce modèle au signal de la parole, on a pu extraire les fréquences des formants. Comparé à la prédiction linéaire le calcul de p/2 paramètres est seulement exigé, mais ces calculs sont ici un peu plus longs (extraction des vecteurs propres). Une solution récurrente est actuellement à l'étude. La contribution de chaque fréquence peut être calculée aussi, donnant naissance à p/2 nouveaux paramètres. Ainsi la synthèse du signal est facile en faisant tout simplement la somme des p/2 signaux sinusoïdaux pondérés.

La méthode donne de très bons résultats qui peuvent être comparés au cepstre avec l'avantage de la modélisation. Elle peut être considérée comme une transformation du type Chirp-z, favorisant la séparation des fréquences voisines.

Mais surtout une importante liaison a été constatée entre la transformée de Fourier et l'identification d'un modèle sans pertes.

REFERENCES.

/1/ F. ITAKURA, S. SAITO : Analysis synthesis telephony based upon the maximum likehood method.
6th Int. Congr. Acoust. Paper C.5.5., Tokyo, 1968.
/2/ J. MARKEL, A. GRAY : On autocorrelation equations as applied to speech analysis.
IEEE Trans. on Audio., Vol. AC-21, n° 2, 1973.
/3/ J. MAKHOUL : Linear Prediction : a tutorial review.
Proceedings of the IEEE, Vol. 63, n° 4, 1975.
/4/ C. GUEGUEN, G. CARAYANNIS : Analyse de la parole par filtrage optimal de Kalman.
Automatisme, Tome 18, n° 3, Dunod, 1973.
/5/ G. CARAYANNIS : Analyse de la Parole par identification récurrente d'un modèle du système de phonation.
Thèse de Docteur Ingénieur, Univ. Paris 7 , 1973.
/6/ J. GIBSON, J. MELSA, S. JONES : Digital speech analysis using sequential estimation techniques.
IEEE Trans. on ASSP, vol. ASSP-23, n° 4, 1975
/7/ G. CARAYANNIS : Modélisation des signaux par transformation Karhumen-Loëve.
Rapport d'Activités de l'Institut de Phonétique de Bruxelles, Avril 1975.
/8/ C. GUEGUEN : The modified linear prediction : a factor analysis approach to speech analysis.
UCLA Report, Dept. of System Science, UCLA-ENG, 7540, 1975.
/9/ C. GUEGUEN : La prédiction linéaire modifiée et son application à la modélisation du signal de parole.
Colloque National sur le Traitement du Signal et ses Applications, Papier 79, Nice 1975.
/10/ G. CARAYANNIS, C. GUEGUEN : The factorial Linear Modelling A Karhumen-Loëve Approach to Speech Analysis.
1976 IEEE International Conference on Acoustics, Speech and Signal Processing.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

DETECTION DES SIX SONS D'UN VOCABULAIRE ARTIFICIEL

M.CARTIER, J.L.COURBON, R.DAGORNE, R.DOUCEN, J.J. LUCAS
CNET/LANNION

RESUME : On décrit un système permettant d'identifier 6 sons différents. Ces six sons sont groupés dans des mots CVCV. On obtient pour un locuteur quelconque un taux de reconnaissance dépassant 80 % et un taux d'erreur de 2 %. On peut ainsi reconnaître plusieurs dizaines de commandes vocales différentes avec un dispositif peu coûteux.

SUMMARY : Six sounds are detected in a system which is described. CVCV words are built with these sounds. Performed recognition score is greater than 80 % and error rate is about 2 %. This result holds for several male and female speakers. 50-100 oral commands may be separated by such a low cost device.

DETECTION DES SIX SONS D'UN VOCABULAIRE ARTIFICIEL

M. CARTIER, J.L. COURBON, R. DAGORNE, R. DOUCEN, J.J. LUCAS
CNET/LANNION

INTRODUCTION

Quand J. DREYFUS-GRAF proposa l'utilisation de "phonocodes" pour des commandes vocales, les avis furent partagés : certains pensèrent que la reconnaissance de six sons était si aisée que le problème ne méritait même pas d'être étudié, mais l'étude parut intéressante à des personnes suffisamment nombreuses et d'origines diverses (*), aussi avons-nous entrepris la réalisation d'un appareillage(**) permettant des expériences en temps réel et étudié des commandes vocales susceptibles d'être reconnues indépendamment du locuteur, homme ou femme.

Partant de la proposition de SOTINA [1] nous avons conservé la répartition des 3 consonnes et 3 voyelles. Des essais d'analyse et d'écoute [2,3,4] ont conduit à choisir [ʃ] plutôt que [s] et à utiliser [k] ou [t] en fonction de la voyelle suivante. Les essais d'écoute n'ont bien sûr pas révélé de confusion entre [ɲ] et les voyelles. Malgré des difficultés [n] - [i] , nous avons conservé [i] tout en choisissant [o] de préférence à [u] à cause de confusions [n] - [u] : les confusions [n] - [i] peuvent se corriger.

Nous verrons comment on peut, au prix de quelques contraintes de structure, reconnaître soixante-dix mots avec un pourcentage d'erreurs de 2 % et une probabilité de reconnaissance sans répétition de 0,83, ceci pour une grande variété de locuteurs et sans réglage individuel. Quelques perfectionnements sont envisageables : nous examinerons leur intérêt et leur complexité .

PRINCIPES D'ANALYSE ET DESCRIPTION DE L'APPAREILLAGE (fig.1)

Le traitement acoustique des signaux est effectué dans des circuits analogiques dont les sorties sont numérisées et envoyées à un microprocesseur.

(*)contrat CRI n° 73 034

(**) appareil baptisé "CHARLES" (marque déposée !)

1. Partie câblée

a/ Analyse spectrale : un compresseur de dynamique régule le signal avant d'attaquer six filtres [3] . Les caractéristiques du compresseur réalisent un compromis entre réponse dynamique, régulation et distorsion [5] . Une seule boucle est utilisée (commande "avant"). Les constantes de temps ont été déterminées de façon que la phase transitoire sur les attaques ne dure pas plus d'une période de fondamental (< 10 ms) et que le temps de recouvrement soit de 20 à 30 millisecondes. La distorsion est de 10 % à 100 Hz et 3 % à 200 Hz. La dynamique de compression est de 26 dB pour ± 1 dB et 30 dB à - 6 dB.

L'analyse spectrale, qui aboutit à la classification des spectres instantanés toutes les 10 ms, s'inspire du principe des compresseurs sélectifs de DREYFUS-GRAF et des travaux de POLS [6] : elle consiste en une description fondée sur des comparaisons énergétiques entre bandes de fréquence, réalisées par des combinaisons linéaires de logarithmes. La figure 2 précise les relations. On voit qu'un taux de compression R important conduit à des fonctions identiques. Les transformations après filtrage et détection sont réalisées par des amplificateurs logarithmiques fonctionnant dans une dynamique de 50 dB, dont les sorties attaquent des codeurs pour les simulations (256 niveaux de 0,2 dB) et des circuits de sommations algébriques pondérées pour le fonctionnement câblé. Chaque combinaison linéaire est envoyée à un circuit à seuils suivi d'une logique combinatoire. Les sorties binaires (une classe toutes les 5 ms) sont envoyées au microprocesseur.

b/ Détection de plosives : un second compresseur fournit une impulsion à chaque apparition d'une plosive. Il diffère du précédent par une réponse plus rapide de l'ordre de quelques millisecondes. Le dépassement d'un seuil par la sortie déclenche une impulsion calibrée.

c/ Segmentation : un canal séparé, indiquant simplement le logarithme de la puissance à court terme du signal, est utilisé pour la détection de parole, la séparation des mots et la détection des plosives internes. Un paramètre dérivé décrit l'évolution du spectre du signal comprimé : les combinaisons linéaires sont dérivées et additionnées de façon quadratique (technique voisine de [7] et [8])

2. Partie programmée

La nécessité d'un fonctionnement autonome en temps réel et de la souplesse d'un traitement programmé impliquait l'utilisation d'un microprocesseur. La phase de mise au point a fait intervenir, avant de figer les programmes en mémoire morte, un système complet avec un télétype et une mémoire vive de 8 K mots (1 K mots pour les données et le programme de traitement proprement dit).

La partie câblée de l'appareil fournit toutes les 5 ms un échantillon constitué de deux mots, le premier indiquant la présence de parole P et la classe détectée, et le second la valeur codée de la courbe de segmentation.

Les mots sont repérés par le signal parole silence P : P = 1 quand la puissance p dans la bande 300 - 3400 Hz dépasse un seuil s_o. P est maintenu à 1 pendant 250 ms après la disparition du signal ($p < s_o$). Une plosive interne est détectée lorsque P = 1 et $p < s_o$ pendant un temps compris entre 80 et 250 ms : on affecte alors la classe K aux échantillons correspondants.

Le programme effectue en permanence l'enregistrement des échantillons dans une table (environ 2 secondes) Dès qu'un échantillon est porteur de l'information de parole P, l'adresse courante est notée comme adresse de début de mot. Ensuite l'identification des segments est automatiquement enchaînée pour finalement aboutir à l'affichage du résultat. Cette opération s'effectue en moins d'une demi-seconde. Dès lors le processus d'enregistrement recommence dans l'attente d'un mot, tandis que l'inscription est maintenue.

Le processus d'identification des segments est fondé sur la longueur et la continuité des suites de classes issues de la table des échantillons. Chaque suite d'éléments identiques dont la durée dépasse 80 ms est reconnue comme un segment, c'est-à-dire un son; on tolère dans de telles suites un échantillon différent isolé (ex : ... AAAA x AAAA...). Les plosives initiales sont traitées séparément. Lorsqu'une suite a été annulée et que deux segments de même nature deviennent ainsi consécutifs, on les confond en un seul avec sommation des durées. A l'issue de ce processus, les caractères représentatifs des segments sont visualisés (ex : CHOKINA) sur un affichage lumineux. Un segment de classes non reconnues est représenté par un point (ex : CH. K. N.).

EXPLOITATION DES RESULTATS D'ANALYSE

1/ Analyse spectrale : détermination des paramètres et des seuils.

La distinction entre [ʃ] et les sons voisés ne présente pas de difficulté. Pour séparer les sons voisés entre eux nous les avons opposés par paires et les échantillons correspondants ont été soumis à une analyse factorielle par P. GRAILLOT [9] . L'équation de la droite optimale de séparation dans le plan principal d'inertie de chaque paire se traduit par une combinaison linéaire des canaux ; celle-ci a été ajustée pour la rendre indépendante du niveau : ainsi les caractéristiques de compression ne sont pas critiques. Le tableau 3 indique les combinaisons retenues. Une variable supplémentaire partage les échantillons en deux catégories : O + A d'une part, N + I d'autre part. Plusieurs conditions logiques sont cumulées pour chaque classe. A chaque combinaison linéaire correspond une seule variable logique. Les seuils ont été déterminés à partir de données d'analyses puis réajustés expérimentalement.

Deux points méritent d'être examinés : la validité des paramètres retenus et l'éventualité d'une adaptation au locuteur, étant entendu que nous excluons a priori l'intervention du contexte phonétique. Nous avons extrait des échantillons représentant les quatre sons voisés prononcés par une voix féminine et deux voix masculines dans trois contextes phonétiques différents (200 échantillons par son). Les valeurs moyennes et écarts types des canaux et du paramètres R_{NI} séparateur de la paire N/I sont les suivants :

	C1	C2	C3	C4	C5	C6	R_{NI}
I	243(11)	202(30)	39(36)	113(48)	176(36)	151(45)	141(25)
N	244(7)	198(20)	83(49)	94(20)	133(26)	90(40)	-63(94)

Le paramètre R_{OA} possède les propriétés semblables 42(27) et 134(33) et met en évidence les variantes de prononciation. Nous avons, sur les mêmes échantillons, calculé les écarts entre valeurs extrêmes pour chaque canal quand varient le locuteur et le contexte phonétique, puis un écart entre locuteurs E_{LOC} en moyennant les écarts précédents sur les canaux et sur les contextes, et un écart de contexte E_{PH} en moyennant sur les canaux et sur les locuteurs. Les résultats sont les suivants :

	A	I	N	O
E_{PH}	35	55	39	32
E_{LOC}	18	40	32	35

En attendant des résultats plus complets en fonction d'une articulation soignée (qui elle-même impliquait le fonctionnement en temps réel ! ...) dans la structure de mots choisie, ces résultats n'incitent pas à attendre des progrès très importants avec une adaptation individuelle indépendante du contexte phonétique.

2/ Détection des plosives

Les indications du détecteur de plosives n'utilisent que l'évolution temporelle de la puissance du signal. On peut tenter pour les débuts de mots d'améliorer les résultats en détectant une évolution spectrale en début de voyelle ou en comparant les instants d'apparition de la plosive et du voisement afin d'éliminer les fausses détections. La première solution étant plus difficile à mettre en oeuvre, nous avons d'abord étudié la seconde. Huit locuteurs (4 hommes et 4 femmes) ont prononcé des mots commençant par [a] , [i] , |o| et par [k] , [p] , [t] , [b] , [d] , [g] , [m] , [n] suivis de [a] , [i] et [o] . Pour chaque mot, après lecture d'un affichage numérique, le

résultat noté était : absence de plosive, voisement précédent la plosive, ou temps plosive-voisement. La figure 4 indique la moyenne et l'écart-type des résultats obtenus. Seuls figurent ceux relatifs aux "voyelles explosées" et aux plosives sourdes, pour lesquelles l'appareil est conçu [2] : de nombreuses plosives sonores n'ont pas été détectées, et dans le cas où elles l'étaient, le voisement était le plus souvent détecté avant l'explosion. Tous les k et t ont été détectés, mais seulement 90 % des p. Sur 96 mots commençant par [n] , 2 plosives ont été détectées (3 pour [m]). Sur 298 mots commençant par une voyelle 51 plosives ont été détectés ; [a] est plus perturbé que [i] et [ɔ] . On voit que les détections de plosives sur des voyelles initiales ne peuvent pas être corrigées avec une sécurité suffisante par la mesure du temps plosive-voisement, et que l'information apportée sur l'identité de la plosive est faible. [k] semble le son le plus favorable à détecter. Cependant la syllabe [t o) amène expérimentalement une plus grande sécurité que [k o] .

Les plosives internes sont détectées par un silence court [7] . Toutes les plosives sont donc correctement détectées, tandis que les voyelles initiales peuvent donner lieu à de fausses détections, mais non les nasales. Nous avons donc adopté le principe de faire commencer tous les mots par une consonne.

3/ Segmentation

L'utilisation du paramètre de segmentation décrit dans [10] a permis d'obtenir des résultats spectaculaires sur des mots de trois sons [3] . La figure 5 donne un exemple où la segmentation permet de corriger des imperfections de classification. Mais d'autres artifices, tels que l'alternance consonne-voyelle, permettent d'éviter cette complication. D'ailleurs, il n'est pas certain que cette segmentation directe, réalisée à partir de paramètres calculés sur des canaux en nombre aussi réduit, soit suffisamment fiable pour que le nombre de mots pratiquement utilisables augmente sensiblement.

EXPERIMENTATION ET CONTRAINTES DE VOCABULAIRE

Des résultats partiels obtenus en simulation ont déjà été publiés [3, 10] . Ils découlaient de procédures de correction variées, dont une segmentation en syllabes à partir des segments détectés [10] . Les résults indiqués dans ce qui suit sont relatifs à une expérimentation en temps réel, menée avec un bruit ambiant de 60 DBA avec un microphone Sennheiser . (l'utilisation d'un téléphone ne perturbe pas sensiblement les sons voisés, mais implique un réglage différent de la détection de voisement). Des essais ont été menés avec 15 locuteurs, dont 4 femmes, qui ont prononcé des listes de 20 mots. La liste L1 comporte des mots de structure K V V , V [ʃ] V et V [n] V (V = voyelle), dont il existe 24 possibilités ; la liste L2 comporte des mots CV CV où la seule restriction consiste à interdire les suites NIN et INI, soit 75 possibilités. Le locuteur a le contrôle immédiat par l'affichage

des segments détectés. Les résultats ont été interprétés par les expérimentateurs (programme de correction en temps réel en cours d'implantation) : pour la liste L1, le résultat n'est pas modifié, par contre, pour L2, on assimile toute suite de segments N et I à N, I, NI ou IN en fonction des segments voisins. Par exemple, KAI est corrigé en KANI, CNA en CHINA, KNICHO en KICHO. Si un mot, après correction éventuelle, ne correspond pas à la structure fixée, on le considère comme rejeté.

Les locuteurs ont d'abord prononcé les deux listes L1 et L2 au cours d'une première séance sans consigne spéciale. Cette séance a été suivie d'explications et d'un entraînement d'une durée de cinq à dix minutes. Puis chaque liste a été prononcée d'affilée après entraînement. Certains locuteurs ont procédé à une 3ème séance après de nouveaux essais.

Les résultats sont en moyenne les mêmes pour les deux sexes. La figure 6 regroupe l'ensemble des 15 locuteurs. Le nombre de reconnaissances erronées est toujours faible, même pour la première séance. On obtient un nombre intéressant de succès dès la seconde séance, ce qui indique un apprentissage rapide. Nous retenons la structure CVCV, qui correspond à 75 mots possibles avec lesquels on obtient dès la deuxième séance 1,5 % de fautes, 18,5% de rejets et 80 % de reconnaissances correctes. Ces pourcentages ont été portés respectivement à 3 %, 12 % et 85 % pour 16 locuteurs prononçant des mots CVCV, chiffres recueillis à la dernière séance (3 à 5 séances en tout).

CONCLUSIONS.

Nous avons montré qu'il est effectivement possible de reconnaître avec très peu d'erreurs plusieurs dizaines de mots bâtis à partir de six sons, sans adaptation au locuteur, avec un apprentissage du locuteur très court. Ceci peut se réaliser avec un appareillage relativement sommaire, dont le prix total des pièces détachées, y compris le traitement programmé, est inférieur à 25 000 francs. Il faut cependant noter que ceci n'est possible que grâce à des contraintes de vocabulaire : il est clair qu'il n'est pas question, même avec six sons, d'utiliser un langage non redondant avec un locuteur quelconque, à moins - peut-être - d'utiliser à tous les niveaux des paramètres nombreux et des traitements aussi élaborés qu'en reconnaissance de la parole continue... Ces contraintes de vocabulaire ne présentent pas d'inconvénient pratique, et nous avons obtenu des résultats utilisables pour des commandes vocales.

Pour obtenir des résultats concrets, il fallait mettre en place tous les éléments de la chaîne : traitement, classification, reconnaissance, affichage, comportement du locuteur. Il reste maintenant à optimiser certains éléments, par ordre de complexité croissante :

- revoir la classification sur des données acquises avec la structure CVCV adoptée, essayer comme l'a suggéré M. DREYFUS-GRAF, le son [e]au lieu de [i],et sur ces données,étudier de façon approfondie l'efficacité des divers paramètres [11].

- essayer une adaptation au locuteur pour améliorer les détections critiques, ceci en fonction des résultats précédents.

- prendre en compte le paramètre de segmentation afin d'augmenter la sécurité - au risque de diminuer le nombre de reconnaissances immédiates - et implanter en temps réel la procédure de reconstitution de syllabes citée [10] .

L'intérêt de ces perfectionnements ne prend de sens qu'en fonction d'une utilisation pratique des commandes vocales artificielles.

BIBLIOGRAPHIE

[1] J.A. DREYFUS-GRAF - Reconnaissance automatique de la parole codée (phonocode), sonore et chuchotée, Revue d'Acoustique n° 25 1973

[2] M. CARTIER, J.L.COURBON, R.DAGORNE, R.DOUCEN, J.J.LUCAS Capteur pour la reconnaissance de six sons. 2e rapport intermédiaire du contrat C.R.I. n° 73034, décembre 1974.

[3] J.A.DREYFUS-GRAF - Codes phonétiques (phonocodes) et commandes verbales, 6èmes J.E. Groupe Communication Parlée GALF, TOULOUSE (mai 1975), p. 396-404.

[4] J.P. KOESTER, J.A. DREYFUS-GRAF - Optimisation de phonocodes par tests d'intelligibilité avec des sujets allemands, id. (6ème J.E) p. 405-410.

[5] J.L. COURBON - Régulation automatique de niveau : étude et réalisation de deux compresseurs de dynamique, Note technique CNET, décembre 1973.

[6] L.C.W. POLS, L.J.T. VAN DER KAMP, R.PLOMP. Perceptual and Physical space of Vowel Sounds, JASA 46, 1969, p.458-467.

[7] L. BUISSON, G.MERCIER et al : Phonetic decoding for automatic recognition of words p. 189 - 196.

[8] J.J. MARIANI, J.S.LIENARD, G.RENARD - Le vecteur delta comme indice phonétique et son application à la reconnaissance automatique de la parole, 6e J.E. GALF TOULOUSE (mai 1975) pp 191-203.

[9] P. GRAILLOT - Projection, compression et reconstitution de données spectrales de parole, Bulletin de l'Institut de Phonétique de GRENOBLE, Vol.III, 1974, p. 53-71.

[10] J.DREYFUS-GRAF - Recognition of Coded Speech (Phonocodes) 1976 IEEE - ICASSP - PHILADELPHIE (mai 1976) session 6.

[11] C.BERGER-VACHON - Conception d'une entrée vocale automatique : prétraitement des signaux ... Thèse (septembre 1975) LYON - p. 76-89.

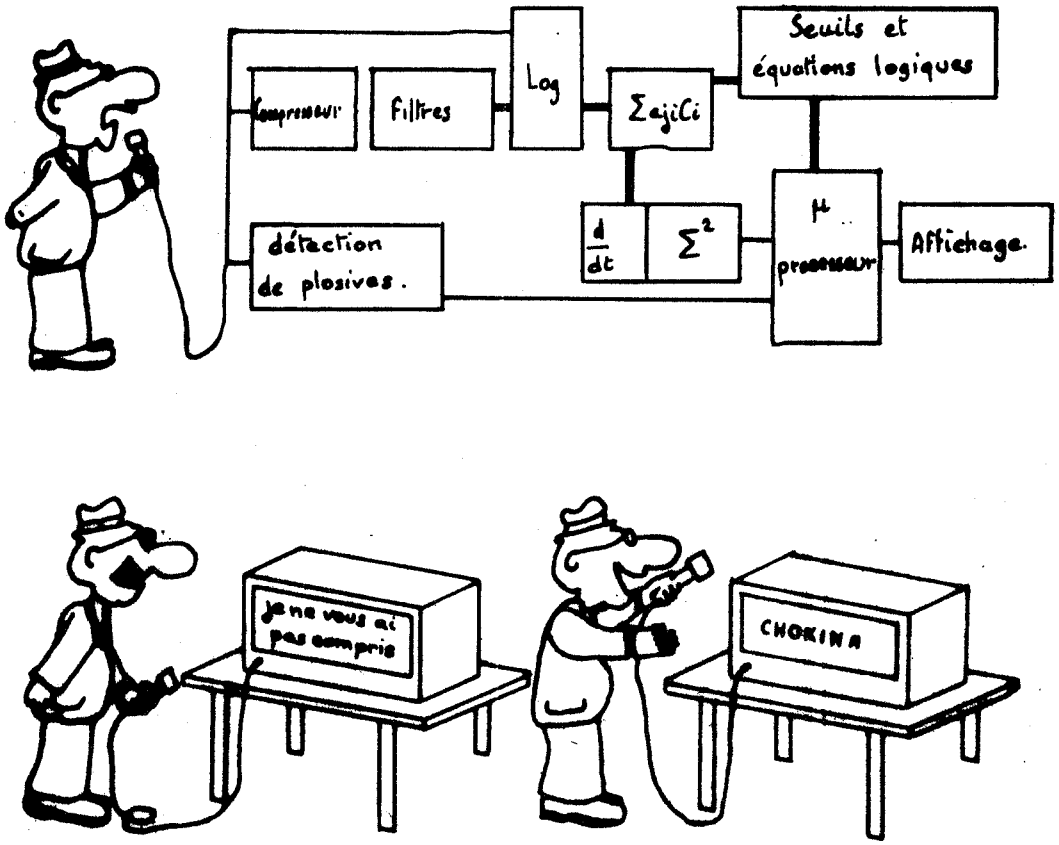
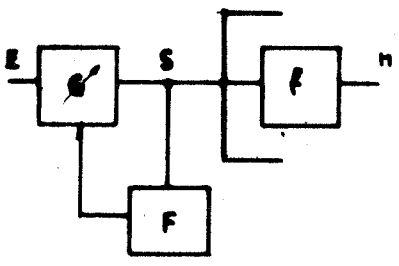


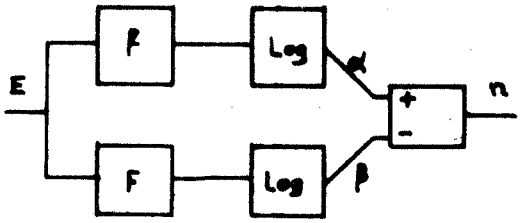
Fig 1 : Schéma d'ensemble.



Compresseur Sélectif :

$n = C^{ste} + \frac{1}{R} E + k_F - k_f$

E : niveau d'entrée
S : " de sortie
n : paramètre de sortie
k_f, k_F : Facteurs de forme spectraux liés aux filtres



Combinaisons linéaires :

$n = C^{ste} + \alpha(E - k_f) - \beta(E - k_F)$

E, n, k_f et k_F sont de forme logarithmique
α, β et $\frac{1}{R}$: coefficients multiplicatifs

$\Delta E / \Delta S = R$

Fig 2 : Principe d'analyse.

	C_1	C_2	C_3	C_4	C_5	C_6	seuil
CH/voisé	-0,33	-0,33	-0,33	0,33	0,33	0,33	L_1 : 65
N+I/O+A	0,3	-0,3	-0,7	0,1	0,3	0,3	L_2 : -15
O/A	0,7	0,3	-0,2	-0,8	0	0	L_3 : 82
N/I	0	0	1	0	-0,3	-0,7	L_4 : -65
N/O	1	-1	0	0	0	0	L_5 : 22
N/A	1	0	-0,6	-0,4	0	0	L_6 : 80
O/I	0,4	0,6	-0,2	-0,3	-0,5	0	L_7 : 20

$CH = L_1 . \bar{K}$

$O = \bar{L}_2 . L_3 . L_4 . \bar{L}_5 . \bar{K} . \overline{CH}$

$A = \bar{L}_2 . \bar{L}_3 . \bar{L}_6 . \bar{K} . \overline{CH} . \bar{O}$

$N = L_2 . L_4 . L_5 . L_6 . \bar{K} .... \bar{A}$

$I = L_2 . \bar{L}_4 . \bar{L}_7 . \bar{K} ..... \bar{N}$

Fig 3 : matrice des combinaisons, seuils et équations logiques.

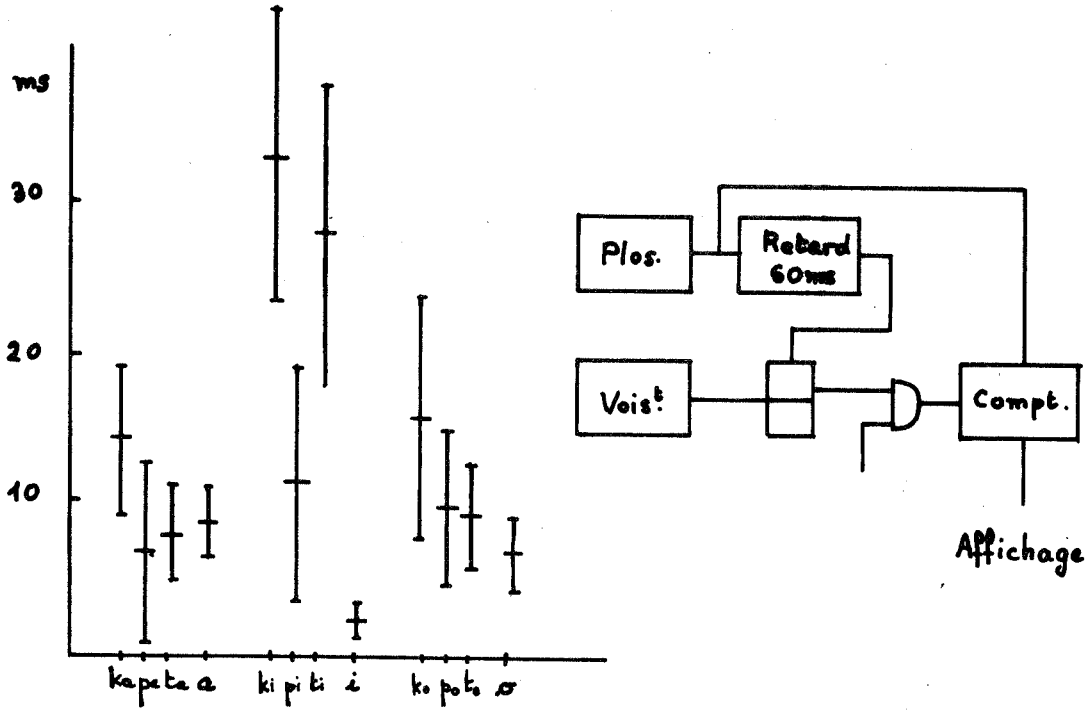


Fig. 4 : temps plosive-voisement : schéma et résultats. ($m \pm \sigma$)

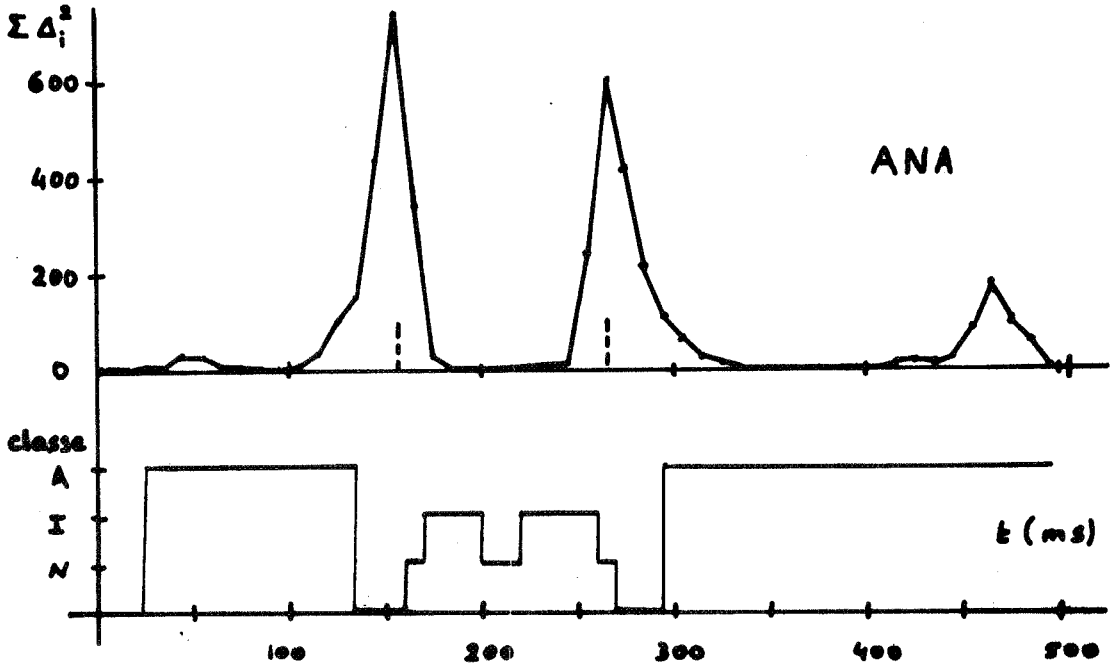


Fig. 5 : exemple de classification erronée avec segmentation correcte

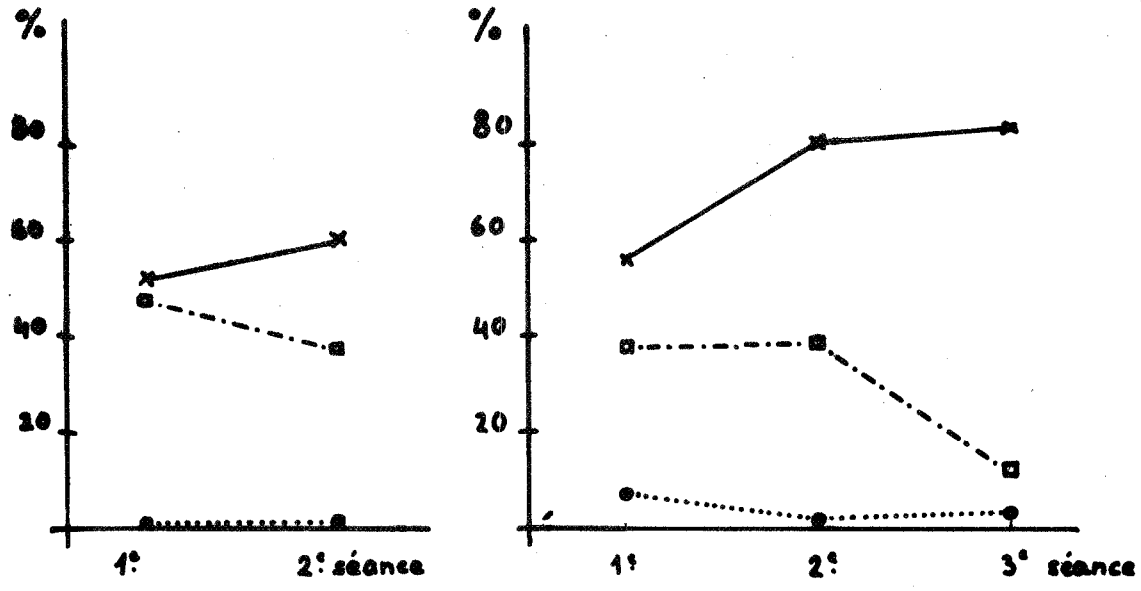


Fig 6 : résultats temps réel ; 15 locuteurs, 20 mots pour chaque liste.

L1 : mots de 3 sons
24 réponses possibles

L2 : mots CVCV
75 réponses possibles

correct

rejet

erreur

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

Deux Méthodes de détermination de la Fonction d'Aire du Conduit Vocal dans le domaine temporel .

Raymond DESCOUT *, Bernard TOUSIGNANT **, Michel LECOURS **

* Département E.T.A.
C.N.E.T.
22301 Lannion

** Département de Génie Electrique
Université Laval
Québec, Canada G1K 7P4

RESUME : Les méthodes consistent à envoyer dans le conduit vocal une onde de pression acoustique impulsionnelle, de la mesure de l'onde réfléchie, après passage dans la cavité buccale, deux traitements sont présentés:

I. La réponse impulsionnelle aux lèvres, obtenue après déconvolution, permet de connaître l'état du système, et de calculer la fonction d'aire en tout point;

II. Par approches successives, on modèle la forme du conduit vocal à partir de la recherche des constrictions par ordre décroissant d'importance, l'identification se fait alors de manière globale.

Les résultats obtenus par ces deux méthodes sont bons, les fonctions d'aire ainsi mesurées sont des mesures absolues. On fera une présentation comparative (avantages et inconvénient de ces deux méthodes complémentaires).

-=-=-=-=-=-

SUMMARY : Two methodes are described for determining the vocal tract area function from masurements at the lips of the response to an impulsive acoustic pressure wave:

I. In one case, the impulse response at the lips obtained by deconvolution is used to compute the area function at all points;

II. Alternatively, the vocal tract is modeled by successive approximation, a search is made for the constrictions by decreasing order of importance rather than sequentially from the lips.

Results obtained with both methods are good, the area functions are measured in absolute values rather than in arbitrary units. Advantages, difficulties and limitations of both methods are discussed.

-=-=-=-=-=-

Deux Méthodes de détermination de la Fonction d'Aire du Conduit Vocal dans le domaine temporel :

Raymond DESCOUT *, Bernard TOUSIGNANT **, Michel LECOURS **

I. INTRODUCTION

L'étude de la forme du conduit vocal et de son évolution en fonction du temps au cours d'une élocution continue est importante pour une meilleure compréhension du phénomène phonatoire et pour diverses applications dans le domaine de la reconnaissance, et de la synthèse de la parole.

Dans le but de commander un **synthétiseur** de parole par simulation du fonctionnement du conduit vocal à partir de paramètres de type articulatoire, il est nécessaire de disposer d'un ensemble de données de fonctions d'aire pouvant permettre la détermination pratique des grandeurs choisies et de leur évolution au cours du temps. C'est dans ce but que l'on s'est attaché dans ces études à recueillir une mesure "absolue" de la fonction d'aire du conduit vocal, et c'est la raison pour laquelle seule la cavité buccale a été mesurée par une méthode acoustique externe.

Dans ce texte deux méthodes sont décrites, dans les deux cas, le dispositif de mesure de l'onde acoustique réfléchie aux lèvres s'inspire de la technique proposée par Sondhi et Gopinath (1,2). Dans le premier cas, la recherche de la réponse impulsionnelle est un problème d'identification de processus; pour résoudre ce dernier, on approxime le système cherché par un modèle dont la structure est un filtre transversal. Dans le second cas, le conduit vocal est modelé de proche en proche par des approximations successives des coefficients de réflexion; les coefficients du conduit sont déterminés par ordre décroissant d'importance, alors que dans la première méthode, ceux-ci sont déterminés de façon séquentielle à partir des lèvres.

Dans la deuxième et troisième partie de ce texte, nous présentons l'aspect théorique du problème, les montages expérimentaux et les difficultés qui s'y rapportent. Dans les quatrième et cinquième partie, les deux méthodes sont exposées, ainsi que certains résultats obtenus.

II. PRINCIPES THEORIQUES ET LIMITATIONS EXPERIMENTALES

Pour déterminer la fonction d'aire du conduit vocal, il est nécessaire et suffisant de connaître l'onde retour, issue du conduit vocal au niveau des lèvres, en réponse à une impulsion acoustique incidente pendant une durée égale à environ 1ms. Ce temps est celui mis par cette dernière pour traverser le conduit vocal jusqu'à la glotte et en revenir. Sous réserve que sa fréquence supérieure soit limitée à 5 kHz , nous pouvons raisonablement faire l'hypothèse de propagation en onde plane, nous supposerons aussi que les pertes dans le conduit vocal sont négligeables ou de forme connue.

De façon à déterminer les pertes, il faut connaître la nature des parois buccales, évaluer l'impédance mécanique qu'elles représentent et tenir compte des pertes par vibrations. La détermination de l'impédance des parois a déjà été abordée par Flanagan (4), Ishizaka (5), Sondhi (6), et Mryati (10). Dans les méthodes que nous présentons ici, il est difficile d'introduire des facteurs variant avec la fréquence, aussi nous avons choisi

un facteur d'atténuation qui en soit indépendant. Comme El Mallawany (8), et Fant(5), nous avons pris une expression empirique que nous avons modifiée après plusieurs expériences sur des moulages et après avoir évalué son influence sur la qualité des résultats :

$$e^{-\alpha \Delta x} \quad , \quad \alpha = K(0,007\sqrt{\pi/A}) \quad \text{népers/cm}$$

où A est l'aire en cm^2, Δx l'unité de longueur en cm et K un facteur arbitraire.

De façon à être assuré que la propagation du son se fait en mode plan, et compte tenu des dimensions du conduit vocal, l'onde de pression acoustique incidente ne doit pas posséder de fréquences supérieures à 4 ou 5 kHz : ces conditions impliquent donc que l'onde aura une durée de l'ordre d'une fraction de ms. Sachant que la vitesse du son est de 34 cm/ms; il est difficile d'obtenir une résolution spatiale meilleure que le cm. Même s'il était possible de négliger la propagation en mode transverse, une autre limitation serait alors imposée par la taille du transducteur (en général un microphone B&K 1/4 de pouce) qui réduit ainsi la résolution de la mesure à 2 mm environ.

Ainsi on approxime le conduit vocal par une succession de sections cylindriques d'égale longueur, chaque section ayant une une longueur unitaire de 1 cm, ce qui permet de modéliser celui-ci en 17 sections environ. Le tube acoustique ainsi défini est alors caractérisé par ses coeffficients de transmission et de réflexion.

III. CONDITIONS EXPERIMENTALES

Nous voulons envoyer dans la bouche, au niveau des lèvres, une onde de pression acoustique plane de courte durée, et nous devons mesurer les réflexions en provenance de cette cavité. Tout ceci doit être réalisé pendant environ 1 ms de façon que le conduit vocal puisse être considéré comme stationnaire au cours de la mesure, de plus comme l'on désire suivre l'évolution dynamique du conduit vocal, il faut que cette impulsion puisse avoir une fréquence de répétition de 20 à 30 ms. Enfin, l'appareil de mesure doit permettre une élocution relativement aisée...

Nous avons vu que pour que l'onde depression puisse être considérée comme plane dans le tube et dans la cavité buccale, le contenu en fréquence de l'impulsion acoustique doit être limité à 5 kHz; cependant, puisque les petites constrictions de la cavité buccale affectent principalement les hautes fréquences, on doit avoir encore quelque énergie jusqu'à 3 ou 4 kHz. Diverses expériences ont été menées avec différents transducteurs électromécaniques et avec des étincelles électriques. Les étincelles présentent trop d'énergie dans les hautes fréquences, de plus il est difficile de contrôler leur répétitivité, et leur périodicité. Nous avons trouvé que avec des tweeters, la traînée est trop longue, et avec des H.P. de trop grand diamètre, il est difficile d'éviter de tomber dans les modes propres du tube acoustique. Au département de Génie électrique à l'Université Laval, nous avons obtenu une impulsion acoustique acceptable (figure 1) avec un haut-parleur "médium" excité par deux impulsions successives, respectivement positive, et négative. Au C.N.E.T. à Lannion, des résultats

équivalents ont été obtenus avec une chambre de compression, après avoir fait les premiers essais avec un H.P. de grand diamètre. Dans les deux cas, il a été montré que, en prenant de grandes précautions, les réflexions secondaires, et les traînées résultantes dans le tube, peuvent être maîtrisées et une nouvelle mesure peut être faite 20 ms après la première.

De façon à réaliser une liaison légère et flexible entre le tube de mesure et la bouche, nous utilisons un petit cône en plastique mou coupé à une extrémité de préférence à un masque respiratoire trop volumineux.

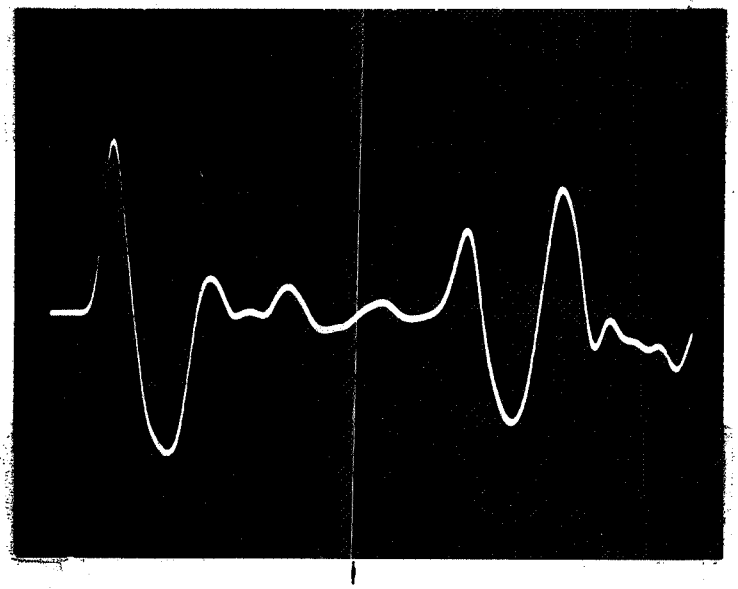


Figure 1

Onde Incidente Onde Réfléchie

IV. DETERMINATION DE LA REPONSE IMPULSIONNELLE

La fréquence de l'onde incidente envoyée dans le tube ne dépassant pas 5 kHZ, celle-ci n'a pas une durée infiniment courte devant le temps mis par l'onde pour parcourir une section de 1 cm. Si nous notons e(t), s(t), et h(t) les ondes incidentes et réfléchie, et la réponse impulsionnelle, celles-ci sont liées par l'équation de convolution suivante :

$$s(t) = \int_0^t e(\tau).h(t-\tau)\, d\tau \qquad -1-$$

Dans le casparticulier qui nous intéresse, le système étant causal, et pour un signal échantillonné, celle-ci s'écrit:

$$s_n = \sum_{i=1}^{N} e_{n-i+1}\ h_i \qquad -2-$$

Chacune des mesures de e et de s étant entachée d'erreur, il est impossible de faire la résolution directe de ce système linéaire. Sachant qu'une équation de convolution dans le domaine temporel se traduit par un produit simple dans le domaine fréquenciel :

$$S(\omega) = H(\omega).E(\omega)$$

il est possible d'obtenir des résultats significatifs en divisant les

spectres du signal de sortie et du signal d'entrée, ceci permet d'avoir accès à la forme grossière de la cavité à mesurer.Mais étant donné la faible durée des signaux mesurés (celle-ci étant liée à une meilleure séparation des ondes incidente et réfléchie), la résolution en fréquence des spectres n'est pas suffisante, et donc les transitions brutales et les faibles constrictions ne sont pas détectées.

L'approche que nous avons poursuivie pour résoudre ce problème consiste à chercher à approximer le système par un modèle ayant la structure d'un filtre transversal. L'optimisation se fait donc en minimisant le carré de l'erreur $\varepsilon(t)$ entre la sortie réelle mesurée du système $s(t)$ et la sortie $\hat{s}(t)$ estimée du modèle (figure 2).

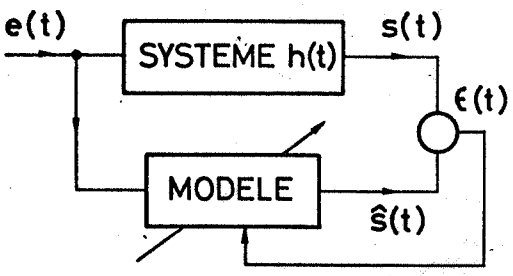


Figure 2

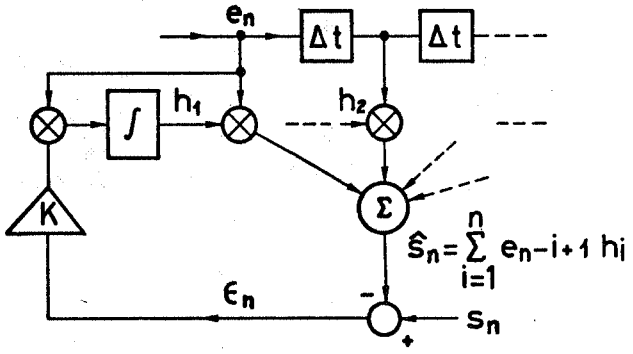


Figure 3

La structure adaptative est illustrée par la figure 3 où les coefficients h_k du filtre transversal représentent la réponse impulsionnelle estimée: leurs valeurs sont alors ajustées par des approximations successives utilisant la méthode du gradient à partir de l'équation de réccurence suivante :

$$h_k^{m+1} = h_k^m + K\varepsilon_n e_{n-k+1} \qquad -3-$$

dans laquelle m et m+1 représentent l'ordre de l'itération. Avec un choix judicieux du coefficient de contre-réaction K, il est possible d'obtenir de bons résultats après un nombre d'itérations de l'ordre d'une centaine. Un système câblé capable de faire cette opération sur une vingtaine de points de réponse impulsionnelle en moins de 100 ms, est en fin d'étude au C.N.E.T. et servira de périphérique de calcul rapide.(9)

V. APPROXIMATIONS SUCCESSIVES DES COEFFICIENTS DE REFLEXION

Considérons sur la figure 4a, une constriction soumise à deux ondes de pression planes A_I et B_I venant de deux directions opposées. Les relations entre les ondes incidentes A_I et B_I et les amplitudes des ondes transmises et réfléchies de la figure 4b seront:

$$A_R = r\, A_I \qquad -4-$$

$$A_T = (1+r)\, A_I \qquad -5-$$

$$B_R = -r\, B_I \qquad -6-$$

$$B_T = (1-r)\, B_I \qquad -7-$$

où r est le coefficient de réflexion qui est lié aux aires adjacentes S_i et S_j du tube par : $r = \dfrac{S_i - S_j}{S_i + S_j}$

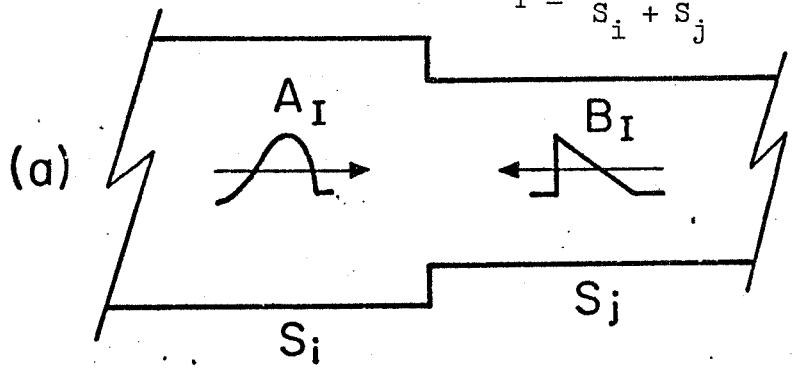
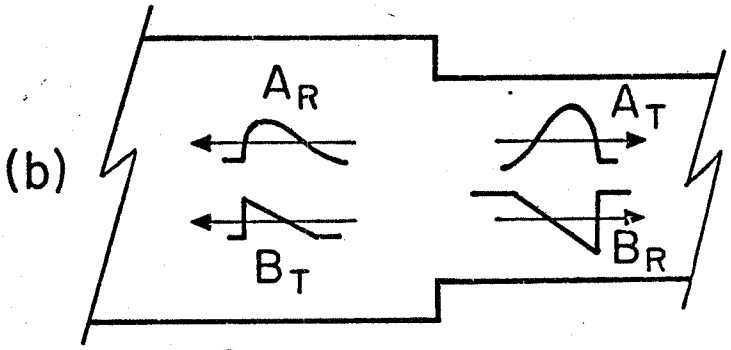


Figure 4



Un coefficient de réflexion peut être ainsi défini pour toutes les intersections entre chaque section du tube. L'équation liant l'échantil-lon à l'instant i et à l'intersection j sur la figure 5 sera donc la suivante :

$$P_R(i,j) = (1+r_j)\, P_R(i-1,j-1) - r_j\, P_L(i-1,j+1) \qquad -9-$$

$$P_L(i,j) = (1-r_j)\, P_L(i-1,j+1) + r_j\, P_R(i-1,j-1) \qquad -10-$$

où P_R et P_L sont respectivement les ondes de pression allant vers la droite et vers la gauche du conduit (figure 5).

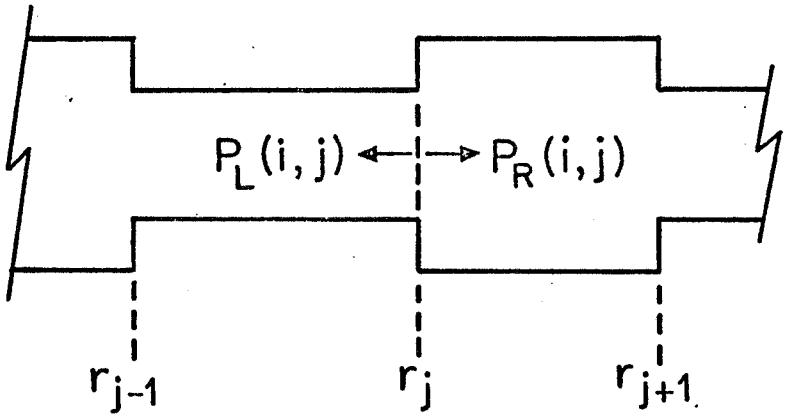


Figure 5

Ainsi, pour un tube initialement au repos, excité par une onde de pression $P_s(t)$, la réponse $P_e(t)$ peut être approchée par l'équation :

$$P_e(t) = \sum_{k=0}^{N} a_k \, P_s(t-k\Delta t) \qquad -11-$$

où $\Delta t/2$ est le temps mis par l'onde acoustique pour traverser une section unitaire, les coefficients a_k étant directement reliés aux coefficients de réflexion pour un système sans pertes..

La connaissance de l'onde de pression $P_s(t)$ et de sa réponse $P_e(t)$ permet, en principe, de localiser les constrictions et de déterminer leur importance. Deux facteurs contribuent à donner à une constriction son importance :

- sa position par rapport au point d'excitation;
- le saut de section que celle-ci représente.

Dans les données expérimentales, les effets dûs à ces deux facteurs sont mélangés, nous identifions alors les constrictions par étapes successives. Ainsi, sur la figure 6 une constriction importante semble exister entre la 3ème et la 4ème section. On évalue alors le coefficient de réflexion, et on compare la réponse mesurée $P_e(t)$ à celle que l'on obtient avec ce coefficient de réflexion évalué; la différence entre ces deux réponses permet alors la détection et la mesure d'une nouvelle constriction. Et ainsi de suite, d'étape en étape, il est possible d'identifier les constrictions majeures qui sont à l'origine de l'onde réfléchie $P_e(t)$. L'introduction d'un coefficient d'atténuation indépendant de la fréquence de type $\alpha = 0.007\sqrt{\pi/A}$ nepers/cm ne présente aucune difficulté.

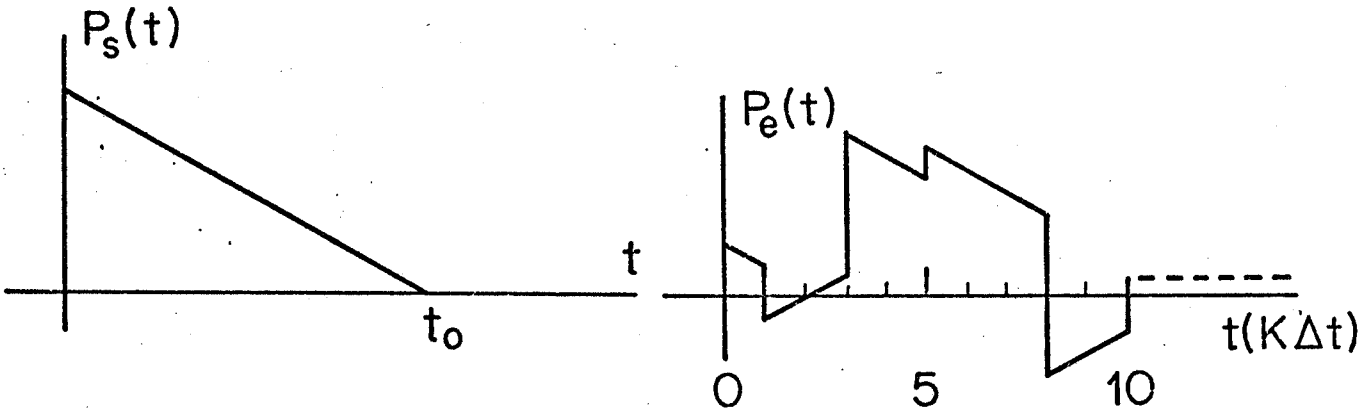


Figure 6

VI. CONCLUSION

Des mesures ont été effectuées sur différents types de moulages, et sur des conduits vocaux réels. Au C.N.E.T. nous avons réalisé, et utilisé des moulages en plastique mou, à l'Université Laval, les expériences ont été menées en utilisant des tubes de plexiglas et des moulages en plâtre. Pour les expériences sur des cavités buccales réelles, il est possible de comparer les résultats obtenus à ceux relevés à partir de mesures aux rayons X ; pour l'instant les fonctions d'aire mesurées sont utilisées comme données pour les synthétiseurs de conduit vocal que nous avons au C.N.E.T.

Les figures 7 à 9 montrent les résultats obtenus à partir de la méthode du gradient pour des moulages en plastique construits à partir des mesures fournies par Fant (7) et pour un conduit vocal réel (figure 9): comme on peut le constater, il y a un bon accord entre la mesure de la cavité réelle (M) et les résultats obtenus par la méthode du gradient (GR). L'algorithme de calcul est implanté sur le calculateur Télémécanique T 1600 du département E.T.A. et son exécution pour 20 coefficients de réflexion, et 150 itérations est de l'ordre de 30 sec. Comme nous l'avons précisé plus haut, un processeur spécialisé doit être prochainement raccordé pour réduire ce temps de calcul.

Les figures 10 et 11 montrent les résultats obtenus avec la méthode des coefficients de réflexion successifs pour un moulage en plâtre, et un conduit vocal réel avec différents coefficients d'atténuation. Ici encore, les résultats semblent bons. La fonction d'aire peut être ajustée à la longueur du conduit par des tests au cours du processus d'itération. Des contraintes physiques, telles que valeur maximale de la fonction d'aire admissible, peuvent aussi aider à accélérer la convergence. Le choix d'un coefficient d'atténuation indépendant de la fréquence, quoique théoriquement non valable, a donné jusqu'ici de bons résultats. On obtient un nombre de coefficients de réflexion suffisant après 10 ou 12 passes de façon à identifier le conduit vocal avec assez de précision. l'erreur de sortie estimée atteint alors son premier minimum et oscille ensuite autour de cette valeur.

Il est encore difficile, arrivés à cette étape de nos études, de comparer les deux méthodes. La méthode du gradient nécessite moins de données sur l'onde de pression réfléchie, elle semble aussi moins sensible aux caractéristiques de l'onde de pression incidente. D'autre part, la méthode des coefficients de réflexion successifs nécessite moins d'itérations et la qualité des résultats diminue avec la distance aux lèvres de façon moins marquée qu'avec l'approche séquentielle: son avantage principal est de déterminer en premier lieu les variations importantes de la fonction d'aire.

Ce type de mesure est bien sûr limité aux configurations de conduit vocal dans lesquelles la bouche est raisonablement ouverte. Par ailleurs, du fait de l'absence de phonation du patient, il y a quelques difficultés à relier une séquence phonétique supposée être dite, à la séquence des fonctions d'aires ainsi déterminée; le sujet doit aussi apprendre à maintenir sa tension musculaire dans le conduit vocal, sans qu'aucun flux d'air ne s'échappe de ses poumons.

Cependant, ces limitations étant acceptées, il y a un espoir réel d'atteindre ainsi par des moyens purement acoustiques , des séquences de fonction d'aire en mesure absolue, pour une utilisation ultérieure à des fins de synthèse et d'études phonétiques.

Nous remercions vivement M.M. Sondhi des Laboratoires Bell qui s'est intéressé et nous a encouragé dans la prémière étape de cette étude. Nous voulons aussi remercier Melle Thérèse Pierrat pour son travail efficace en tant qu'assistant de recherche sur ce projet.

BIBLIOGRAPHIE

1. M.M. Sondhi, B. Gopinath, "Determination of Vocal-Tract Shape from impulse response at the lips", J. Acoust. Soc. Amer., 49, n°6 (prt 2), 1867-1873, June 1971.
2. M.M. Sondhi, B. Gopinath,"Determination of the Shape of a Lossy Vocal Tract", 7th Int. Congress on Acoustics, Budapest 1971,pp. 165-167
3. J.L. Flanagan, K. Ishizaka, K.L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell Syst. Tech. J.,Vol. 54, n°3, March 1975, pp. 485-506.
4. J.L. Flanagan, "Speech Analysis, Synthesis and Perception", 2nd Edition, New York: Springer Verlag, 1972.
5. K. Ishizaka, J.C. Frech, J.L. Flanagan, "Direct Estimation of Vocal Tract Wall Impedance", J. Acous. Soc. Amer. ,55 (April 1974), p.579(A).
6. M.M. Sondhi, "A model for wave propagation in a Lossy Vocal-Tract", J. Acous. Soc. Amer. , 55 (May 1974), pp. 1070-1075.
7. G. Fant, "Acoustic Theory of Speech Production", Mouton 1970.
8. I. El Mallawany, "Fonction de transert et fonction d'aire du conduit vocal", Analyse et Synthèse de la Parole, Vol I, Rapport annuel 72-73 CNET pp. 105-122.
9. J.C. Le Viol, "Identification de la fonction de ponderation h(t)", Recherches/acoustique, CNET, Vol II, pp. 151, 1975.
10. M. Mryati, "Modèlisation du Conduit Vocal, et étude des pertes ", Thèse de troisième cycle, Grenoble Fev. 1976.

-=-=-=-=-=-=-=-=-=-=-=-=-=-

Résultats expérimentaux :

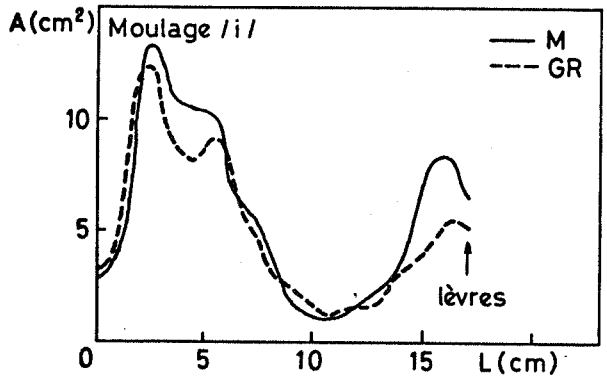


Figure 7

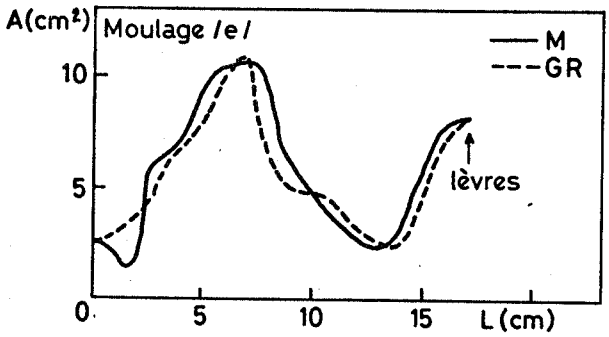


Figure 8

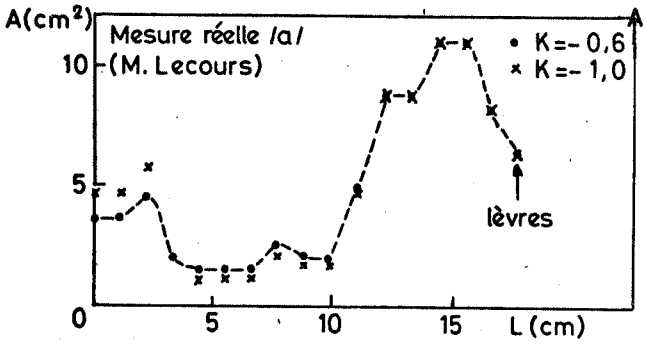


Figure 11

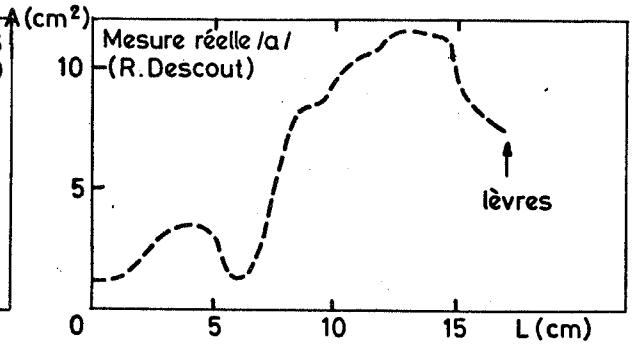


Figure 9

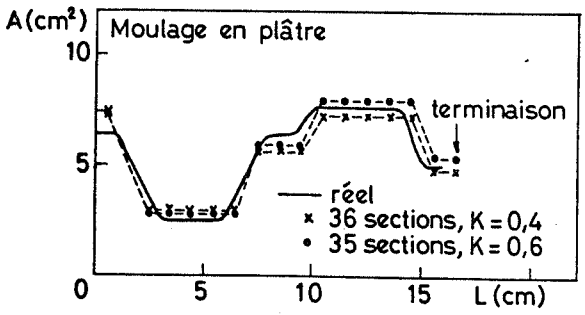


Figure 10

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 AU 21 MAI 1976

APPROCHES A LA DETECTION DE ET A L'ANALYSE SUR L'INTERVALLE DE FERMETURE DE LA GLOTTE

PAR

I.I. EL MALLAWANY - CNET-LANNION

RESUME :

Deux méthodes non itératives pour la détection de l'intervalle de fermeture de la glotte sont présentées. La première fondée sur l'interprétation de l'erreur normalisée obtenue par une méthode de prédiction linéaire s'est révélée très fiable mais fait appel à des calculs trop nombreux. La deuxième basée sur la détection de l'intervalle dont le spectre à très court terme a une dynamique minimale est plus simple et nécessite un temps de calcul nettement inférieur. L'application de cette dernière approche à l'analyse de la parole est décrite et des résultats sont présentés.

SUMMARY :

Two non iterative algorithms for the detection of the closed glottis interval are described. The first approach, based on the interpretation of the normalized error evaluated using linear predictive coding, has proved to be reliable but time consuming. The alternative method, which calls for the detection of the interval for which the very short time spectrum has a minimal dynamic range, is simpler and less time consuming. The application of the latter approach to speech analysis is described and results are given.

TITRE : APPROCHES A LA DETECTION DE ET A L'ANALYSE SUR L'INTERVALLE DE FERMETURE DE LA GLOTTE

AUTEUR : I.I. EL MALLAWANY

I - INTRODUCTION

Dans le codage de la parole en vue de sa transmission à de faibles débits, l'analyse de la parole est du type global. Cependant, dans les autres applications l'accent est mis sur l'extraction de paramètres pertinents, qui décrivent soit la phonation, soit l'articulation. L'analyse asynchrone du signal de parole est bien adaptée au codage du type global et trouve sa justification dans l'évolution relativement lente du spectre à court terme. Néanmoins, les résultats d'analyse sont imprécis en raison de la non stationnarité et de la pseudo-périodicité du signal. Ces inconvénients sont surmontés en portant l'analyse sur une période entière de la mélodie. Cependant, il subsiste un inconvénient majeur, à savoir, la variabilité de l'impédance à la glotte et le taux d'intéraction, qui s'ensuit, entre la source de phonation et le conduit vocal (CV). En effet, les résonances de l'appareil vocal sont différentes dans les deux temps d'ouverture et de fermeture de la glotte. Dans l'intervalle de fermeture de la glotte (IFG) l'impédance à la glotte est infinie et les résonances sont caractéristiques du seul CV donc de l'articulation uniquement. Pendant le temps d'ouverture de la glotte les résonances de l'appareil vocal diffèrent en raison de l'intéraction entre le CV et les cavités sous-glottales. Par conséquent, cette analyse synchrone ne permet pas de séparer la phonation de l'articulation et subit, par ailleurs, l'effet des variations prosodiques.

De ces diverses considérations, on peut conclure que la meilleure approche au problème de l'extraction de paramètres pertinents est fondée sur l'analyse dans l'IFG. Cependant, un tel traitement n'est exploitable que si trois conditions sont réunies : simplicité relative des calculs, fiabilité de la détection de l'IFG et pertinence des résultats d'analyse. Nous présentons ci-dessous une contribution dans cette voie.

II - LA PREDICTION LINEAIRE

Le principe de la méthode [1] est qu'en l'absence temporaire d'une source d'excitation, un échantillon de parole, s_n, peut être prédit avec exactitude à l'aide d'une somme pondérée linéairement de p échantillons le précédant immédiatement. Dans cette hypothèse, la fonction de transfert de l'appareil vocal peut être représentée par un filtre numérique récursif ne possédant que des pôles [1] de la forme

$$H(z) = G/(1 + \sum_{k=1}^{p} a_k z^{-k}) \qquad (1)$$

Sur un intervalle de temps donné l'erreur quadratique totale de prédiction est donnée par

$$E = \sum_n e_n^2 = \sum_n \left(s_n + \sum_{k=1}^{p} a_k \; s_{n-k} \right)^2 \qquad (2)$$

L'analyse consiste à calculer un ensemble de coefficients, a_k, du prédicteur qui minimise la grandeur E au sens des moindres carrés, dans l'hypothèse de stationnarité du CV dans l'intervalle d'analyse. La minimisation de E conduit au système d'équations [1]

$$\sum_{k=1}^{p} a_k \cdot R_{|i-k|} = -R_i \quad , \quad 1 \le i \le p \qquad (3)$$

où

$$R_i = \sum_{n=0}^{N-|i|} s_n \, s_{n+|i|} \qquad (4)$$

Ce système d'équations peut être résolu par récurrence comme suit

$$E_o = R_o \qquad (5)$$

$$k_i = - \left(R_i + \sum_{j=1}^{i-1} a_j^{(i-1)} \, R_{i-j} \right) / E_{i-1} \qquad (6)$$

$$a_i^{(i)} = k_i \qquad (7)$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i \, a_{i-j}^{(i-1)} \quad , \quad 1 \le j \le i-1 \qquad (8)$$

$$E_i = (1-k_i^2) \, E_{i-1} \qquad (9)$$

Ces relations sont évaluées par récurrence avec i=1,2,...,p, la solution étant donnée par $\{ a_k = a_k^{(p)} \}$. L'erreur quadratique totale minimale de prédiction est donnée par

$$E_p = R_o + \sum_{k=1}^{p} a_k \, R_k \qquad (10)$$

L'erreur normalisée est définie par

$$V_p = E_p \, / \, R_o = \prod_{i=1}^{p} (1 - k_i^2) \qquad (11)$$

De cette analyse on obtient deux ensembles de paramètres : les $\{a_k\}$ qui sont les coefficients d'un filtre récursif classique, et les $\{k_i\}$ qui sont les paramètres d'un filtre à structure en échelle.

Les $\{k_i\}$ représentent les coefficients de réflexion aux jonctions consécutives entre les sections cylindriques d'égale longueur constituant le conduit acoustique équivalent à l'appareil vocal. Ce conduit peut approcher [1] la fonction d'aire du CV dans le cas des sons sonores sans nasalité. Si A_i représente l'aire d'une section, i, il vient

$$k_i = (A_i - A_{i+1}) / (A_i + A_{i+1}) \qquad (12)$$

III - METHODE FONDEE SUR LA PREDICTION LINEAIRE

Cette méthode a été décrite dans le détail ailleurs [1], et seules les principales considérations seront abordées ici. Le principe de la méthode est fondé sur une mesure de la dépendance linéaire de certains intervalles du signal, dans l'hypothèse où l'analyse porte sur une séquence de 2p échantillons du signal bien plus petite que la période de mélodie, T_M. En principe, l'erreur de prédiction est importante dans des intervalles où l'excitation est non nulle, ce qui signifie que la dépendance linéaire est relativement faible. Par contre, cette erreur est très faible sur des séquences dans l'intervalle de fermeture de la glotte où la dépendance linéaire est nettement plus grande. Cette erreur est, néanmoins, fortement dépendante de l'amplitude du signal, et il nous est apparu plus fiable de baser la méthode sur l'interprétation de l'erreur normalisée, V_p, où p est l'ordre du prédicteur utilisé.

Afin que la dépendance linéaire soit faible pendant le temps d'ouverture de la glotte, il est nécessaire d'appliquer la mesure sur un intervalle de la vitesse volumique aux lèvres, $v(t)$, et non de la pression au capteur, $s(t)$. Nous savons que $v(t)$ est donnée approximativement par l'intégral de $s(t)$. Ainsi, l'erreur normalisée, V_p, correspond à la vitesse volumique à la glotte et non à une dérivée de celle-ci.

Il est possible d'interpréter le comportement de V_p de la manière suivante. V_p dépend entièrement de la forme du spectre du signal. Si l'énergie spectrale est bien répartie sur l'ensemble de la bande de fréquence utile, V_p tend vers l'unité. Si, au contraire, toute l'énergie est concentrée en certaines régions du spectre, V_p tend vers zéro. Par conséquent, dans les intervalles d'ouverture de la glotte où l'énergie spectrale est fortement concentrée dans les basses fréquences on aura V_1 très faible et V_p légèrement plus faible d'où un écart $(V_1 - V_p)$ faible. Par contre, au niveau de l'IFG, où le signal prend la forme de la réponse d'un filtre en oscillation libre et où les hautes fréquences ont des énergies relativement élevées, V_1, sera relativement important

et du fait que la réponse du CV peut être approchée à l'aide d'un filtre ne possédant que des pôles, V_p, sera sensiblement plus faible d'où $(V_1 - V_p)$ relativement important. Par conséquent, l'écart $e=(V_1-V_p)$ représente une mesure qualitative de la dépendance linéaire. Cette mesure est calculée sur une fenêtre rectangulaire cadrant une séquence de 2p échantillons. Cette fenêtre est décalée d'une période d'échantillonnage, T, à chaque fois ce qui permet la détermination de la fonction de temps e(nT), où n représente l'instant du premier échantillon de la séquence.

Cette interprétation du comportement général de l'écart, e, fait appel à la remarque suivante : la valeur de e peut être importante dans des intervalles où l'excitation est non nulle mais dans lesquels le signal de parole prend l'allure de la réponse d'un filtre en oscillation libre. Ce cas se présente dans des intervalles qui correspondent à une crête de la pulse glottale où la valeur de la vitesse volumique à la glotte, δ (nT) est pratiquement constante. De ce fait, dans l'hypothèse de vibrations vigoureuses des cordes vocales, on détectera deux maxima dans une période de mélodie. Cette ambiguité peut être éliminée en ayant recours au signal d'excitation, δ (nT), qu'on peut calculer par une déconvolution entre le filtre modèle calculé sur l'intervalle supposé être l'IFG et le signal v(t). La pointe marquant le début de l'IFG sera déterminée en fonction de la position du minimum de δ(nT), qui doit obligatoirement se situer dans la séquence d'analyse de 2p qui a permis le calcul de la valeur de la pointe. Cette dernière opération ne permet pas, uniquement, de lever l'ambiguité dans ce cas, mais permet de garantir le fonctionnement fiable de la procédure dans le cas où les vibrations des cordes vocales ne sont pas vigoureuses. Dans ce dernier cas, les pointes ont tendance à se multiplier et afin de parvenir à une localisation correcte de l'IFG on est obligé de faire appel à trois pointes du signal d'erreur e(nT).

Cette méthode a été largement utilisée et s'est avérée très fiable. Cependant, l'importance des calculs rend difficile son exploitation dans le domaine des applications. Il reste, néanmoins, un autre inconvénient à cette approche, qui est la nécessité de se référer au signal d'excitation. En effet, il convient de prendre certaines précautions avant d'appliquer l'algorithme dans le cas où la parole est reproduite par magnétophone. En général, ces appareils engendrent une distorsion du temps de propagation de groupe au niveau des basses fréquences, qui est de nature à rendre le signal d'excitation inexploitable à moins d'une égalisation de cette distorsion. Afin d'éviter ce genre d'écueil et de simplifier les calculs, il est possible de valider la pointe du signal d'erreur marquant le début de l'IFG en se basant sur l'énergie du signal dans l'intervalle élémentaire d'analyse correspondant. L'intervalle de plus grande énergie représente l'IFG. Ce critère s'est révélé opérant. Néanmoins, aucune étude statistique n'a été menée pour vérifier sa fiabilité.

IV - APPROCHE FONDEE SUR LA PENTE DU SPECTRE

Compte tenu des inconvénients de la méthode précédante, nous avons recherché un moyen simple pour détecter l'IFG. A cette fin, on ne dispose que de deux critères susceptibles de servir de base de départ à la détection de cet intervalle : un spectre à très court terme dont les hautes fréquences ont une énergie relativement élevée, et une très grande énergie du signal.

Le premier critère peut être fondé sur l'évaluation de la pente moyenne du spectre. La détection de l'IFG sera marquée par un minimum de cette mesure. Le principe de l'estimation de l'indice de pente utilisé fait appel au calcul du premier coefficient de prédiction. En effet, si la séquence d'échantillons, s_n, dans l'intervalle élémentaire est délimitée par n=1 à n=N, ce coefficient est donné par

$$a = \sum_{n=2}^{N} s_n s_{n-1} \Big/ \sum_{n=1}^{N-1} s^2_n$$

Le prédicteur dans ce cas est donné par $(1-az^{-1})$. Par conséquent, si "a" est positif l'énergie est concentrée dans les basses fréquences et vice-versa. Néanmoins, ce seul coefficient est insuffisant pour l'étude de tous les cas qui peuvent se présenter, car il ne peut pas faire la différence entre pentes dépassant 6dB/octave. De ce fait, il est nécessaire d'effectuer un filtrage inverse sur la séquence s_n, soit

$$s'_n = s_n - as_{n-1}$$

Cette opération représente l'annulation de la partie mesurée de la pente spectrale ou en d'autres termes une égalisation de celle-ci. On réitère les calculs sur s'_n et s''_n. En effet l'estimation de trois valeurs de "a" consécutives suffit pour bien marquer le minimum recherché dans tous les cas. L'indice, I_p, est donné par la somme algébrique des trois mesures de a.

L'indice d'énergie, Ig, peut être donné par $\sum_n s_n^2$. Cependant cette indication n'a qu'une valeur qualitative dans la mesure où l'on est tributaire d'un convertisseur qui peut être décentré.

Le traitement doit être basé sur l'évaluation de ces indices sur la totalité de l'intervalle d'analyse considéré (10 à 30ms). Plus précisément, il faut évaluer ces grandeurs sur une fenêtre représentant l'intervalle élémentaire N, et décaler progressivement cette fenêtre d'un échantillon de sorte à balayer tout l'intervalle d'analyse. Il convient, ensuite, de localiser l'instant n=k pour lequel la fonction Ip_n marque un minimum de pente mais d'énergie maximale vérifiée par Ig_n. Cet instant marque le début de l'effet de fermeture de la glotte au niveau des lèvres. Cependant, l'instant de fermeture de la glotte intervient plus tôt ce qui tient au temps de propagation dans le CV. Par conséquent, cet instant

se situe au voisinage de k-p où p représente l'ordre du prédicteur, tel que pT soit à peu près égal au temps nécessaire à la propagation d'une onde acoustique pour effectuer l'aller-retour entre la glotte et les lèvres (T=période d'échantillonnage).

Il reste deux points à préciser : comment estimer p et quelle est la longueur N de la fenêtre. D'après les considérations précédantes il apparait que p dépend de T et de la longueur, L, du CV [1] qui est inconnue. Par conséquent, une solution très approximative à ce problème consiste à se fonder sur la mélodie, f_M, moyenne du locuteur. La valeur de p que nous retenons est toujours évaluée par excès afin de tenir compte des variations de L. Si la valeur de p est importante pour l'extraction de la fonction d'aire du CV, de faibles écarts ne compromettent pas le bon fonctionnement de l'algorithme de détection de l'IFG. A titre d'exemple, pour $75<f_M<125 \rightarrow$ L = 22cm et $175<f_M<275 \rightarrow$ L = 18cm, et $p=2.L/(350.T)$ où L est en mètres.

La longueur de la fenêtre est choisie en fonction de l'analyse que l'on compte effectuer. Dans ce cas le choix s'est porté sur l'application de la méthode de covariance de la prédiction linéaire. Par conséquent, si les p échantillons précédant l'instant k fournissent les conditions initiales de l'analyse il est nécessaire que $N \geq p$. Nous avons retenu $N = p + (p+1)/2$. Il faut noter que N et p doivent être liés afin d'assurer des performances relativement stables de l'algorithme pour tous locuteurs.

Cette méthode s'est avérée praticable. Cependant, si les calculs sont moindres que dans le cas précédant, l'algorithme requiert quand même un temps de calcul relativement important. Afin de réduire davantage ces calculs, l'attention s'est portée sur l'intégration des deux indices en un seul en faisant appel à la longueur de la trajectoire du signal inscrit dans la fenêtre élémentaire. Une trajectoire sera d'autant plus longue que l'énergie sera plus importante et les oscillations (énergie dans hautes fréquences) seront plus nombreuses. L'indice It, de la longueur de cette trajectoire est donné par

$$It = \sum_{n=2}^{N} (s_n - s_{n-1})^2$$

On notera que cette mesure n'est pas affectée par un convertisseur décentré. La valeur de It est localisée dans le temps par coïncidence avec le premier échantillon de l'intervalle élémentaire. L'IFG sera localisé par la plus grande valeur de It dans l'intervalle à analyser. Cependant, l'instant j ainsi fixé ne représente, ici encore, que l'effet au niveau des lèvres et coïncide le plus souvent avec le minimum de pente détecté à l'aide de Ip. Néanmoins, il peut y avoir de faibles écarts. Par conséquent, il est nécessaire d'évaluer Ip dans le voisinage direct de l'instant jT soit dans l'intervalle $\lceil j \pm (P/2) \rceil . T$ afin de situer le minimum k de Ip le plus proche.

V - L'ANALYSE

Nous avons analysé le signal de parole, principalement, en vue de l'extraction de la fonction d'aire du CV. Le mode de calcul de ce profil d'aire est décrit ailleurs [1]. A cet effet, nous avons utilisé la méthode de covariance qui est la mieux adaptée pour l'évaluation des coefficients, a_k, du filtre modèle du CV. La séquence d'échantillons à analyser va de n=1 à n=p+(p+1)/2 = N, et les échantillons de n=-p+1 à n=0 représentent les valeurs initiales. Le choix de N représente un compromis entre la précision souhaitable dans l'estimation des a_k (N grand) et le souci de limiter l'intervalle d'analyse au temps pendant lequel l'intéraction entre le CV et la source est nulle ou négligeable.

Les deux problèmes qui se posent dans ce type d'analyse sont la stabilité du filtre modèle et la régularité des formes des fonctions d'aire. L'application directe de la prédiction linéaire sur la séquence, s_n, mène à des configurations du CV qui sont cohérentes et assez réalistes mais la régularité n'est pas parfaite pour des voyelles soutenues. Par conséquent, nous avons envisagé d'appliquer une apodisation au signal afin de rendre ces formes plus régulière, de diminuer le risque d'instabilité et de parvenir à des configurations réalistes. En effet, en l'absence d'apodisation il y a d'importantes interférences entre bandes de fréquence et le spectre global est moins contrasté. L'apodisation partielle utilisée est dérivée de la fenêtre de Hamming, soit

$$W_n = 0.54 - 0.46.\ \cos\ (2\ \pi\ .\ (n+p+3)/(N+p+6))$$

Avant l'analyse il convient d'effectuer l'égalisation de la pente spectrale. Cette opération est réalisée selon le mode décrit en § IV et en portant l'évaluation des facteurs "a" sur l'intervalle n=1,...,N. Afin de lisser légèrement les configurations du CV et de réduire davantage les risques d'instabilité on augmente les largeurs de bande des formants d'environ 50 Hz en remplaçant [1] les a_k par $[a_k.\exp(-c.k)]$ où $c = 50\ \pi\ T$. Des résultats d'analyse sur des diphtongues sont illustrés dans les figures suivantes.

VI - CONCLUSION

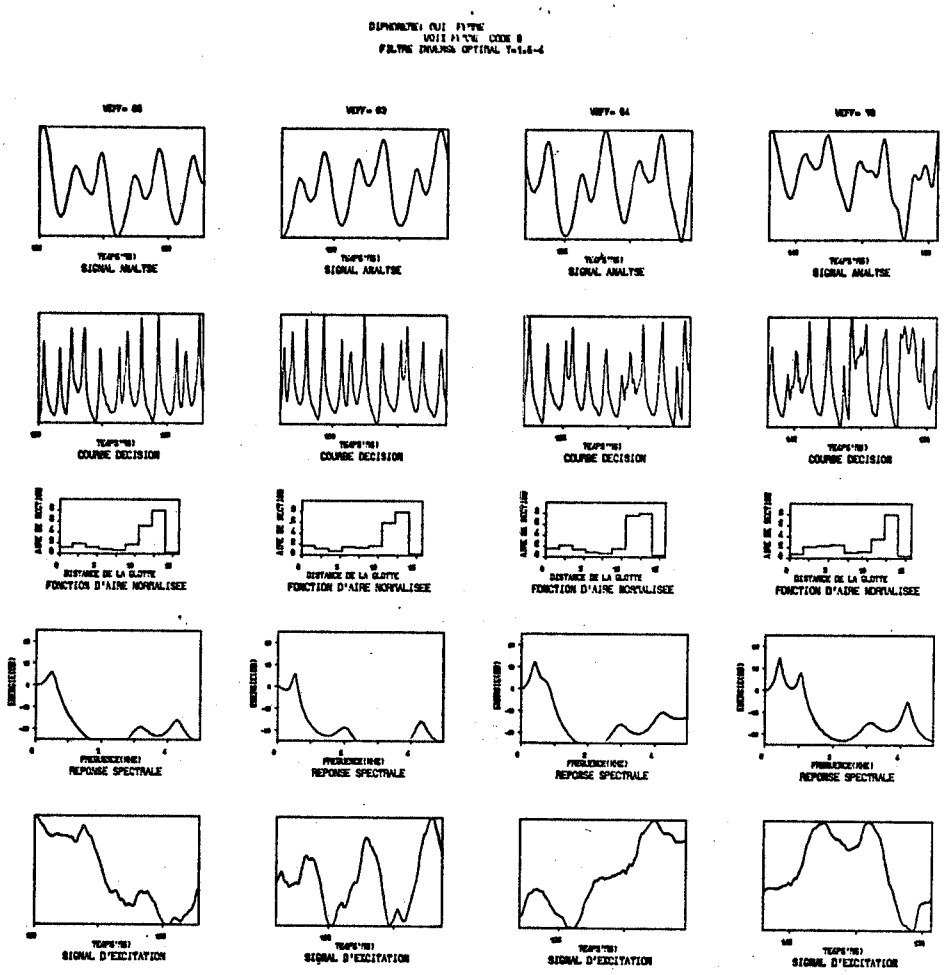
La méthode de détection de l'IFG fondée sur la pente du spectre à très court terme ne fait appel qu'à des calculs relativement simples et n'exige pas de temps de calcul important. Au stade actuel de l'étude il n'est pas possible de se prononcer sur sa fiabilité. Néanmoins, l'algorithme a été appliqué à des voyelles orales et nasales ainsi qu'à des diphtongues et s'est avéré très satisfaisant. Cependant, le résultat le plus intéressant est que l'analyse de la parole sur l'IFG à l'aide de la prédiction linéaire permet l'extraction de configurations du CV qui sont stables et réalistes.

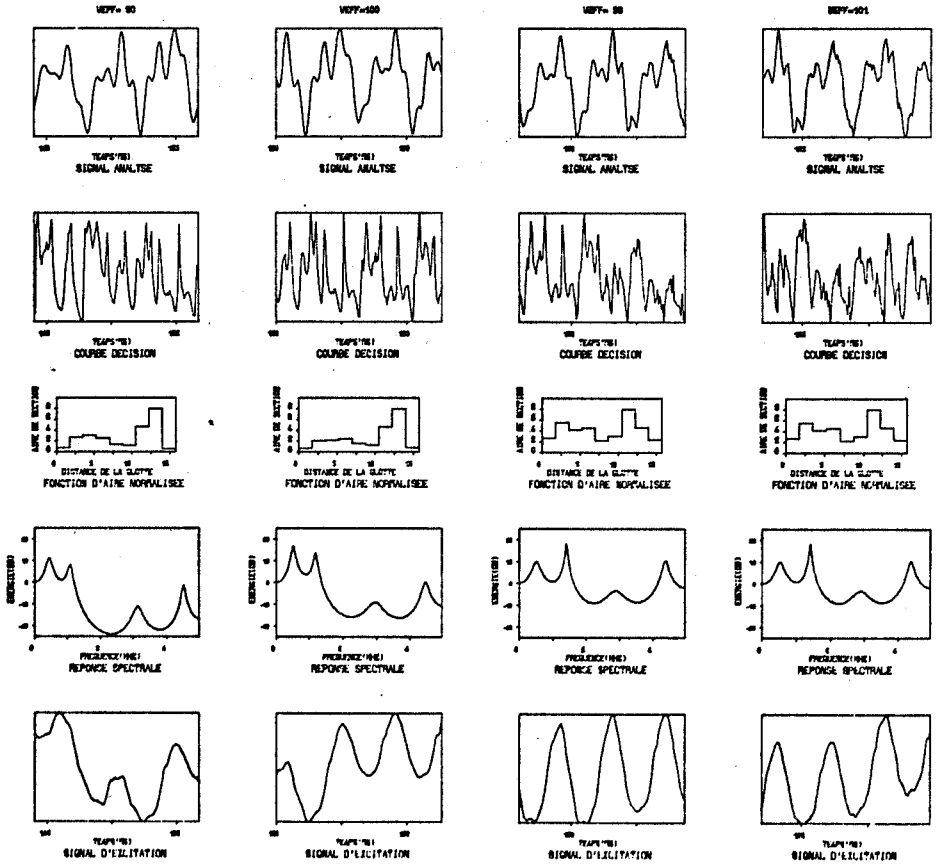
VII - REFERENCE

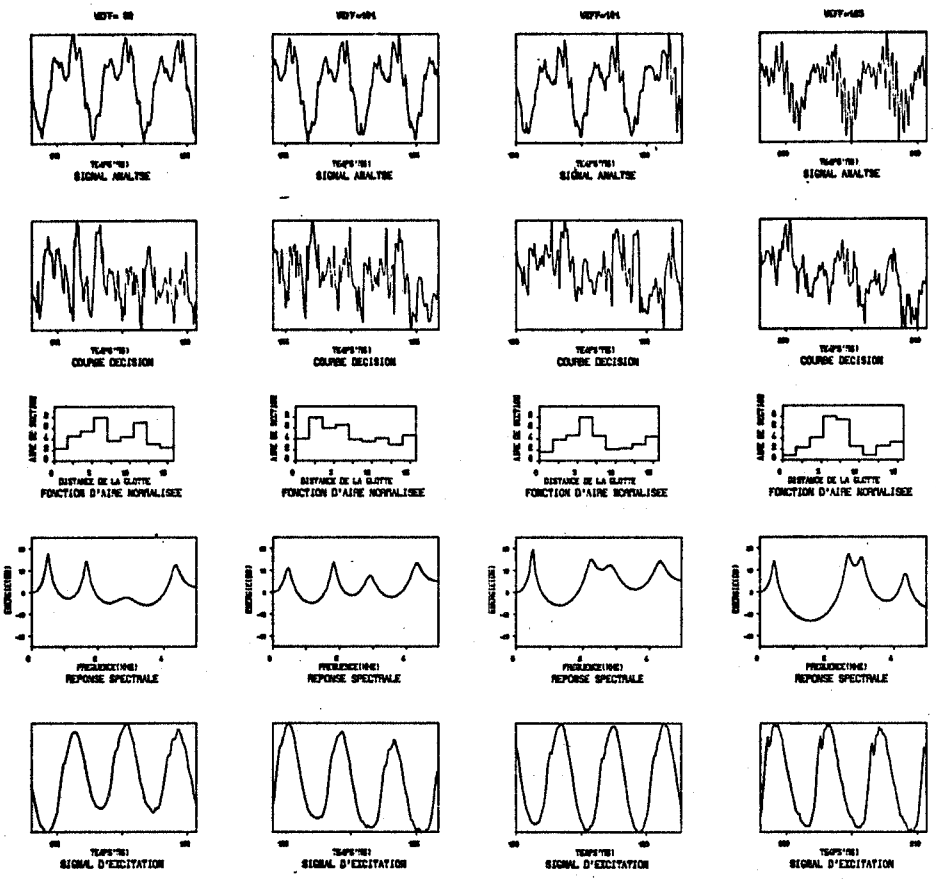
[1] I. EL MALLAWANY : "Etude de vocoders à prédiction linéaire,..., Extraction de la fonction d'aire du conduit vocal". USMG-INPG Grenoble (septembre 1975), n° d'inscription aux archives originales du centre de documentation du C.N.R.S. A.O. 11.924.

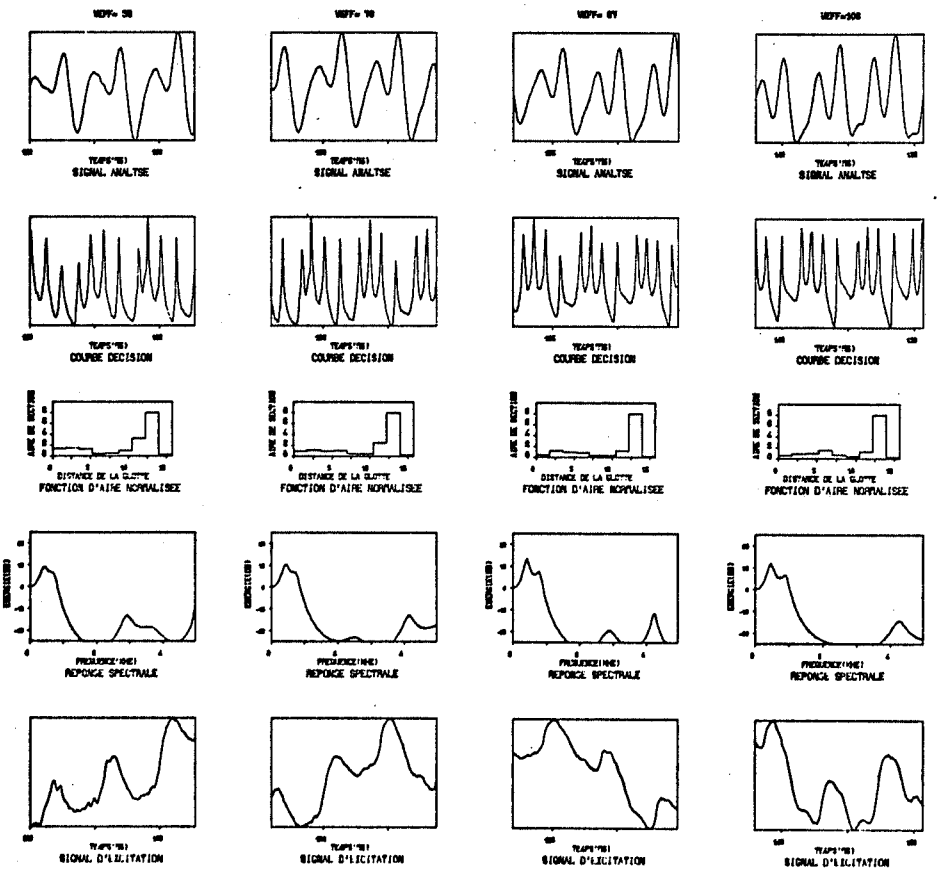
VIII - RESULTATS

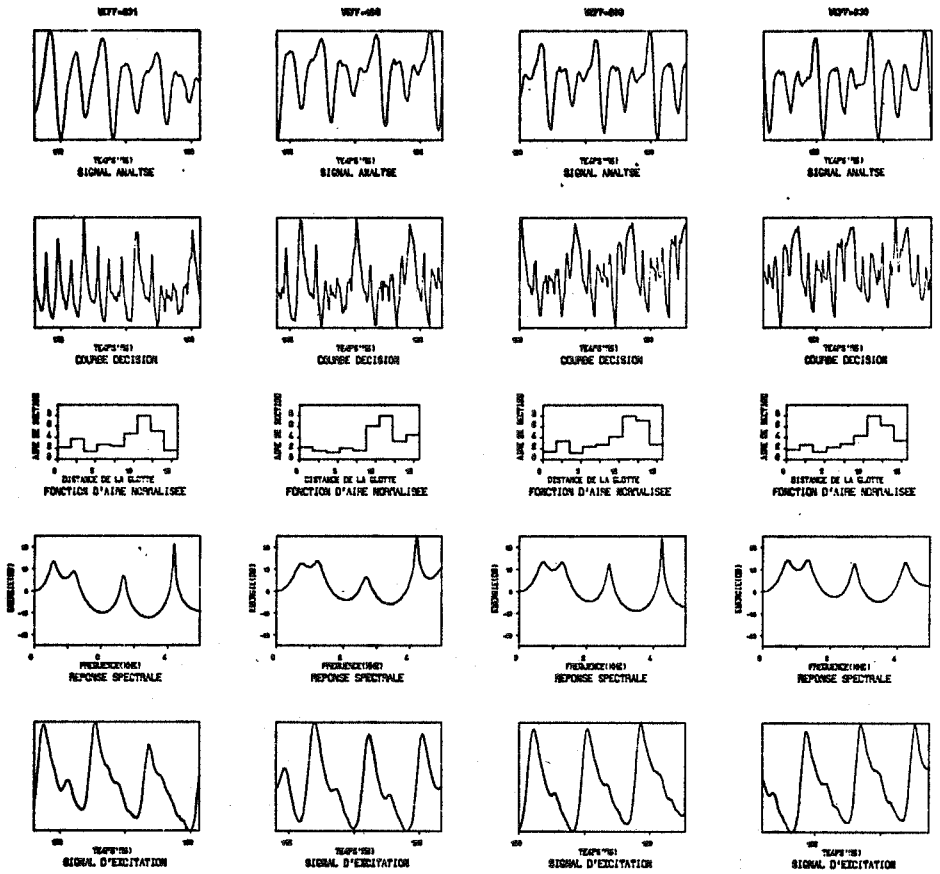
Les figures ci-dessous illustrent les résultats d'analyse obtenus sur deux diphtongues /ui/ et /ua/ respectivement. Chaque figure comprend 5 lignes et 4 colonnes. Chaque colonne représente les résultats d'un intervalle d'analyse. Le temps écoulé entre deux colonnes est de 6ms. La première ligne représente le signal analysé, la 3ème ligne la fonction d'aire extraite, la 4ème ligne le spectre correspondant et la 5ème le signal d'excitation obtenu par déconvolution.

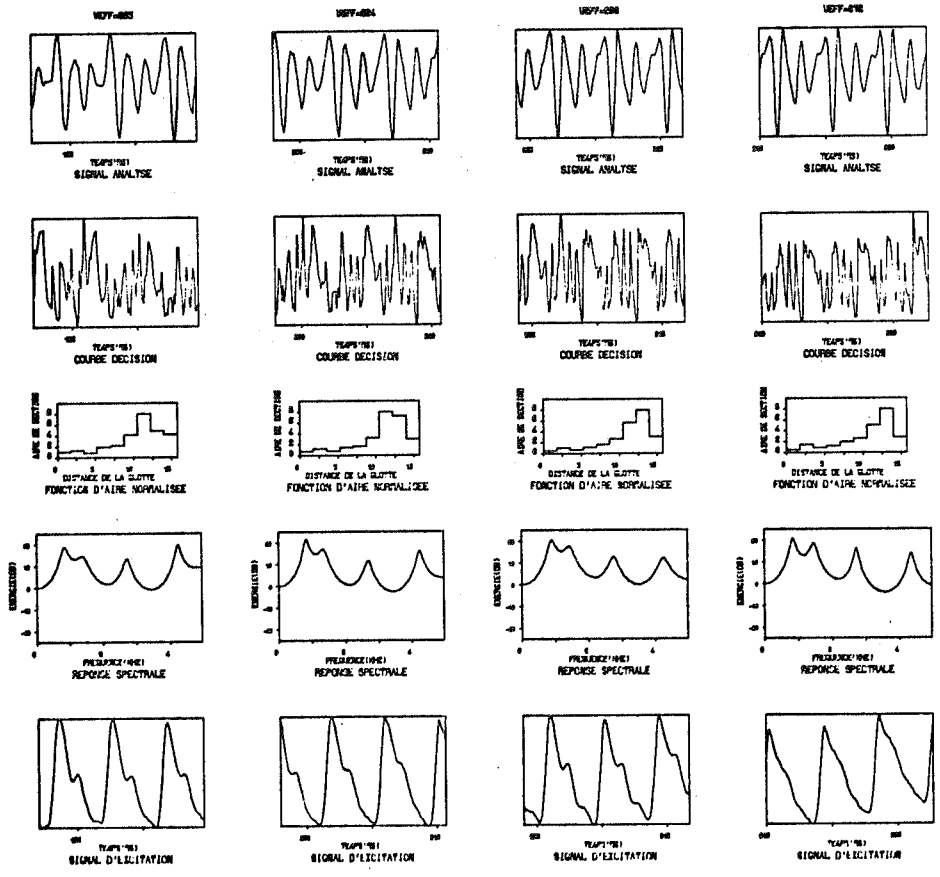

SIGNAL ANALYSE
COURBE DECISION
FONCTION D'AIRE NORMALISEE
REPONSE SPECTRALE
SIGNAL D'EXCITATION


SIGNAL ANALYSE
COURBE DECISION
FONCTION D'AIRE NORMALISEE
REPONSE SPECTRALE
SIGNAL D'EXCITATION


SIGNAL ANALYSE
COURBE DECISION
FONCTION D'AIRE NORMALISEE
REPONSE SPECTRALE
SIGNAL D'EXCITATION


SIGNAL ANALYSE
COURBE DECISION
FONCTION D'AIRE NORMALISEE
REPONSE SPECTRALE
SIGNAL D'EXCITATION


SIGNAL ANALYSE
COURBE DECISION
FONCTION D'AIRE NORMALISEE
REPONSE SPECTRALE
SIGNAL D'EXCITATION


SIGNAL ANALYSE
COURBE DECISION
FONCTION D'AIRE NORMALISEE
REPONSE SPECTRALE
SIGNAL D'EXCITATION

7èmes JOURNEES D' ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

DECTECTION DE LA MELODIE PAR AUTOCORRELATION NON LINEAIRE

C. GALAND — C.E.R. I.B.M. - LA GAUDE
D. ESTEBAN — C.E.R. I.B.M. - LA GAUDE
F. DUBUS — LABORATOIRE 190 C.N.R.S. Université de Nice

RESUME :

On étudie une méthode de détection de la mélodie essentiellement basée sur des pondérations de similitude entre séquences binaires.

Après avoir décrit cette méthode, on la compare, du point de vue mise en oeuvre, à la méthode de cepstre et à la méthode d'autocorrélation. Puis, des résultats de simulations sont exposés et conduisent à la définition d'algorithmes supplémentaires qui permettent la détection de la mélodie avec un rendement voisin de l'unité.

La fonction d'autocorrélation non linéaire est ensuite appliquée au codage prédictif de la parole. Des enregistrements de codage optimal par blocs à 10.000 éléments binaires par seconde seront présentés à la conférence.

SUMMARY :

Here is described a pitch-extractor which is based on bit-patterns comparisons.

A comparison is made between the relative complexity of three methods : the non linear autocorrelation analysis, the cepstrum analysis, and the short-term autocorrelation analysis. Then, the results of simulations are presented. In order to improve the quality of the pitch detection, some additional algorithms are defined.

Then, the non linear autocorrelation function is applied to the predictive coding of speech signals. Resulting records from an optimal block coding at a bit rate of 10 kbits/s will be played at the conference.

DETECTION DE LA MELODIE PAR AUTOCORRELATION NON LINEAIRE

C. GALAND, D. ESTEBAN, F. DUBUS

I.- INTRODUCTION

La mélodie, ou variation en fonction du temps de la période du fondamental des sons voisés, est un des paramètres essentiels dans des applications telles que l'analyse et la synthèse de la parole.

Les sons voisés résultent de l'excitation du conduit vocal par le son glottal qui est engendré par un phénomène de relaxation des cordes vocales. Ce signal est riche en harmoniques et sa fréquence fondamentale caractérise la hauteur du son émis. Le conduit vocal (pharynx, cavités buccales et nasales) est un ensemble de résonateurs dont les fréquences propres varient à chaque instant par suite des modifications ou obturations des différentes cavités. Le spectre du signal résultant du filtrage du signal glottal par le conduit vocal comporte donc des maxima, appelés formants, qui correspondent aux fréquences de résonance du conduit vocal et qui caractérisent le timbre de la voix.

Les méthodes d'extraction du fondamental sont extrêmement variées. On distingue cependant la méthode de cepstre et la méthode d'autocorrélation. Le cepstre est une double analyse spectrale du signal vocal associé à un traitement non linéaire qui permet de séparer les effets du signal glottal de ceux du conduit vocal. Cette méthode a été largement développée par A.M. NOLL et M.R. SCHROEDER |1,2,3|. La méthode d'autocorrélation est une des plus anciennes, et a été utilisée dans ce domaine comme dans d'autres où l'on cherche à mettre en évidence une périodicité quelconque ; elle a été expérimentée en particulier par V.N. SOBOLEV |4|, J.S. GILL |5|, et M.M. SONDHI |6|. Ces deux méthodes donnent de bons résultats, mais au prix d'une mise en oeuvre complexe et difficilement réalisable en temps réel.

La méthode d'autocorrélation non linéaire, qui est décrite ici, a été adaptée d'un algorithme proposé par A.H. FREI et al |7|. Cette méthode s'apparente, par de nombreux points communs, à la méthode d'autocorrélation, mais bénéficie sur celle-ci d'une mise en oeuvre beaucoup plus simple.

Cet article décrit deux applications de la méthode d'autocorrélation non linéaire. La première est la détection de la mélodie nécessaire à la synthèse par vocodeurs, et a été réalisée en temps réel |8|. La deuxième application est orientée vers le codage prédictif de la parole et a fait l'objet de simulations sur ordinateur.

II. - FONCTION D'AUTOCORRELATION NON LINEAIRE

Le signal vocal f(t) est échantillonné à la fréquence 1/T et modulé par impulsions et codage (MIC). L'échantillon f(nT) sera noté f_n.

Soit une séquence $\{f_1, f_2, \ldots, f_{2N}\} = \{f_n\}_1^{2N}$ de 2 N échantillons. Suivant le même principe que la méthode d'autocorrélation, la séquence $\{f_n\}_1^N$ est comparée à la séquence $\{f_n\}_{1+k}^{N+k}$, le décalage k variant de 0 à N. Soit les deux éléments binaires S_n et D_n repré-

sentant respectivement le signe de l'échantillon f_n et le signe de $f_n - f_{n-1}$. La méthode consiste à déterminer la similitude entre $\{S_n\}_1^N$ et $\{S_n\}_{1+k}^{N+k}$ d'une part et $\{D_n\}_1^N$ et $\{D_n\}_{1+k}^{N+k}$ d'autre part, k variant de 0 à N. Les critères de cette similitude sont basés sur la définition de deux paramètres : la "pente" P_n (table I) et la "courbure" C_n (table II).

A chaque décalage k, on fait N comparaisons de signe, de pente et de courbure, c'est-à-dire que l'on compare S_n et S_{n+k}, P_n et P_{n+k}, C_n et C_{n+k}, n variant de 1 à N. Chacune de ces N comparaisons donne lieu à un nombre $I_{n,k}$ (table III) et la similitude entre les séquences $\{f_n\}_1^N$ et $\{f_n\}_{1+k}^{N+k}$ est pondérée par la somme :

$$R(k) = \sum_{n=1}^{N} I_{n,k}$$

Le nombre $I_{n,k}$ peut prendre l'une des cinq valeurs $0, \pm 1, \pm 2$. La valeur 2 est obtenue lorsque le signe, la pente et la courbure des points f_n et f_{n+k} sont égaux. Les autres valeurs peuvent être obtenues de plusieurs manières, mais il est clair que $I_{n,0} = 2$ quel que soit n et que, par conséquent, $R(0) = 2N$. La fonction R(k) admet un maximum pour k=0, quel que soit le signal analysé $\{f_n\}_1^{2N}$. Si ce signal est périodique de période MT, on aura $f_{n+M} = f_n$ quel que soit $n \in [1,N]$ et par conséquent $R(M) = 2N = R(0)$.

Si le signal est sinusoïdal de période MT, et en supposant M pair, on aura $f_{n+\frac{M}{2}} = f_n$ quel que soit n, et $R(\frac{M}{2}) = 2N$.

Ces dernières remarques soulignent l'analogie de la fonction d'autocorrélation non linéaire R(k) avec la fonction d'autocorrélation de la séquence $\{f_n\}_1^{2N}$. La figure 1 représente ces deux fonctions pour une fenêtre de 140 échantillons de signal sinusoïdal de fréquence 200 Hz, échantillonné à 8kHz.

La table IV donne une évaluation comparative des charges de calcul nécessaires à la détermination de la fonction d'autocorrélation, du cepstre, et de la fonction d'autocorrélation non linéaire d'une séquence de 256 échantillons d'un signal f(t), la dynamique de codage étant de 12 bits. Les hypothèses pour le calcul du cepstre sont les suivantes : coefficients de FOURIER et de HAMMING codés sur 16 éléments binaires et table de logarithmes de 64 valeurs codées sur 8 éléments binaires. La réalisation en temps réel de l'une de ces méthodes peut s'envisager soit par microprocesseur, soit par l'emploi d'un circuit spécialisé. Dans l'une ou l'autre des éventualités, la fonction d'autocorrélation non linéaire bénéficie des avantages suivants :

- le calcul ne fait appel à aucune multiplication,
- la normalisation est automatique : quelle que soit la

séquence analysée, R(0) = 2N,

- la taille de la mémoire nécessaire est relativement faible.

A titre de comparaison, la figure 2 représente la fonction d'autocorrélation, le cepstre, et la fonction d'autocorrélation non linéaire d'une fenêtre de 32ms d'un son voisé.

III. - DETECTION DE LA MELODIE

La méthode d'autocorrélation et la méthode d'autocorrélation non linéaire ont été appliquées à la détection de la mélodie. Dans les deux cas, l'analyse du signal se fait séquence par séquence ; pour chaque séquence, on doit déterminer la nature voisée ou non voisée du signal, et, le cas échéant, la période du fondamental. Celle-ci pouvant atteindre 15ms, la longueur de la fenêtre d'analyse doit être supérieure ou égale à 30ms. Si l'on suppose une fréquence d'échantillonnage de 8 kHz, le nombre d'échantillons par séquence analysée peut être choisi égal à 256.

La détermination de la période du fondamental se fait de la façon suivante : de façon à doubler la définition, on interpole le signal, puis on calcule la fonction d'autocorrélation (ou la fonction d'autocorrélation non linéaire) et on détecte la position du maximum de cette fonction, hormis évidemment le maximum à l'origine. Si la valeur de ce maximum est inférieure à un certain seuil de corrélation pré-fixé, la fenêtre est considérée comme non voisée. Dans le cas contraire, la période du fondamental est donnée par la position du maximum.

Une étude systématique a permis de déterminer les possibilités et les limites de ces deux détecteurs. Il résulte de cette étude que les deux méthodes donnent des résultats comparables, avec un rendement de l'ordre de 80%, c'est-à-dire que sur 100 fenêtres voisées analysées,on obtient en moyenne 20 valeurs erronnées de la période du fondamental. Cette estimation a été faite en prenant pour référence la mélodie détectée sur le signal glottal [9].

Les erreurs de détection sont essentiellement de trois types. Les plus nombreuses proviennent de la présence, dans certains sons, d'harmoniques d'amplitude très élevée, ce qui entraîne la détermination d'une valeur inférieure à la période du fondamental. Le deuxième type d'erreurs est le dédoublement qui peut intervenir lorsque la corrélation est mieux marquée sur deux périodes du fondamental, par exemple lors d'une variation très rapide de la mélodie. Le troisième type d'erreurs se rencontre parfois dans les transitions son voisé/silence ou inversement. Dans ce cas, la fonction d'autocorrélation peut marquer un maximum assez important pour une valeur quelconque du décalage.

Ces constatations nous ont conduits à rechercher des solutions tendant à réduire le nombre de déterminations erronnées de la période du fondamental. La méthode adoptée pour les erreurs dues aux harmoniques est un prétraitement du signal vocal à l'aide d'un

filtre passe-bas dont les caractéristiques sont fonction des limites fréquentielles du fondamental. La fréquence de ce dernier étant comprise entre 70 Hz et 400 Hz, on utilise un filtre dont le module de la réponse harmonique décroît approximativement de 12 dB par octave entre 70 Hz et 800 Hz. Il en résulte une atténuation systématique de 12dB des harmoniques par rapport à la fréquence fondamentale, quelle que soit cette fréquence. La figure 3 donne un exemple d'une fenêtre de 32ms. Sa fonction d'autocorrélation non linéaire (fig.4) présente des pics dûs aux harmoniques, et l'analyse de cette fenêtre peut conduire à une erreur de détection. Les figures 5 et 6 représentent le même signal après filtrage et sa fonction d'autocorrélation non linéaire. Les pics parasites ont disparu, la corrélation est mieux marquée, et la détermination de la période du fondamental ne présente plus d'ambiguité. Dans tous les cas analysés, ce prétraitement a permis d'éliminer complètement les erreurs dues aux harmoniques, portant ainsi le rendement à 97%.

Les erreurs du deuxième type peuvent être éliminées en grande partie si l'on effectue une détection plus sophistiquée du maximum de la fonction d'autocorrélation non linéaire. En particulier, on peut considérer un seuil de corrélation non plus fixe, mais adaptatif, dépendant tout à la fois de la valeur précédente du maximum de corrélation et de la décision précédente de voisement.

Les erreurs du troisième type peuvent être éliminées si l'on détecte les transitions entre les sons voisés et les silences, et inversement. L'algorithme est le suivant ; pour chaque fenêtre $\{f_n\}_1^{2N}$ de 2N échantillons, on détermine :

$$M_1 = \max_{n=1,N} |f_n|$$

$$M_2 = \max_{n=N+1,2N} |f_n|$$

Si le rapport M_1/M_2 est plus grand que 2 ou plus petit que 0,5, et si la fenêtre précédente n'était pas voisée, on classe la fenêtre actuelle comme non voisée ; si la fenêtre précédente était voisée, on augmente le seuil de corrélation.

La méthode d'autocorrélation non linéaire, associée au prétraitement du signal vocal, et aux algorithmes heuristiques qui ont été décrits, permet une détection de la mélodie avec un rendement voisin de l'unité.

IV. - PREDICTION A LONG TERME

La méthode d'autocorrélation non linéaire a été appliquée au codage prédictif à long terme de la parole. Les sons voisés présentant une grande similitude d'une période du fondamental à l'autre, la méthode consiste à prédire le signal, puis à quantifier l'erreur de prédiction. Ceci permet de diminuer l'énergie du signal d'erreur de quantification, et par conséquent d'augmenter le rapport signal sur bruit. Cette méthode, proposée par B.S. ATAL et M.R.SCHROEDER

|10| a été appliquée en particulier par D. ESTEBAN et J. MENEZ |11| au codage par blocs.

Soit $\{f_n\}_1^N$ un bloc de N échantillons à coder. La similitude de ce bloc par rapport aux échantillons passés est déterminée par autocorrélation non linéaire. La fonction d'autocorrélation non linéaire R(k) est calculée pour L valeurs du décalage k du bloc sur les échantillons passés. Pour chaque décalage k, on calcule R(k) par N incrémentations :

$$R(k) = \sum_{n=1}^{N} I_{n,k} \qquad k \in [0,L]$$

En pratique, la longueur des blocs à coder est de 8ms, ce qui correspond à N=64 échantillons à 8kHz, et la similitude est cherchée sur une longueur de 24ms (L=192). La position du maximum de R(k) donne le décalage M de plus grande similitude. La valeur M est le plus souvent égale à la période du fondamental, mais correspond parfois à un multiple ou à une fraction de cette période.

Les échantillons prédits sont donnés par $\{\beta f_{n-M}\}_{n=1}^{M}$, le facteur d'homothétie β permettant de tenir compte des variations d'énergie d'un bloc à l'autre. La détermination de β se fait en minimisant l'énergie de l'erreur de prédiction :

$$E = \sum_{n=1}^{N} (f_n - \beta f_{n-M})^2 \quad \text{minimum}$$

$$\frac{\partial E}{\partial \beta} = 0 \Longrightarrow \beta = \frac{\sum_{n=1}^{N} f_n . f_{n-M}}{\Sigma f_{n-M}^2}$$

La figure 7 représente un exemple de codage par bloc, avec ou sans prédiction à long terme. La longueur du bloc est de 8ms et chaque échantillon est codé à l'aide d'un élément binaire. L'emploi d'un prédicteur à long terme améliore le rapport signal sur bruit d'environ 4dB en moyenne, pour le type de codage considéré (environ 10.000 éléments binaires/seconde).

V. - CONCLUSION

Une méthode d'extraction de la mélodie a été décrite. Bien qu'étant basée sur le même principe que l'analyse par autocorrélation, la méthode d'autocorrélation non linéaire bénéficie sur cette dernière d'une mise en oeuvre beaucoup plus simple. Des simulations ont permis de déterminer les possibilités et les limites de la détection. Après examen des résultats obtenus, des solutions simples ont été proposées pour améliorer le rendement. La méthode d'autocorrélation, associée à ces algorithmes, permet la détection de la mélodie avec un taux d'erreur très faible.

D'autre part, la méthode a été appliquée au codage prédic-

tif à long terme de la parole. Des enregistrements présentés à la conférence permettront d'apprécier l'amélioration apportée par un prédicteur à long terme dans le cas d'un codage optimal par blocs à 10.000 éléments binaires par seconde.

VI. - REFERENCES

1 A.M. NOLL "SHORT-TIME SPECTRUM AND CEPSTRUM TECHNIQUES FOR VOCAL-PITCH EXTRACTION" J.A.S.A. Vol. 36 p 296-302 - 1964

2 A.M. NOLL and M.R. SCHROEDER "SHORT-TIME CEPSTRUM PITCH DETECTION" - J.A.S.A. Vol. 36 p 1030 - 1964

3 A.M. NOLL "CEPSTRUM PITCH DETERMINATION" J.A.S.A. Vol. 41 n° 2 pp 293-309 - 1967

4 V.N. SOBOLEV "EXPERIMENTAL INVESTIGATIONS OF THE CORRELATION METHOD FOR THE DISCRIMINATION OF THE FUNDAMENTAL SPEECH TONE" Soviet Physics-Acoustics Vol. 14 n° 3 - January-March 1969

5 J.S. GILL "AUTOMATIC EXTRACTION OF THE EXCITATION FUNCTION OF SPEECH WITH PARTICULAR REFERENCE TO THE USE OF CORRELATION METHODS" Proc. Third Internat'l Congr. Acoust. Stuttgart, Germany Sept. 1959

6 M.M. SONDHI "NEW METHODS OF PITCH EXTRACTION" I.E.E.E. Trans.on Audio and Elect.Vol.AU-16 n° 2 June 1968

7 A.H. FREI, H.R. SCHINDLER, P. VETTIGER, E. VON FELTEN "ADAPTATIVE PREDICTIVE SPEECH CODING BASED ON PITCH-CONTROLLED INTERRUPTION/REITERATION TECHNIQUES - Int. Conf. on Communications, June 11, 12, 13, 1973. Seattle, Washington.

8 C. GALAND "ETUDE ET REALISATION D'UN DETECTEUR DE MELODIE EN TEMPS REEL" - Thèse 3ème cycle Nice - Nov.1973

9 J.P. PECKELS, V. RISO "ON THE USE OF AN ACCELEROMETER AS A PITCH EXTRACTOR" 5ème Congrès International d'Acoustique - A-25 Liège Sept. 1965

10 B.S. ATAL, M.R. SCHROEDER "ADAPTATIVE PREDICTIVE CODING OF SPEECH SIGNALS" B.S.T.J. Oct.1970, pp 1973-1986

11 D. ESTEBAN, J. MENEZ "LOW BIT RATE VOICE TRANSMISSION BASED ON TRANSVERSAL BLOCK CODING" 91st ASA Meeting, Washington, April 1976.

TABLE I

D_{n-1}	D_n	P_n
1	1	1
1	0	0
0	1	0
0	0	-1

TABLE II

D_{n-1}	D_{n+1}	C_n
1	1	0
1	0	1
0	1	-1
0	0	0

TABLE III

P_n ou C_n	P_{n+k} ou C_{n+k}	$I^1_{n,k}$ ou $I^1_{n,k}$
1	1	$1-2\theta$
1	0	0
1	-1	-1
0	1	0
0	0	$1-2\theta$
0	-1	0
-1	1	-1
-1	0	0
-1	-1	$1-2\theta$

$\theta = 0$ si $S_n = S_{n+k}$

$\theta = 1$ si $S_n \neq S_{n+k}$

$$I_{n,k} = I^1_{n,k} + I^2_{n,k}$$

TABLE IV. Evaluation comparative de la charge de calcul nécessaire à la détermination de la fonction d'autocorrélation, du cepstre, et de la fonction d'autocorrélation non linéaire d'une séquence de 256 échantillons . Le nombre d'opérations est calculé par instant d'échantillonnage . La taille des mémoires est indiquée en nombre de mots x nombre d' e.b par mot.

	Autocorrélation	Cepstre (DFT)	Cepstre (FFT)	Autocorrélation Non Linéaire
Additions	128	2046	34	128
Multipli-cations	128	2055	39	0
Lectures de table	0	10		128
Mémoire vive	512x12	256x12		512x1
Mémoire morte	0	256x16 + 128x16 + 64x8		64x3

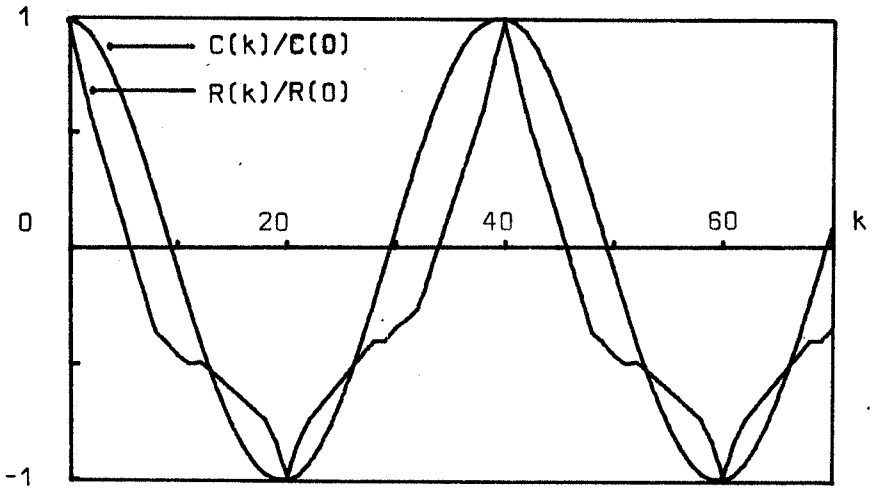
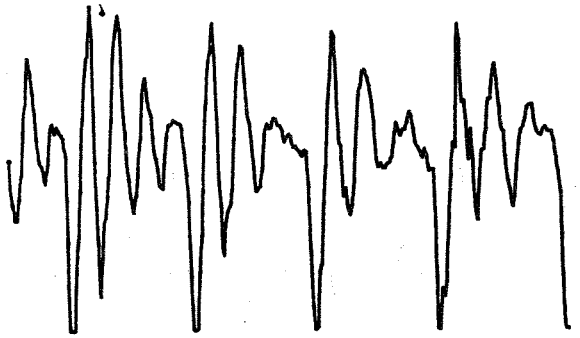


Fig 1. Fonction d'autocorrélation C(k)/C(0) et Fonction d'autocorrélation non linéaire R(k)/R(0) d'une fenêtre de 140 échantillons d'un signal sinusoidal de fréquence 200 Hz échantillonné à 8 kHz.

$$R(k) = \sum_{n=1}^{n=N} I_{n,k} \qquad C(k) = \sum_{n=1}^{n=N} f_n . f_{n+k}$$



Fenêtre de 32 ms de signal vocal

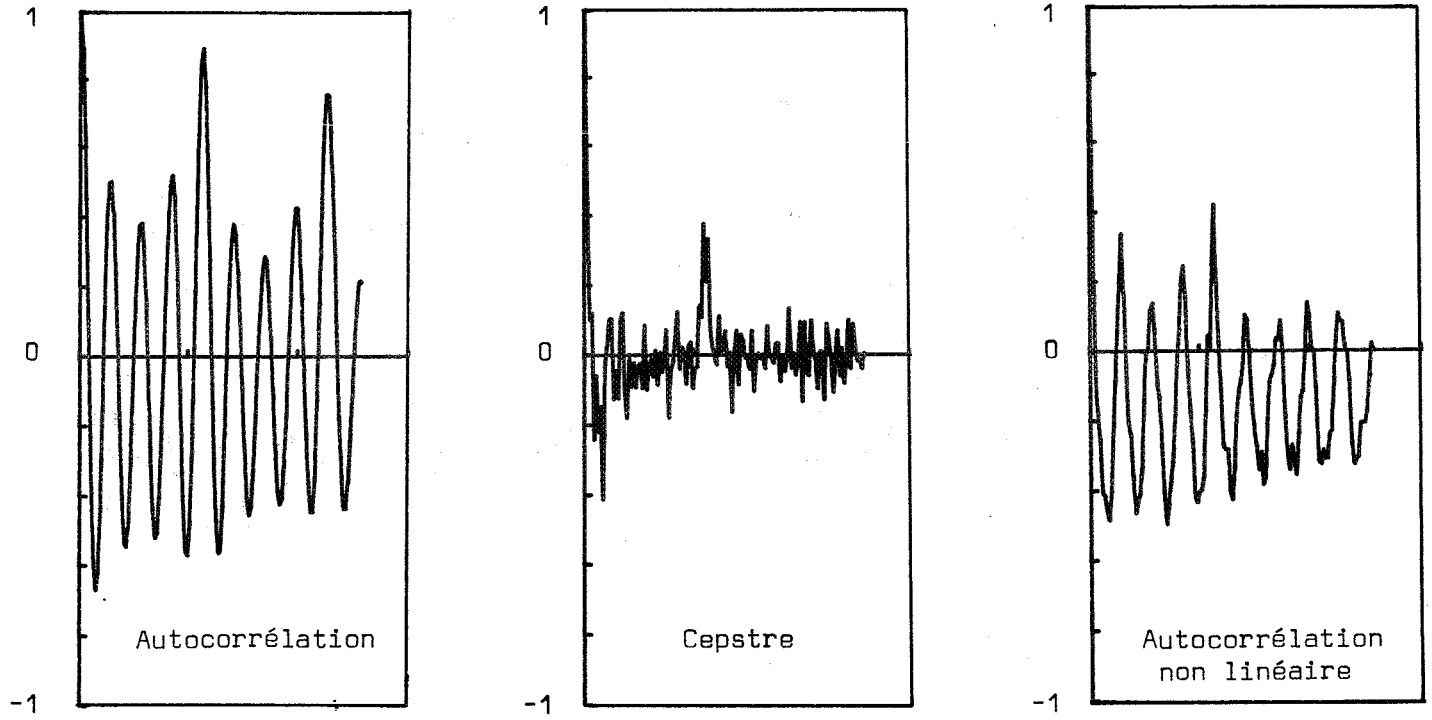


Fig 2. Comparaison entre la fonction d'autocorrélation, le cepstre, et la fonction d'autocorrélation non linéaire d'une fenêtre de 32 ms de signal vocal. Les trois fonctions sont normalisées à leur valeur à l'origine.

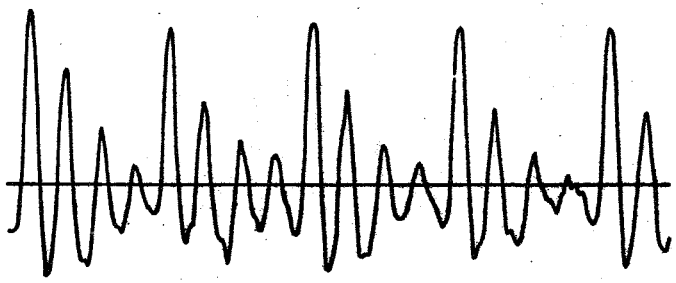


Fig 3. Fenêtre de 32 ms de signal vocal.

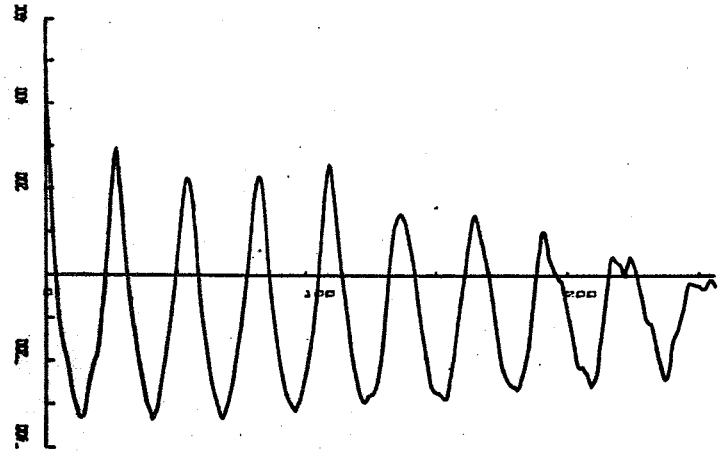


Fig 4. Fonction d'autocorrélation non linéaire du signal vocal.

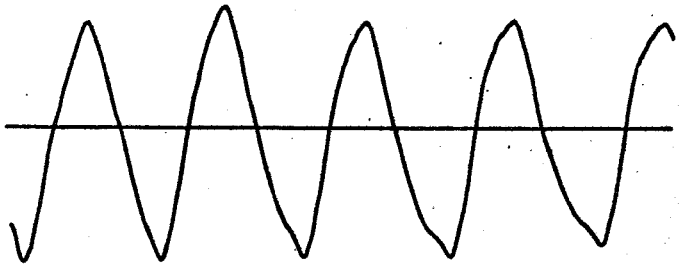


Fig 5. Signal vocal prétraité.

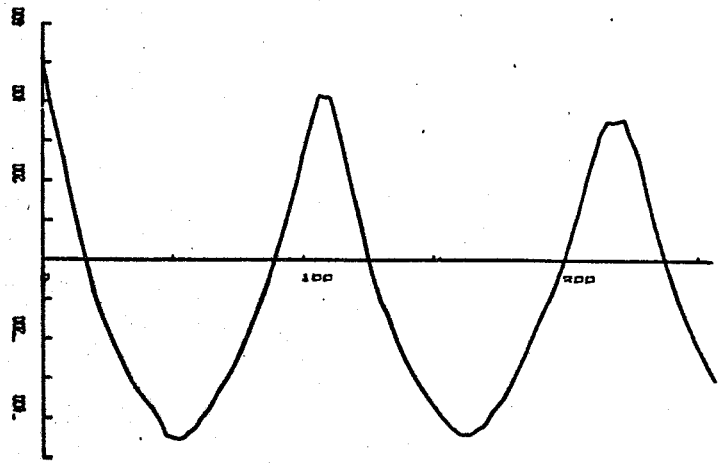
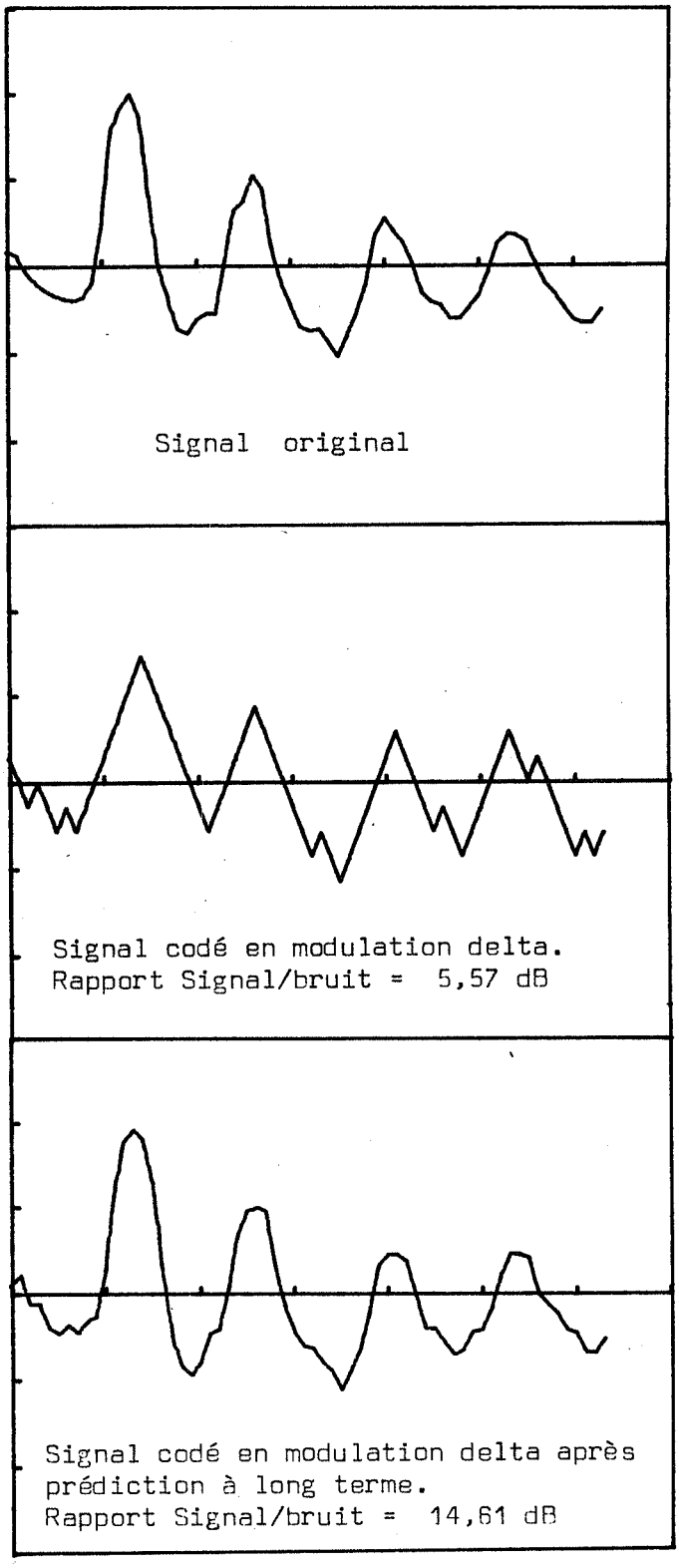


Fig 6. Fonction d'autocorrélation non linéaire du signal prétraité.

Fig 7. Exemples de codage optimal à 1 élément binaire par échantillon d'un bloc de 8 ms d'un signal vocal échantillonné à 8 kHz.



7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

UNE METHODE DE REPRESENTATION DU SIGNAL VOCAL EN BASE ADAPTATIVE

Marie-Christine HATON et Jean-Paul HATON
Laboratoire d'Informatique
Université de NANCY 1
C.O. 140 54037 - NANCY Cedex

RESUME

On propose une méthode de représentation du signal vocal dans une base de filtres exponentiels. La base optimale est obtenue par une technique de filtrage inverse en considérant, dans le plan complexe, des points situés à l'intérieur du cercle de rayon 1.

Cette base peut être adaptée à un vocabulaire donné, ou à un locuteur donné, selon les applications.

Un exemple est donné dans le cadre du projet SIRENE de rééducation vocale des mal-entendants.

SUMMARY

A method is proposed for representing the speech signal in an exponential filters base. An inverse filtering technique yields the optimal base by considering points inner the unit circle in the complex plane.

This base can be adapted to a given vocabulary, or a given speaker, according to the application.

Some preliminary results have been obtained in our SIRENE project for speech reeducation of deaf children.

UNE METHODE DE REPRESENTATION DU SIGNAL VOCAL EN BASE ADAPTATIVE

Marie-Christine HATON - Jean-Paul HATON
Laboratoire d'Informatique
Université de Nancy 1

1 . INTRODUCTION

La caractérisation d'un signal, du signal de parole en particulier, par un ensemble de paramètres dépend essentiellement des propriétés sur lesquelles on veut mettre l'accent. De façon générale, on peut classer les paramètres en deux grandes catégories : les uns permettront de caractériser les propriétés de transmission d'un système invariant dans le temps ; les autres, faisant abstraction du caractère redondant du signal, correspondront dans la mesure du possible à ses traits distinctifs. Bien entendu, les deux catégories peuvent interférer. C'est ainsi que les méthodes de prédiction linéaire (1) appliquées à la parole, tout en isolant des paramètres pertinents, permettent de retrouver la configuration du conduit vocal pendant l'émission du signal. Pour un système invariant dans le temps, la classe de représentations du signal la plus importante est celle où il se résoud en ses composantes sinusoidales (base des polynômes de Tchebycheff par exemple (2)); la décomposition en série de Fourier qui peut se calculer facilement grâce aux algorithmes de transformée rapide de Fourier (FFT) est la plus couramment utilisée des méthodes numériques spectrales.

Pourtant, cette décomposition est mal adaptée à la représentation des signaux transitoires. La figure 1 donne le spectre de fréquences calculé par FFT d'un signal composé de trois sinusoides amorties de la forme $\exp(-\alpha t).\sin(\omega t+\varphi)$

Le signal est échantillonné à la fréquence f_e=10 kHz = 1/T.
Chaque sinusoide échantillonnée est de la forme :

$$f(nT) = \exp(-\alpha nT).\sin(n\omega T+\varphi) = \rho^n.\sin(n\theta+\varphi)$$

Les quantités ρ et θ caractérisent l'amortissement et la fréquence.

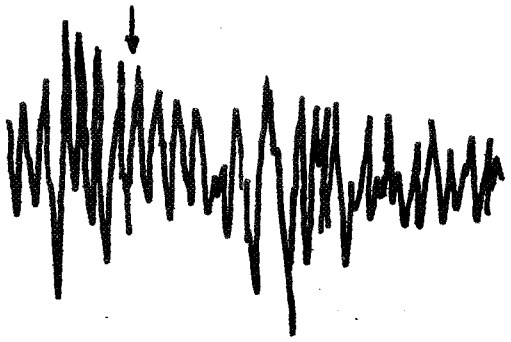


fig. 1

On peut remarquer que la deuxième composante n'apparaît que comme une légère déformation de la courbe.

Nous développons ici une méthode de caractérisation du signal faisant intervenir ses projections sur un ensemble de filtres numériques exponentiels; chacun de ces filtres correspond à une fréquence et un amortissement donnés.

La construction de cet ensemble de filtres de base dépend de l'application envisagée (reconnaissance d'éléments de parole, reconnaissance de locuteurs...). Notre application rentre dans le cadre du système SIRENE (Système Interactif de Rééducation des Enfants Non-Entendants) que nous développons actuellement au Laboratoire d'Informatique. La base optimale pourra être construite à partir de productions acceptables de l'élève dans le cas idéal ou, à défaut, à partir des productions d'un autre sujet de même âge présentant des caractéristiques vocales analogues.

2 . CONSTRUCTION DES FILTRES NUMERIQUES

L'analyse par filtrage inverse appliquée à un signal linéaire conduit à rechercher la fonction de transfert du filtre dont la réponse impulsionnelle se confond le mieux (au sens d'un certain

critère) avec ce signal. Pour les sons voisés non nasalisés, le conduit vocal peut être assimilé à un filtre récursif d'ordre P du type $H(z) = \sigma / 1+\sum_{p=1}^{P} a_p z^{-p}$, ce qui revient à considérer que la sortie du filtre à un instant donné ne dépend que de la valeur de l'entrée à ce même instant et des valeurs de sortie aux P instants précédents soit : $S(nT) + \sum_{p=1}^{P} a_p S((n-p)T) = \sigma . e(nT)$.
Pour les consonnes nasales pures, le conduit oral, correspondant à une branche fermée, provoque l'apparition de zéros dans la fonction de transfert réelle. Dans le cas des voyelles nasalisées, c'est le conduit nasal qui, intervenant comme une branche ouverte, fait apparaître des pôles et des zéros supplémentaires. Comme pour les sons non voisés, les zéros sont situés à l'intérieur du cercle-unité dans le plan z : il est alors possible (3) d'approcher chaque facteur du numérateur par un facteur au dénominateur de la fonction de transfert.

Le modèle H(z) précédent est ainsi adapté à la représentation de l'ensemble des signaux vocaux. L'analyse par filtrage inverse fourn it la suite des paramètres $\{a_p\}$ de H(z).

Le calcul de $|H(z)|^2$ pour z se déplaçant dans le plan z sur le cercle unité donne le spectre de puissance de la réponse impulsionnelle du filtre inverse. La figure 2a correspond à l'exemple précédent. On peut remarquer que la deuxième composante n'est pas mieux différentiée que sur la figure 1. On évite seulement la modulation du spectre due à la fréquence fondamentale lorsque le son est voisé. De plus on assiste (tout comme sur le spectre de Fourier calculé directement) à un déplacement des maxima par rapport aux valeurs exactes qui peut être important si les amortissements sont grands.

Si l'on fait le calcul du module carré de H(z) sur des demi-cercles de rayon ρ inférieur à 1, le spectre obtenu évolue pour l'exemple choisi comme le montre la figure 2b. On voit que pour la valeur du rayon correspondant à l'amortissement de chacune des composantes, on obtient un maximum local accentué.
La deuxième composante, en particulier, est maintenant bien différentiée.

$\begin{cases} f_1 = 700 \text{ Hz} & \rho_1 = 0.90 \\ f_2 = 1200 \text{ Hz} & \rho_2 = 0.82 \\ f_3 = 2750 \text{ Hz} & \rho_3 = 0.86 \end{cases}$

fig. 2a

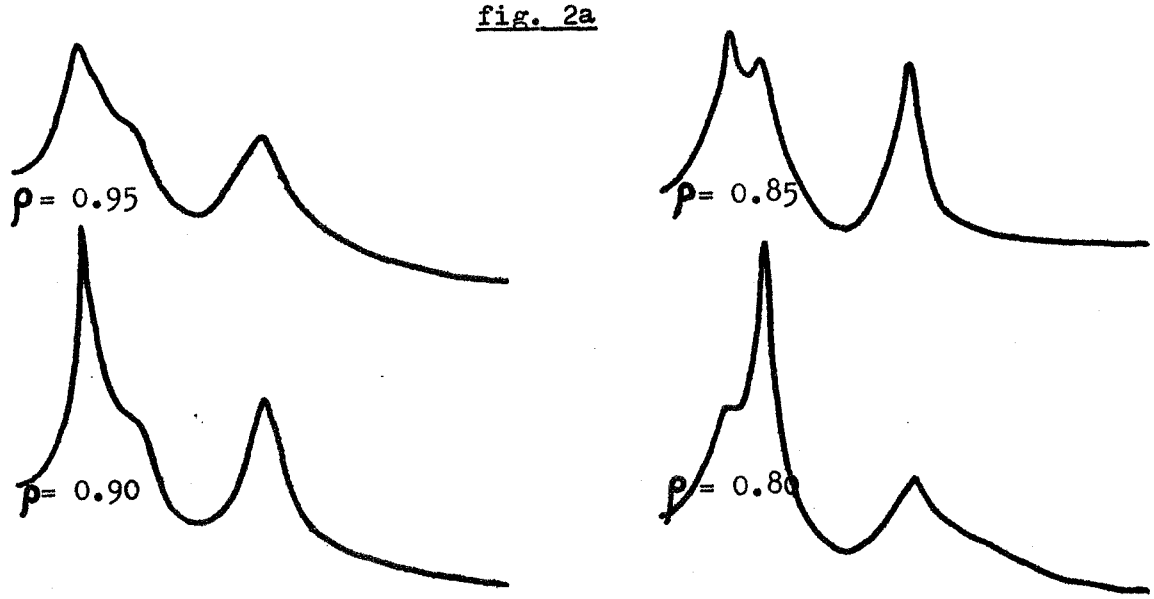


fig. 2b

Ce résultat présente également un intérêt dans les cas où la présence d'un zéro peut masquer un formant (c'est le cas des nasales par exemple où le deuxième formant est souvent escamoté dans le spectre).

Soit $A(z)$ = dénominateur de $H(z) = 1 + \sum_{p=1}^{P} a_p z^{-p}$.

Un minimum dans le calcul de $|A(z)|^2$ correspondra à la proximité d'un pôle de la fonction de transfert. Soit $z_o = \rho . \exp(i\theta)$ un point du plan z situé dans le cercle-unité ($0<\rho< 1$). Par la suite R et M désigneront les quantités :

$R = 2\rho.\cos\theta$ = 2 x partie réelle de z_o

et $M = \rho^2$ =module carré de z_o .

On aura donc : $(1-z_o z^{-1})(1-z_o^* z^{-1}) = (1-Rz^{-1}+Mz^{-2})$.

La quantité $A(z_o)$ a pour expression :

$$A(z_o) = 1 + \sum_{p=1}^{P} a_p \cdot \frac{\exp(-ip\theta)}{\rho^p}$$

$$A(z_o) = (1 + \sum_{p=1}^{P} a_p \frac{\cos p\theta}{\rho^p}) - i (\sum_{p=1}^{P} a_p \frac{\sin p\theta}{\rho^p})$$
$$= A_r(z_o) - i A_i(z_o)$$

$A(z_o) = 0$ si le point $\exp(i\theta)$ est un zéro de $A(z)$ donc un pôle de $H(z)$.

Nous avons recherché les expressions en z des fonctions de transfert $F_1(z)$ et $F_2(z)$ des filtres numériques qui, recevant à l'entrée les données 1, a_1, a_2, .., a_P, fournissent respectivement à la sortie à l'instant P x T les quantités $A_r(z_o)$ et $A_i(z_o)$.
Leurs réponses impulsionnelles doivent être respectivement égales à :

$$\sum_{p=0}^{P} \frac{\cos p\theta}{\rho^p} . \delta(t-pT) \quad \text{et} \quad \sum_{p=0}^{P} \frac{\sin p\theta}{\rho^p} . \delta(t-pT) .$$

On peut remarquer que ρ étant inférieur à 1 ces filtres ne vérifient pas le critère de stabilité. Ce point n'est pas gênant pour le calcul puisque nous ne nous intéressons qu'à la P+1 ème valeur de la sortie.

Le calcul nous a conduit aux expressions suivantes :

$$F_1(z) = \frac{M - R/2\, z^{-1}}{M - R\, z^{-1} + z^{-2}} \quad \text{et} \quad F_2(z) = \frac{\sqrt{M - R^2/4}\; z^{-1}}{M - R\, z^{-1} + z^{-2}}$$

Nous appellerons filtres (R,M) l'ensemble des deux filtres F_1 et F_2 correspondant à des valeurs données de la fréquence $f = \frac{\theta}{2\pi T}$ et de l'amortissement, donc à des valeurs données de θ et de ρ.

3 . RECHERCHE D'UNE BASE OPTIMALE DE FILTRES (R,M)

Pour construire une base de représentation optimale pour un ensemble de signaux, on recherche par approximation les minima des quantités $|A(z_o)|^2$ calculées pour des points z_o situés sur des cercles de rayon inférieur à l'unité.

Une solution identique consiste à rechercher les pôles des fonctions $H(z)$ et de construire les filtres à partir des valeurs obtenues.

Une étude statistique systématique permet de sélectionner

un ensemble de couples de filtres (R,M) qui constituera la base optimale pour les signaux étudiés.

La recherche de cette base dépendra de l'application considérée :

- en reconnaissance vocale, elle pourra être construite à partir de formes de référenses émises par plusieurs locuteurs,

-en reconnaissance de locuteurs, les différents filtres tiendront compte des caractéristiques de chacun des locuteurs pour un vocabulaire limité,

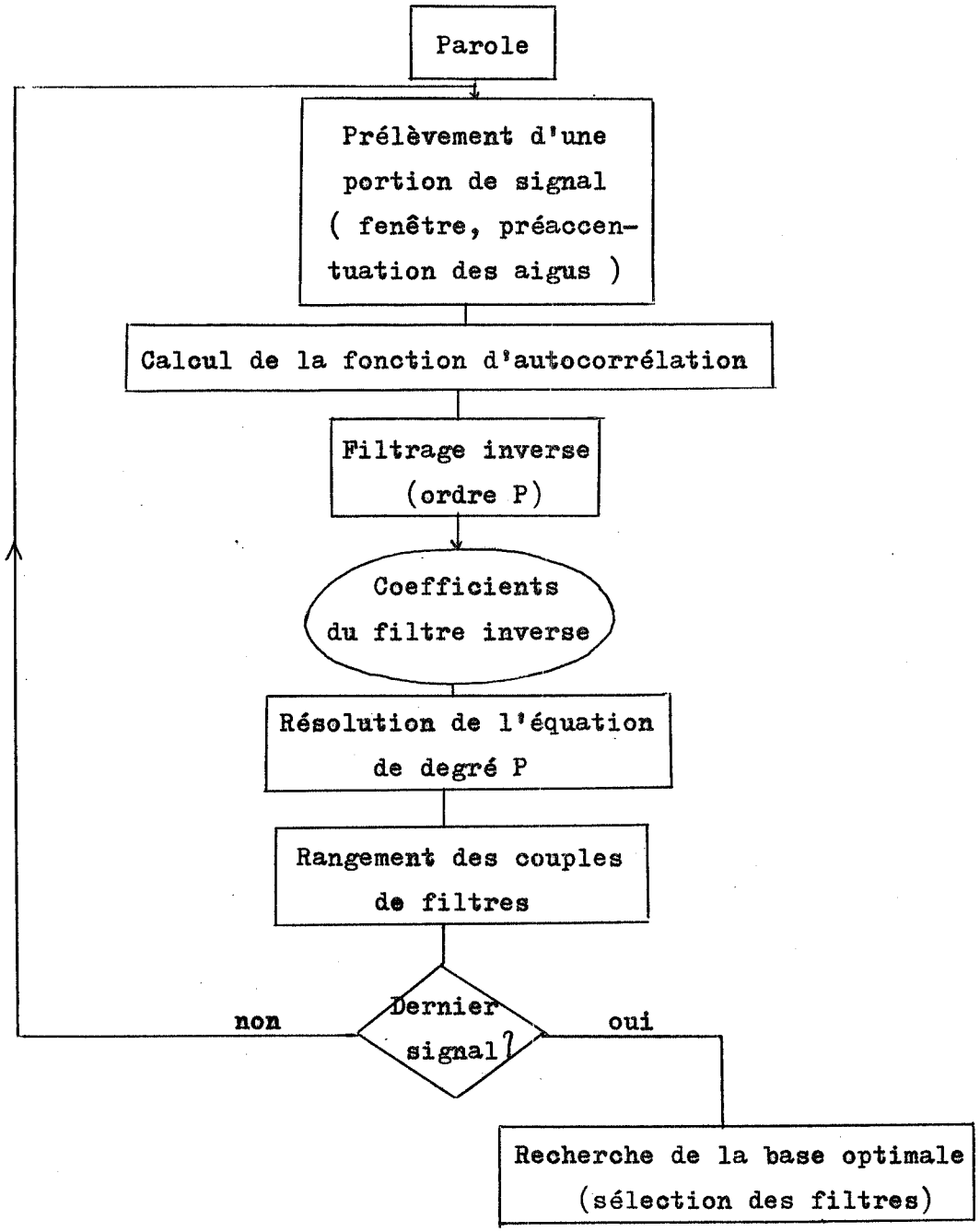
- en rééducation vocale, qui nous intéresse plus spécialement dans cette étude, la base retenue tient compte des caractéristiques vocales désirées pour l'élève à un instant donné de sa rééducation. Cette base peut être obtenue à partir d'un ensemble limité de sons ou de mots pour permettre une meilleure discrimination sur cet ensemble. Elle peut en particulier évoluer au cours de la rééducation pour mieux s'adapter au locuteur.

4 . PROJECTIONS D'UN SIGNAL SUR UNE BASE DE FILTRES (R,M)

Nous appelons projections d'un signal caractérisé par la suite $\{1, a_1, a_2, \ldots, a_p\}$ sur un couple de filtres (R,M), les quantités A_r et A_i telles qu'elles ont été définies auparavant. Pour identifier leur calcul avec celui de la convolution du signal d'entrée et des réponses impulsionnelles des filtres numériques, il est nécessaire d'entrer les coefficients dans l'ordre inverse $a_p, a_{p-1}, \ldots, a_1, 1$.

Sur une base de N couples de filtres, un signal sera caractérisé par une chaîne de 2N éléments. Dans notre application, la représentation par une chaîne de ce type des productions de l'élève fournit une appréciation de leur qualité et permet d'évaluer les progrès réalisés.

A titre d'exemple, nous avons recherché une base optimale pour la représentation de dix classes de voyelles orales pronocées par un locuteur. La figure 3 montre le principe général du calcul.



Recherche de la base optimale

fig. 3

La base retenue est constituée de sept filtres (R,M). Sur la figure 4 sont représentées les quantités $1 / A_r^2 + A_i^2$ calculées à partir des projections sur cette base, qui correspondent en quelque sorte aux "spectres" des signaux donnés par les filtres.

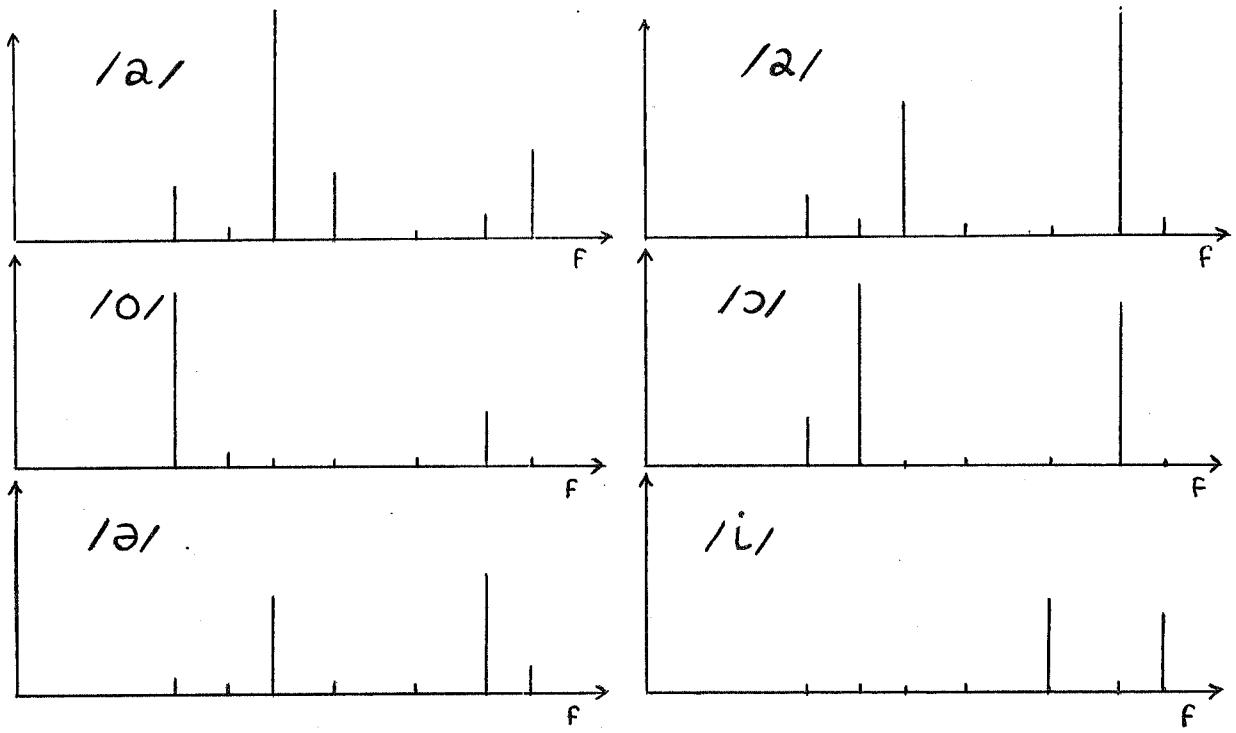


fig. 4

On constate une certaine disparité dans les "spectres" des différentes voyelles représentées. Pour l'instant, les critères de sélection des filtres (R,M) en vue de la constitution de la base optimale sont empiriques. Il est clair que la formalisation de ces critères doit permettre d'adapter cette base, aussi bien à la caractérisation d'un locuteur qu'à la reconnaissance d'un ensemble de signaux.

5 . CONCLUSION

Nous avons présenté une méthode de représentation du signal vocal dans une base optimisée de filtres exponentiels. L'examen des premiers résultats obtenus dans la représentation de voyelles montre l'intérêt de la méthode pour caractériser une voix ou un ensemble de sons. Pour réduire le temps de calcul, tout en conservant des principes analogues, nous recherchons maintenant une représentation applicable directement au signal temporel en construisant une base optimale de filtres inverses.

REFERENCES

(1) J. D. MARKEL, A. H. GRAY "Linear prediction of speech", S.C.R.L. Monograph Nr 10.

(2) M. C. HATON, J.P. HATON "Essai de caractérisation des voix d'enfants sourds par analyse polynomiale de la mélodie", 6èmes Journées d'Etudes sur la Parole, Toulouse, Mai 1975.

(3) B. S. ATAL, S. L. HANAUER "Speech analysis and synthesis by linear prediction of the speech wave", J.A.S.A., vol. 50, pp 637-655, August 1971.

7èmes JOURNEES D'ETUDE SUR LA PAROLE
NANCY 19 au 21 mai 1976

OPTIMISATION DU CALCUL DES COEFFICIENTS DE CORRELATION PARTIELLE

J. LE ROUX

Laboratoire de Théorie des Systèmes
E.N.S.T.
46 rue Barrault - PARIS Cedex 13 - 75634

RESUME :

Cet article présente un nouvel algorithme de calcul des coefficients de corrélation partielle d'un système linéaire quand on connaît la covariance de sa sortie, l'entrée étant un bruit blanc. Bien que dérivé de l'algorithme de LEVINSON, l'algorithme proposé n'utilise pas les paramètres usuels de l'équation récurrente de la prédiction linéaire. En conséquence, il nécessite moins de calculs et peut être utilisé sur un calculateur en virgule fixe. L'algorithme est appliqué à l'analyse du signal vocal.

SUMMARY :

This paper introduces a new computational algorithm for the partial correlation coefficients of a linear system given the covariance of its output when excited by a white input noise. Although derived from LEVINSON's well known procedure, the proposed algorithm does not make use of the usual parameters in the linear prediction recursion. Consequently, it requires less computations and is implemented using fixed point arithmetics. Application to speech is emphasized.

OPTIMISATION DU CALCUL DES COEFFICIENTS DE CORRELATION PARTIELLE

J. LE ROUX
Laboratoire de Théorie des Systèmes
E.N.S.T.
46 rue Barrault - PARIS Cedex 13 - 75634

I - INTRODUCTION

Les recherches récentes en analyse du signal vocal utilisent le codage par prédiction linéaire (LPC). Cette approche peut être interprétée comme la modélisation des échantillons du signal s_n comme la sortie d'un système linéaire d'ordre p représenté par l'équation récurrente :

$$\sum_{i=0}^{p} a_i^p \, s_{n-i} = u_n^p \qquad (1a)$$

$$a_o^p = 1 \qquad (1b)$$

où u_n^p est l'entrée du système que l'on suppose être un bruit blanc gaussien.

Plusieurs auteurs /1/, /2/, /3/, déduisent les paramètres a_i^p du modèle auto-régressif (ne comportant que des pôles) de l'autocorrélation du signal.

Soient r_k les coefficients de corrélation du signal :

$$\forall k \qquad r_k = E(s_n \, s_{n-k}) \qquad (2a)$$

$$r_k = r_{-k} \qquad (2b)$$

et R_p la matrice de corrélation d'ordre p du signal (matrice de TOEPLITZ) :

$$R_p = \begin{bmatrix} r_o & r_1 & \cdots & r_{p-1} \\ r_1 & r_o & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ r_{p-1} & \cdots & \cdots & r_o \end{bmatrix} \qquad (3)$$

On montre que les coefficients a_i^p sont la solution de l'équation de YULE WALKER :

$$(a_1^p, \ldots, a_p^p)\ R_p + (r_1, \ldots, r_p) = 0 \qquad (4)$$

où $(a_1^p, \ldots, a_p^p)$ et $(r_1, \ldots, r_p)$ sont des vecteurs-lignes de dimension p.

LEVINSON /4/ et ROBINSON /5/ donnent une solution itérative de l'équation (4) qui utilise comme paramètres intermédiaires les coefficients de corrélation partielle (PARCOR). Ces coefficients de corrélation partielle ont des propriétés remarquables ce qui permet leur utilisation en transmission et en synthèse de la parole /6/, /7/, /8/.

Nous présentons ici une nouvelle méthode de calcul des PARCOR. Elle n'utilise pas explicitement les paramètres a_i^p de l'équation récurrente et utilise des calculs intermédiaires qui sont en fait les coefficients d'intercorrélation notés e_m^h entre l'entrée et la sortie des modèles d'ordre successif.

$$e_m^h = E(u_n^h\ s_{n-m}) \qquad (5)$$

En multipliant (1a) par s_{n+k} et en prenant la moyenne, la définition (5) donne :

$$\forall\, m : e_m^p = \sum_{i=0}^{p} a_i^p\ r_{m-i} \qquad (6)$$

On montre que l'utilisation de l'équation (6) réduit le nombre de calculs à effectuer. De plus, toutes les variables intermédiaires sont comprises entre -1 et +1 si la variance du signal est inférieure à 1.

II - UNE AUTRE FORMULATION DE LA SOLUTION ITERATIVE

Utilisant les notations des équations (3), (4), (6), la solution à l'ordre h de l'équation (4) est supposée connue.

$$(r_1, \ldots, r_h) + (a_1^h, \ldots, a_h^h)\ R_h = 0 \qquad (7)$$

Alors le vecteur $(a_1^{h+1}, \ldots, a_{h+1}^{h+1})$ solution de

$$(r_1, \ldots, r_{h+1}) + (a_1^{h+1}, \ldots, a_{h+1}^{h+1})\ R_{h+1} = 0 \qquad (8)$$

se calcule à partir de (6) et (7) par les deux égalités

$$(r_1, \ldots, r_{h+1}) + (a_1^h, \ldots, a_h^h, 0)\ R_{h+1} = (0, \ldots, 0, e_{h+1}^h) \qquad (9)$$

$$(a_h^h, \ldots, a_1^h, 1)\ R_{h+1} = (0, \ldots, 0, e_o^h) \qquad (10)$$

L'équation (8) s'obtient en multipliant (10) par $-\dfrac{e_{h+1}^h}{e_o^h}$ et en l'ajoutant à (9).

Donc la solution d'ordre (h+1) est déduite de la solution d'ordre h par

$$(a_o^{h+1}, a_1^{h+1}, \ldots, a_h^{h+1}, a_{h+1}^{h+1}) = (a_o^h, a_1^h, \ldots, a_h^h, 0) - \frac{e_{h+1}^h}{e_o^h}(0, a_h^h, \ldots, a_1^h, a_o^h) \qquad (11)$$

Les coefficients a_i^h sont les coefficients de la relation récurrente (1). En appliquant (1) aux coefficients de corrélation, utilisant (5) et (11) et définissant k_h par :

$$k_h = -e_{h+1}^h / e_o^h \qquad (12a)$$

Les deux équations suivantes sont obtenues par tout m :

$$e_m^{h+1} = e_m^h + k_h e_{h-m+1}^h \qquad (12b)$$

$$e_{h-m+1}^{h+1} = k_h e_m^h + e_{h-m+1}^h \qquad (12c)$$

En conséquence, les PARCOR $k_o, \ldots, k_{p-1}$ sont calculés par l'algorithme suivant :

i - les valeurs des intercorrélations e_m^o sont initialisées à r_m pour

ii - la valeur de k_{h-1} est calculée par (12a)

iii - les valeurs de $e_{h+1-p}^h, \ldots, e_o^h, e_{h+1}^h, \ldots, e_p^h$ sont calculées par (12b) et (12c)

(ii) et (iii) sont répétées pour h = 1, ..., p - 1.

LLe calcul habituel des PARCOR est fondé sur l'utilisation de l'équation (6). Les k_h sont déduits en fonction des coefficients a_i^p dont on connaît mal les propriétés /2/, /3/. Au contraire, dans cette nouvelle formulation, l'utilisation des intercorrélations précédentes conduit à une interprétation physique des variables utilisées en (12). De plus, cette formultation est très proche de la réalisation du modèle linéaire par la structure de filtre en treillis /7/, /8/, /9/, /10/.

III - CALCUL EN VIRGULE FIXE

Dans de nombreuses applications, le temps de calcul doit être limité et la prédiction linéaire doit être implantée sur un calculateur de petite taille (sans processeur virgule flottante). Pour ces deux raisons un calcul en virgule fixe paraît particulièrement intéressant.

L'inégalité de SCHWARTZ appliquée à (5) donne :

$$|E(u_n^h s_{n-m})|^2 \leq E(u_n^h)^2 E(s_{n-m})^2 \qquad (13)$$

De plus, il a été montré /2/ que :

$$E(u_n^h)^2 = r_o \prod_{j=o}^{h-1} (1 - k_j^2) \qquad (14)$$

En utilisant (2) et (14), (13) conduit à la majoration

$$|e_m^h| \leq r_o \prod_{j=o}^{h-1} (1 - k_j^2)^{1/2} \qquad (15)$$

Les coefficients de corrélation partielle k_j déduits de valeurs exactes de la covariance sont compris entre -1 et +1 /2/ ; en conséquence, (15) entraîne :

$$|e_m^h| \leq r_o \qquad (16)$$

Avec l'hypothèse supplémentaire : $r_o < 1$, la relation (16) montre que tous les calculs intermédiaires sont entre -1 et +1 ce qui permet une implantation en virgule fixe.

(15) donne une majoration plus précise des intercorrélations après la hième itération :

$$\forall h' \geq h, \forall m \qquad |e_m^{h'}| \leq r_o \prod_{j=o}^{h-1} (1 - k_j^2)^{1/2} \qquad (17)$$

comme

$$(1 - k_j^2)^{1/2} \leq (1 - k_j^2 / 2) \qquad (18)$$

la relation (17) implique :

$$\forall h' > h, \forall m \qquad |e_m^{h'}| \leq r_o \prod_{j=o}^{h-1} (1 - k_j^2 / 2) \qquad (19)$$

Cette dernière majoration, plus facile à calculer que la précédente donne le domaine de variation des intercorrélations. Les équations (12b) et (12c) étant homogènes, il est possible de faire un décalage des valeurs en mémoire de façon à sauvegarder la précision des calculs.

IV - RESULTATS

La méthode nécessite p^2-p+1 multiplications et divisions (soit 91 pour un ordre 10) et $(2p+3)$ mots-mémoire. Elle a été implantée, en virgule fixe sur un calculateur 16 bits standard (PACER 100 EAI). Le calcul des PARCOR à l'ordre 10 prend 3.5 ms ce qui permet un calcul en temps réel pour l'analyse de la parole.

La majoration simplifiée augmente peu la qualité des résultats (décalage maximum de deux bits). La relation (16) a toujours été vérifiée et le système trouvé a toujours été stable /2/ pour 3000 calculs des PARCOR à partir de la covariance du signal vocal.

Cette méthode a été comparée à l'algorithme habituellement utilisé /2/, /3/. Comme ce dernier utilise l'équation récurrente habituelle où interviennent les a_i^p, il nécessite leur calcul en tant que valeurs intermédiaires. En conséquence, il nécessite $(p^2+3.5p+1)$ multiplications et divisions (136 pour un ordre 10). La comparaison est résumée dans les tableaux I, II et III où "a" représente les résultats du programme donné par MARKEL et GRAY /2/ et "b" les résultats de l'algorithme présenté ici. Différents paramètres, ayant des sensibilités variables sont montrés : les PARCOR k_i (I), les coefficients de l'équation récurrente a_i^{10} (II) et les pôles du modèle (III).

La figure 1 montre le spectre du signal, la figure 2, les spectres déduits du système linéaire /2/, /3/ et la figure 3 les racines du polynôme en z calculé par les deux méthodes.

Pour plus de 100 expériences, la différence entre les deux résultats est inférieur à .005 pour k_{10} et ne donne pas de différence sensible dans le spectre. L'écart en fréquence ne dépasse pas 1 Hz pour les pôles les plus significatifs, la fréquence d'échantillonnage étant de 10 KHz.

V - CONCLUSION

Les coefficients de corrélation partielle se sont montrés d'un grand intérêt dans l'analyse, la synthèse et la transmission de la parole et aussi comme intermédiaires pour le calcul de paramètres qui leur sont liés /6/, /7/, /8/, /11/.

Cette approche a été facilement étendue à d'autres types de signal et a été appliquée avec succès /3/, /10/. Le gain de temps apporté par cet algorithme sera apprécié surtout lorsque les calculs doivent être répétés : de plus, la possibilité d'effectuer les calculs en virgule fixe peut contribuer au développement dans de nombreuses applications de processeurs spécialisés utilisant l'approche de la prédiction linéaire en temps réel.

REFERENCES

/1/ F. ITAKURA and S. SAITO : "A statistical method for estimation of speech spectral density and formant frequencies"
Electronics and Communications in Japan, Vol 53 a, n° 11970

/2/ J. D. MARKEL and A. H. GRAY : "On autocorrelation equations as applied to speech analysis"
IEEE Trans. on Audio. and electroacoustics, Vol AU-21, n° 2, April 1973

/3/ J. MAKHOUL : "Linear prediction, a tutorial review"
Proceedings of the IEEE, Vol 63, n° 4, April 1975

/4/ N. LEVINSON : "The Wiener RMS error criterion in filter design and prediction"
J. Math. Phys., Vol 25, n° 4, p 261, 1947

/5/ E. A. ROBINSON : "Statistical communication and detection with special reference to digital data processing of radar and seismic signals"
Hafner, New York, 1967, p 274

/6/ J. MAKHOUL, R. VISWANATHAN, L. COSELL, W. RUSSEL : "Natural communication with computers : speech compression research at BBN"
Final Report, Vol II, BBN Report n° 2976 - Bolt, Beranek and Newman Inc., Cambridge, Massachussets, Dec. 1974

/7/ A. H. GRAY and J. D. MARKEL : "Digital lattice and ladder filter synthesis"
IEEE Trans. on Audio.and electroacoustics, Vol AU-21, n° 6, Dec. 1973

/8/ F. ITAKURA and S. SAITO : "Digital filtering techniques for speech analysis and synthesis"
Seventh International Congress on Acoustics, 25 C-1, Budapest, 1971

/9/ M. D. SRINATH and M. M. VISWANATHAN : "Sequential algorithm for identification of parameters of an autoregressive process"
IEEE Trans. on Automatic Control, Vol AC-20, n° 4, August 1975

/10/ M. MORF, G. S. SIDHU, T. KAILATH : "Some new algorithms for recursive estimation in constant linear discrete time series"
IEEE Trans. on Automatic Control, Vol AC-19, n° 4, August 1974

/11/ C. GUEGUEN, J. LE ROUX, J. C. DOMENGER, C. BENCHIMOL, J. F. BELLEC : "Un synthétiseur à structure programmable"
6e Journées d'Etude sur la Parole, Toulouse (France), Mai 1975 (In French)

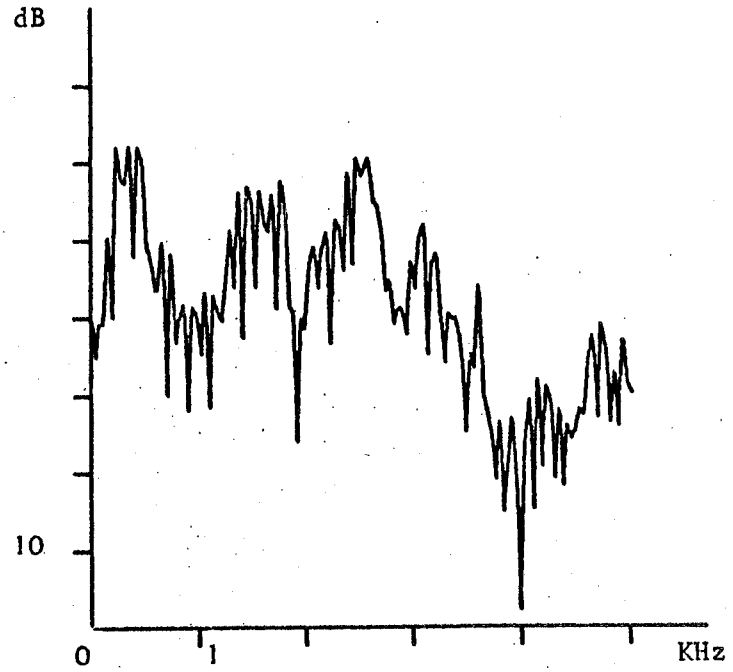
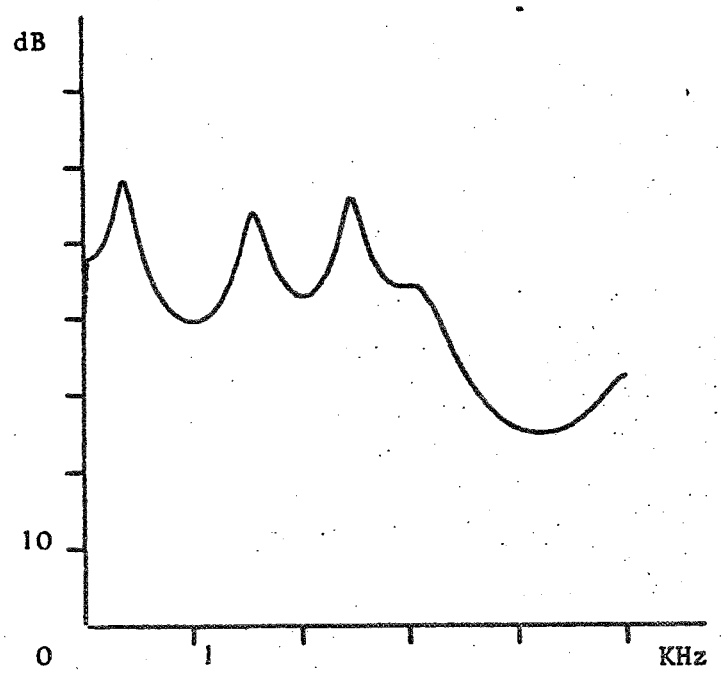
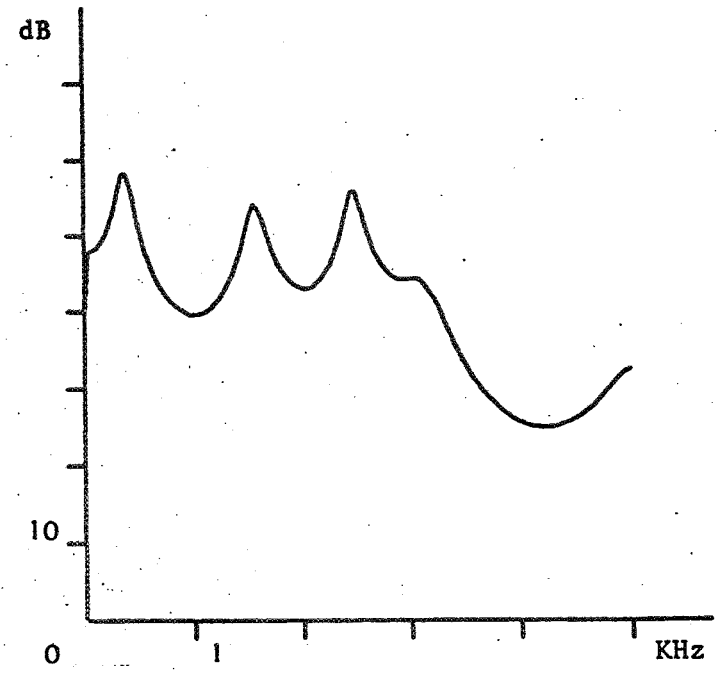


Figure 1 - SPECTRE ORIGINAL DU SIGNAL VOCAL

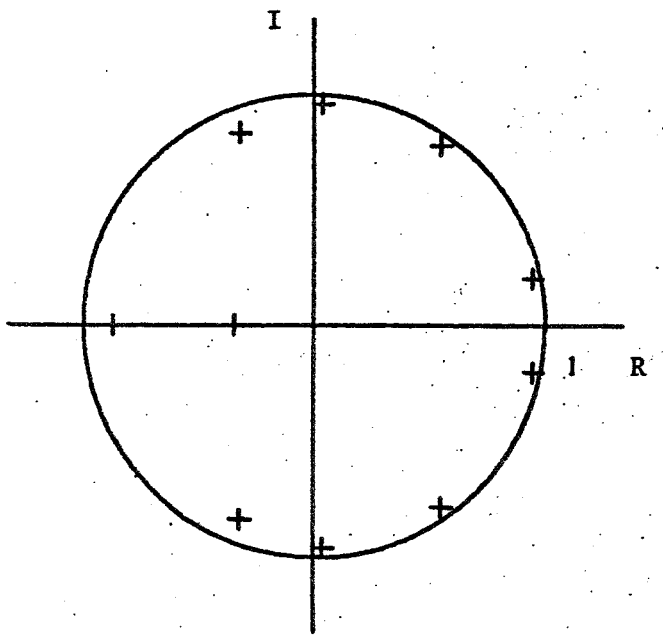


a - Algorithme de DURBIN

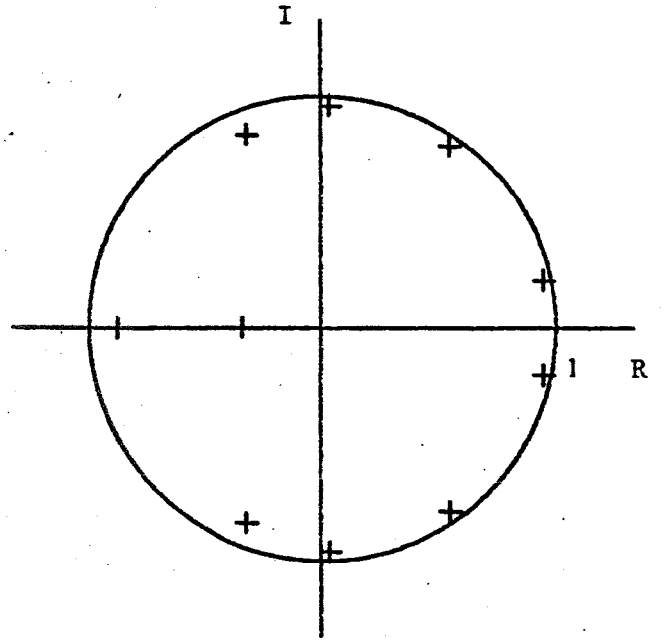


b - Algorithme utilisant l'intercorrélation

Figure 2 - Spectres calculés à partir des coefficients obtenus par les deux méthodes



a - Algorithme de DURBIN



b - Algorithme utilisant l'intercorrélation

Figure 3 - Pôles calculés à partir des coefficients obtenus par les deux méthodes

a	b
-.561739	-.561737
.422128	.422272
-.670678	-.670898
.299176	.300262
-.184051	-.185181
-.142878	-.141998
.153059	.153748
.027713	.027191
.534487	.536377
.184168	.182190

a - Méthode de DURBIN

b - Méthode utilisant l'intercorrélation

Table I - Coefficients de corrélation partielle calculés à partir de la covariance par les deux méthodes

1.000000
-1.215774
1.278934
-1.002074
.121015
.331589
- .785682
.620881
- .423939
.292451
.184168

a - Méthode de DURBIN

1.000000
-1.218343
1.283374
-1.006944
.124316
.332556
- .790242
.627815
- .429932
.296603
.182190

b - Méthode utilisant l'intercorrélation

Table II - Coefficients de l'équation récurrente déduits des coefficients de corrélation partielle

Réel	Imaginaire
-.340385	.000000
.945065	±.201624
.037303	±.958678
-.866220	.000000
.547068	±.783280
-.318246	±.833570

a - Méthode de DURBIN

Réel	Imaginaire
-.335610	.000000
.945086	±.201771
.037247	±.958785
-.867771	.000000
.547008	±.783390
-.318480	±.834006

b - Méthode utilisant l'inter-corrélation

Table III - Pôles du polynôme en z déduits des coefficients de l'équation récurrente par les deux méthodes

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 au 21 mai 1976

ANALYSE DU SIGNAL VOCAL EN VUE DU CODAGE DE LA PAROLE A FAIBLE DEBIT D'INFORMATION

J. MENEZ — Laboratoire 190 du CNRS, Université de Nice
D. ESTEBAN — CER IBM FRANCE, La Gaude
F. DUBUS — Laboratoire 190 du CNRS, Université de Nice

RESUME :

Après un rappel des méthodes fondamentales utilisant plus particulièrement des techniques de codage par prédiction linéaire, nous exposons une méthode d'analyse-synthèse à structure transversale utilisant à la fois les propriétés du codage par blocs et les techniques de prédiction linéaire. Il ressort de cette étude que la structure transversale est bien appropriée au codage de la parole à faible débit d'information.

Des enregistrements à différents débits d'information (12/16/20 Kbps) seront présentés.

SUMMARY :

After a review of linear predictive coding and adaptive predictive coding techniques, it is described an analysis-synthesis method with transversal structure derived from block companded coding and linear prediction techniques. From this study it is shown that the transversal structure is well-suited for low bit rate speech encoding in the range of 10 Kbps to 20 Kbps.

Tape recordings at different bit rates (12/26/20 Kbps) will be played.

ANALYSE DU SIGNAL VOCAL EN VUE DU CODAGE DE LA PAROLE A FAIBLE TAUX D'INFORMATION

J. MENEZ, D. ESTEBAN, F. DUBUS

I. - INTRODUCTION

Pour coder numériquement et efficacement la parole, il existe deux types de systèmes de compression :

- d'une part les systèmes qui cherchent à reproduire l'onde temporelle du signal originel ; ces systèmes ont tous la même structure bouclée de la Modulation en Impulsions Codées Différentielles (DPCM) : parmi les systèmes appartenant à cette catégorie, on peut citer le codeur prédictif évolutif (APC) proposé par ATAL et SCHROEDER |1,2| et qui fut le premier à tenir compte de la nature très redondante du signal vocal ; sa mise en oeuvre nécessite la détermination périodique de paramètres qu'il faut transmettre, en plus du signal binaire issu du quantificateur, pour reconstituer le signal au récepteur : le débit d'information requis par ce procédé de codage est d'environ 10 kbps pour une bande de base de 3,3 kHz. Ultérieurement, STROH |3| propose un codeur ADPCM à prédicteur évolutif dans lequel il n'est plus nécessaire de transmettre les paramètres pour reconstituer le signal au récepteur. Le codeur résiduel proposé par MELSA et al |4,5| est une version plus élaborée du procédé de codage précédent : ce codeur requiert des débits moyens d'information de 9,6 kbps et 16 kbps.

- d'autre part, les systèmes qui permettent de générer un signal ayant des propriétés spectrales "proches" de celles du signal originel, tel le vocodeur à prédiction linéaire (LPC) |6,7| qui permet d'obtenir une parole de synthèse parfaitement intelligible pour des débits d'information variant entre 1,2 kbps et 4,8 kbps : il faut toutefois remarquer que la qualité de la parole obtenue par ce procédé varie considérablement selon le locuteur.

Le procédé de codage que nous proposons possède une structure qui s'apparente à celle du vocodeur à prédiction linéaire ; elle en diffère cependant par le fait que l'on code l'onde temporelle du signal d'excitation : cette structure semble particulièrement bien adaptée au codage de la parole pour des débits d'information compris dans l'intervalle 10 kbps à 20 kbps (pour une bande de base de 4kHz) : les résultats obtenus par simulation tendent à confirmer cette hypothèse.

II. - PREDICTION LINEAIRE DU SIGNAL VOCAL : APPLICATIONS AU CODAGE NUMERIQUE DE LA PAROLE

Les méthodes de prédiction linéaire sont basées sur l'hypothèse qu'un échantillon S_n du signal vocal, correspondant au $n^{ième}$ instant d'échantillonnage, peut être prédit à partir d'une combinaison linéaire des p échantillons précédents selon l'expression

$$\hat{S}_n = \sum_{i=1}^{p} a_i S_{n-i} \qquad (1)$$

où les a_i sont des coefficients de pondération.

Si on désigne par e_n l'écart entre la valeur réelle de l'échantillon S_n et sa valeur prédite, on obtient la relation suivante

$$e_n = S_n - \hat{S}_n = S_n - \sum_{i=1}^{p} a_i S_{n-i} \qquad (2)$$

e_n représente l'erreur de prédiction.

Le problème consiste à déterminer un ensemble de coefficients a_i ,$i \leqslant 1 \leqslant p$ qui minimisent au sens des moindres carrés l'erreur quadratique moyenne de prédiction. Plusieurs méthodes de prédiction linéaire permettent de résoudre ce problème. Ces méthodes sont pour l'essentiel :

- les méthodes nécessitant un traitement par bloc d'échantillons et dont il existe deux formulations différentes connues respectivement sous le nom de méthode de covariance |6| et méthode d'autocorrélation |8,9|. Ces méthodes sont dites du type global.

- les méthodes d'estimation séquentielle pour lesquelles le traitement s'effectue échantillon par échantillon : on peut citer la méthode du filtre de KALMAN |10,11| et la méthode d'approximation stochastique |10|.

En appliquant la transformée en Z à l'équation (2) on obtient

$$E(Z) = S(Z)\left[1 - \sum_{i=1}^{p} a_i Z^{-i}\right] = S(Z).\, G(Z) \qquad (3)$$

et donc

$$\frac{S(Z)}{e(Z)} = \frac{1}{G(Z)} = H(Z) \qquad (4)$$

Cette équation montre que le signal S_n représente le signal de sortie d'un filtre linéaire dont le signal d'entrée serait le signal e_n. Ce filtre a comme fonction de transfert H(Z) et ne possède donc que des pôles. Ce filtre est stable si tous ses pôles sont situés à l'intérieur du cercle unité. Selon les travaux de FANT |12| et de FLANAGAN |13| la fonction H(Z) représente la fonction de transfert du modèle du conduit vocal pour la production des sons sans nasalité.

Les coefficients a_i qui peuvent être déterminés par les différentes méthodes de prédiction linéaire caractérisent le modèle du conduit vocal.

La comparaison de ces différentes méthodes n'a de signification réelle que par rapport à l'usage qu'on désire en faire ; toutefois, dans le cadre de l'extraction en temps réel des paramètres caractéristiques du conduit vocal les contraintes suivantes s'imposent :

- simplicité de l'algorithme mis en oeuvre sur le plan du nombre des opérations et de l'encombrement mémoire,

- convergence de l'algorithme vers un minimum pour un intervalle de temps donné dans le cas des techniques d'estimation séquentielle,

- stabilité du modèle obtenu.

L'équation (3) peut s'écrire encore

$$\frac{E(Z)}{S(Z)} = 1 - \sum_{i=1}^{p} a_i Z^{-i} = G(Z) = \frac{1}{H(Z)} \qquad (5)$$

Sous cette forme le signal e_n représente la sortie d'un filtre G(Z) dont le signal d'entrée serait S_n ; or la fonction G(Z) est l'inverse de la fonction de transfert H(Z) du modèle du conduit vocal. Une fois les coefficients a_i déterminés, le signal e_n représentant l'erreur de prédiction peut être obtenu par filtrage inverse: cette opération correspond à la déconvolution du signal S_n par le filtre H(Z). C'est pourquoi le signal d'erreur e_n est encore appelé signal d'excitation.

II-1 Analyse-synthèse par prédiction linéaire

L'analyse a pour but l'extraction de paramètres fondamentaux qui caractérisent le processus de la parole et au moyen desquels il est possible de décrire l'évolution du signal vocal dans le domaine temporel. Ces paramètres sont pour l'essentiel :

- d'une part, ceux qui caractérisent le conduit vocal,à savoir sur le plan de l'articulation les aires de section de ce conduit et sur le plan de l'acoustique les formants et par là même les coefficients de la fonction de transfert du conduit vocal ; on peut au demeurant passer des coefficients de la fonction de transfert aux sections d'aire et inversement par l'utilisation des coefficients de réflection |14|. L'extraction de ces paramètres est réalisée par l'utilisation des techniques de prédiction linéaire.

- d'autre part, ceux qui caractérisent le signal d'excitation, à savoir la mélodie, le voisement et enfin un facteur de gain qui permet dans le cadre de la synthèse d'obtenir un signal ayant même énergie que le signal originel.

Les domaines d'application de l'analyse sont essentiellement :

- d'une part la synthèse de la parole qui nécessite des débits d'information $\leq$ 4,8 kbps,

- d'autre part la reconnaissance de la parole ou l'identification du locuteur.

La figure 1 représente les éléments fonctionnels d'un vocodeur utilisant des techniques d'analyse-synthèse par prédiction linéaire.

Tous les travaux de recherche sur la synthèse de la parole semblent montrer que c'est l'information contenue dans le signal d'excitation qui confère à la parole son naturel. Or aucune des méthodes d'extraction de la mélodie proposées à ce jour ne peut garantir une détection parfaite de la mélodie ; la paramétrisation du

signal d'excitation est toujours entachée d'erreurs, soit au niveau de la décision de voisement, soit au niveau de la période du fondamental. La paramétrisation trop stricte et parfois erronée du signal d'excitation entraîne une perte de naturel dans la parole de synthèse. Par ailleurs, la qualité obtenue par ce procédé de codage dépend énormément du locuteur, limitant ainsi les applications de ce système de compression de la parole dans le domaine de la transmission de la parole.

II-2 Codage prédictif et évolutif de l'onde temporelle du signal vocal

Le signal vocal est de nature très redondante : les deux causes principales de redondance sont, d'une part la présence de formants dans le spectre du signal vocal, d'autre part la quasi périodicité de l'onde temporelle durant la production des sons voisés. Pour réduire l'information redondante contenue dans le signal vocal ATAL et SCHROEDER [1,2] proposent d'utiliser deux prédicteurs en cascade (figure 2). Le premier prédicteur permet de réduire la redondance. La détermination des coefficients de ce prédicteur est basée sur l'hypothèse que dans le cas de la production d'un son voisé un échantillon du signal S_n peut être prédit à partir d'un échantillon S_{n-M} déjà passé. Le retard M est lié à la périodicité du signal.

On peut donc écrire

$$\tilde{S}_n = \beta S_{n-M} \qquad (6)$$

où β représente un facteur de gain.

Soit V_n l'erreur de prédiction ; V_n est donnée

$$V_n = S_n - \beta S_{n-M} \qquad (7)$$

La détermination de M s'obtient par localisation du maximum de la fonction d'autocorrélation normalisée

$$\rho(M) = \frac{\sum_{n=1}^{N} S_n . S_{n-M}}{\left(\sum_{n=1}^{N} S_n^2 . \sum_{n=1}^{N} S_{n-M}^2 \right)} \qquad (8)$$

β est ensuite obtenu en minimisant l'erreur quadratique moyenne de V_n

$$\beta = \frac{\sum_{n=1}^{N} S_n . S_{n-M}}{\sum_{n=1}^{N} S_{n-M}^2} \qquad (9)$$

Le second prédicteur permet de réduire la redondance spectrale contenue dans le signal V_n issu du premier prédicteur.

Les coefficients du second prédicteur sont déterminés à l'aide de la méthode de covariance, en se basant sur l'hypothèse que l'échantillon V_n peut être prédit à partir d'une combinaison linéaire des p échantillons précédents :

$$\tilde{V}_n = \sum_{i=1}^{p} \alpha_i V_{n-i} \qquad (10)$$

Pour tenir compte de l'évolution du signal vocal, les paramètres des deux prédicteurs doivent être déterminés périodiquement à partir du signal S_n.

La structure du codeur prédictif évolutif (APC) proposé par ATAL et SCHROEDER est similaire à celle du codeur prédictif DPCM représenté par la figure 3. Dans ce type de codeur le signal résiduel e_n obtenu par différence entre la valeur réelle de l'échantillon S_n et sa valeur prédite $\tilde{S}_n$ est codé par un quantificateur. L'utilisation d'un prédicteur très élaboré dans le codeur APC permet de coder le signal résiduel avec un quantificateur à deux niveaux dont le pas est déterminé périodiquement de façon à minimiser la puissance moyenne du bruit de quantification. La reconstruction du signal s'effectue à l'aide de la séquence binaire issue du quantificateur, et des paramètres suivants : pas de quantification, coefficients β, M et α_i, $i \leqslant 1 \leqslant p$ des deux prédicteurs ; ces paramètres doivent donc être transmis au récepteur.

Une des particularités de ce type de codeur est de générer, en l'absence d'erreur de transmission, un bruit égal à celui créé par le quantificateur. Si l'on désigne (voir figure 3) par $\hat{S}_n$ et $\hat{S}'_n$ les échantillons du signal reconstruit respectivement à l'émetteur et au récepteur et si l'on tient compte de la restriction précédente, on peut écrire

$$\hat{S}_n - S_n = \hat{S}'_n - S_n \qquad (11)$$

et

$$e_n = S_n - \tilde{S}_n \qquad (12)$$

Si l'on désigne par $\hat{e}_n$ la valeur obtenue après quantification de l'échantillon e_n on peut encore écrire

$$\hat{e}_n = e_n + q_n \qquad (13)$$

où q_n représente le bruit de quantification : or

$$\hat{S}_n = \tilde{S}_n + \hat{e}_n \qquad (14)$$

et donc

$$\hat{S}_n - S_n = \tilde{S}_n + e_n + q_n - S_n = q_n \qquad (15)$$

Pour évaluer les performances de ce type de codeur, on utilise comme mesure quantitative de la qualité le rapport signal sur

bruit de quantification défini par

$$S/Bq = \frac{\langle s_n^2 \rangle}{\langle q_n^2 \rangle} \qquad (16)$$

où $\langle s_n^2 \rangle$ et $\langle q_n^2 \rangle$ représentent respectivement la puissance moyenne du signal originel et la puissance moyenne du bruit de quantification.

Il est à noter cependant que si le rapport signal sur bruit est une mesure acceptable de la qualité dans le cas de codeurs prédictifs utilisant des quantificateurs possédant un grand nombre de niveaux, il n'en est plus de même pour les codeurs dont les quantificateurs ont un nombre restreint de niveau ; ceci peut s'expliquer par le fait que le bruit de quantification est d'autant plus correlé au signal d'entrée du quantificateur que le nombre de niveaux de ce dernier diminue |15,16|. Dans ce cas, pour juger de la qualité réelle du système de codage, il faut utiliser des tests subjectifs. Néanmoins, dans ce type de codeur, le bruit de quantification demeure de nature granulaire et devient parfaitement audible voire même désagréable pour des débits d'information faibles (de 10 à 16 kbps). Un autre inconvénient du codeur APC réside dans le fait qu'il n'existe pas de détermination directe du pas optimal du quantificateur permettant de minimiser la puissance moyenne du bruit de quantification.

III. - VOCODEUR A EXCITATION RESIDUELLE

Le système de compression que nous proposons utilise à la fois les propriétés du vocodeur à prédiction linéaire et celles du codage prédictif ; l'étude de ce procédé de codage a été entreprise dans le but d'obtenir une parole de bonne qualité pour des débits d'information compris dans l'intervalle 12 kbps-20 kbps (figure 4).

III-1 Analyse : détermination des paramètres du système

La structure du système de compression proposé s'apparente à celle d'un vocodeur à prédiction linéaire dont il en diffère par le fait que l'on code à l'aide de techniques de codage prédictif l'onde temporelle du signal d'excitation obtenue par filtrage inverse. Le traitement s'effectue par bloc d'échantillons de longueur N.

III-1-a) Réduction de la redondance spectrale par prédiction linéaire

La méthode de prédiction linéaire utilisée est la méthode d'autocorrélation |8,9|. Cette méthode n'est pas directement appliquée sur les échantillons S_n du signal originel, mais sur les échantillons X_n obtenus par la relation récurrente

$$X_n = S_n - \alpha S_{n-1} \qquad (17)$$

La détermination du coefficient α est obtenue par minimisation de la valeur quadratique moyenne du signal X_n

$$\alpha = \frac{R_1}{R_0} \qquad (18)$$

où R_0 et R_1 représentent les deux premiers points de la fonction d'autocorrélation du signal S_n.

En appliquant la transformée en Z à l'équation (17) on trouve

$$X(Z) = (1 - \alpha Z^{-1}) \quad S(Z) = S(Z).D(Z) \qquad (19)$$

La fonction D(Z) représente la fonction de transfert d'un filtre de pré-emphase. Le signal X_n ainsi obtenu conserve dans son spectre de puissance les mêmes résonances formantiques que celles contenues dans le spectre de puissance du signal S_n. Cependant, contrairement à ce que l'on peut observer pour le spectre de puissance du signal S_n, l'enveloppe des résonances formantiques du signal X_n est plate.

La méthode d'autocorrélation appliquée au signal X_n permet d'obtenir les coefficients a_i, $i \leqslant 1 \leqslant p$ de la fonction de transfert d'un modèle du conduit vocal. Le prétraitement du signal S_n a pour but de favoriser l'extraction des résonances formantiques supérieures |7|.

Soit H(Z) la fonction de transfert du filtre modèle du conduit vocal ainsi obtenu. Par filtrage inverse ou déconvolution du signal S_n on obtient un signal d'excitation e_n donné par l'équation :

$$e_n = S_n - \sum_{i=1}^{p} a_i S_{n-i} \quad \text{identique à l'équation (2)}$$

Le signal d'excitation e_n résultant est codé par des techniques de codage prédictif.

Les coefficients a_i sont déterminés toutes les 16ms, ce qui correspond à un bloc de 128 échantillons pour une fréquence d'échantillonnage de 8 kHz.

III-1-b) Réduction de la redondance temporelle par prédiction à long terme du signal d'excitation e_n

La réduction de la redondance à long terme est basée sur l'hypothèse qu'un échantillon e_n du signal d'excitation peut être prédit à partir d'un échantillon passé e_{n-M} où M est un retard lié à la périodicité du signal. La valeur de M est déterminée par une méthode d'autocorrélation non linéaire |17|. Dans la mesure où l'on impose à M d'être supérieure à la longueur L du bloc d'échantillons e_n à traiter, on peut écrire

$$e_n = \beta \hat{e}_{n-M} + \delta_n \qquad (20)$$

où $\hat{e}_{n-M}$ représente un échantillon du signal d'excitation reconstitué, β un facteur de gain et δ_n un échantillon du signal d'erreur de pré-

diction. β est calculé par minimisation de l'erreur quadratique moyenne du signal δ_n.

On trouve

$$\beta = \frac{\sum_{n=1}^{L} e_n . \hat{e}_{n-M}}{\sum_{n=1}^{L} (\hat{e}^2_{n-M})^2} \qquad (21)$$

Cela revient à chercher la plus grande similitude entre le bloc d'échantillons e_n à traiter et un bloc d'échantillons $\hat{e}_{n-M}$ reconstitués. Une fois M et β déterminés, on obtient, par la relation (20), une séquence d'échantillons δ_n représentant le signal d'erreur de prédiction que l'on va coder à l'aide d'un quantificateur à deux ou quatre niveaux.

III-1-c) Codage optimal du signal d'erreur δ_n

Le fait que M soit supérieure à L entraîne la connaissance a priori du signal δ_n. On cherche une représentation temporelle du signal δ_n de la forme

$$C + Q\ F_n \qquad (22)$$

c et q sont des paramètres qui caractérisent le quantificateur et F_n une séquence binaire. Le problème consiste à déterminer d'une part les paramètres C et Q, d'autre part la séquence binaire F_n de façon à minimiser l'erreur quadratique **totale** donnée par la relation

$$E = \sum_{n=1}^{L} (\delta_n - C - Q\ F_n)^2 \qquad (23)$$

La minimisation de E n'est pas directe. On utilise une méthode de substitution : pour cela on commence par s'imposer une séquence F_n : par exemple dans le cas d'un codage par modulation MIC à 2 éléments binaires on choisit comme séquence initiale F°_n la séquence donnée par $F^\circ_n = \text{signe}(\delta_n)$. Une fois la séquence F°_n fixée, on peut minimiser E par rapport à C et Q. Soient C_o et Q_o les valeurs de C et Q trouvées. On remplace C et Q par ces valeurs dans l'équation (23); on cherche une nouvelle séquence F^1_n qui réduit E.

Le système converge au bout de six à huit itérations. Ces travaux ont fait l'objet d'une présentation récente (18).

La détermination des paramètres M, β, C et Q de même que celle de la séquence F_n est effectuée toutes les 8ms, ce qui correspond à un bloc de 64 échantillons pour une fréquence d'échantillonnage de 8 kHz.

III-2 Synthèse de la parole

La figure 5 représente le schéma du synthétiseur. Le signal d'excitation est d'abord reconstruit à partir des paramètres β, M, C et Q et de la séquence F_n, puis reconvolé par le filtre modèle du conduit vocal. Il faut remarquer que la structure bouclée du système de codage utilisé pour coder le signal d'excitation e_n fait que le bruit généré par ce système est égal au bruit de quantification. Le signal de synthèse peut donc être considéré comme la superposition du signal originel et d'un signal obtenu en convoluant le bruit de quantification par la réponse impulsionnelle du filtre modèle du conduit vocal.

La transmission des paramètres a_i, β, M, C et Q, requiert un débit d'information d'environ 4 Kbps. La transmission de la séquence F_n peut se faire à 8, 12 et 16 Kbps.

IV. - CONCLUSION

Les techniques de codage prédictif appliquées au signal vocal permettent de coder efficacement la parole pour des débits d'information compris dans l'intervalle 10 Kbps - 20 Kbps. Toutefois, la qualité de parole obtenue avec des systèmes de compression qui utilisent ces techniques est altérée par un bruit granulaire de quantification plus ou moins audible.

Avec le procédé de codage que nous proposons la qualité semble meilleure du fait que la distorsion de signal engendrée par ce système de compression correspond au filtrage du bruit de quantification par le filtre modèle du conduit vocal.

La densité spectrale de puissance de cette distorsion a donc la même forme d'enveloppe que l'enveloppe formantique : cette propriété semble rendre moins audible cette distorsion.

1 B.S. ATAL and M.R. SCHROEDER "PREDICTIVE CODING OF SPEECH SIGNALS : Proceedings of the Conference on Speech Communication and Processing Cambridge, Mass.6-8 nov. 1967

2 B.S. ATAL and M.R. SCHROEDER : "ADAPTIVE PREDICTIVE CODING OF SPEECH SIGNALS" : B.S.T.J. Oct. 1970 Vol.49 n° 8 pp 360-361

3 R.W. STROH "OPTIMUM AND ADAPTIVE DIFFERENTIAL PCM" Ph.D dissertation, Polytech.Inst.Brooklyn,Farmingdale N.Y. 1970

4 J.D. GIBSON, S.K.JONES and J.L.MELSA "SEQUENTIALLY ADAPTIVE PREDICTION AND CODING OF SPEECH SIGNALS", IEEE Trans. Commun. vol COM-22 pp 1789-1797 Nov. 1974

5 J.D. GIBSON, S.K.JONES and J.L.MELSA "SEQUENTIALLY ADAPTIVE PREDICTIVE CODING OF SPEECH" Proc.of the 1974 National Elec. Conf. Oct. 16, 1974, Chicago,111

6 B.S. ATAL and S.L. HANAUER : "SPEECH ANALYSIS AND SYNTHESIS BY LINEAR PREDICTION OF THE SPEECH WAVE" J.A.S.A. August 1971 vol. 50 n° 2 part 2 pp 637-655

7 J.D. MARKEL , A.H. GRAY "A LINEAR PREDICTION VOCODER SIMULATION BASED UPON THE AUTOCORRELATION METHOD" IEEE Trans. ASSP-22 n° 2 p. 124 (april 1974)

8 J.D. MARKEL, A.H. GRAY "ON AUTOCORRELATION EQUATIONS AS APPLIED TO SPEECH ANALYSIS" IEEE Trans. AU 21 n° 2 pp 69-79 (april 1973)

9 F. ITAKURA, S. SAITO "A STATISTICAL METHOD FOR ESTIMATION OF SPEECH SPECTRAL DENSITY AND FORMANT FREQUENCIES" Elec. Comm. in Japan, 53.A n° 1, pp 36-43 (1970)

10 J.D. GIBSON, J.L. MELSA, S.K. JONES "DIGITAL SPEECH ANALYSIS USING SEQUENTIAL ESTIMATION TECHNIQUES" IEEE Trans. on ASSP Vol ASSP-23 n° 4 August 1975

11 C. GUEGUEN, G. CARAYANNIS "ANALYSE DE LA PAROLE PAR FILTRAGE OPTIMAL DE KALMAN" Automatisme - Tome XVIII n° 3, p99 (mars 1973)

12 G. FANT "ACOUSTIC THEORY OF SPEECH PRODUCTION" Mouton and Co The Hague (1960)

13 J.L. FLANAGAN "SPEECH ANALYSIS, SYNTHESIS AND PERCEPTION" Springer Verlag - Berlin (1965)

14 F. ITAKURA, S. SAITO "ON THE OPTIMUM QUANTIZATION OF FEATURE PARAMETERS IN THE PARCOR SPEECH SYNTHESIZER" Conf. on Speech Commun. and Proc., Boston pp 434-437 (april 1972)

15 A.V. OPPENHEIM and R.W. SCHAFER "DIGITAL SIGNAL PROCESSING" Prentice-Hall, chap. 11, pp. 562-570

16 S.N. JAYANT "DIGITAL CODING OF SPEECH WAVEFORMS : PCM, DPCM AND DM QUANTIZERS" Proceedings of the IEEE, Vol. 62, N°5, pp. 611-632, May 1974

17 C. GALAND "NON-LINEAR AUTOCORRELATION FUNCTION (NLAF) - REAL TIME PITCH EXTRACTOR" 91st Spring Meeting of ASA, April 1976, Washington

18 J. MENEZ, D. ESTEBAN and J.P. TEMIME "OPTIMUM ONE-BIT BLOCK-QUANTIZING" 91st Spring Meeting of ASA, April 1976, Washington

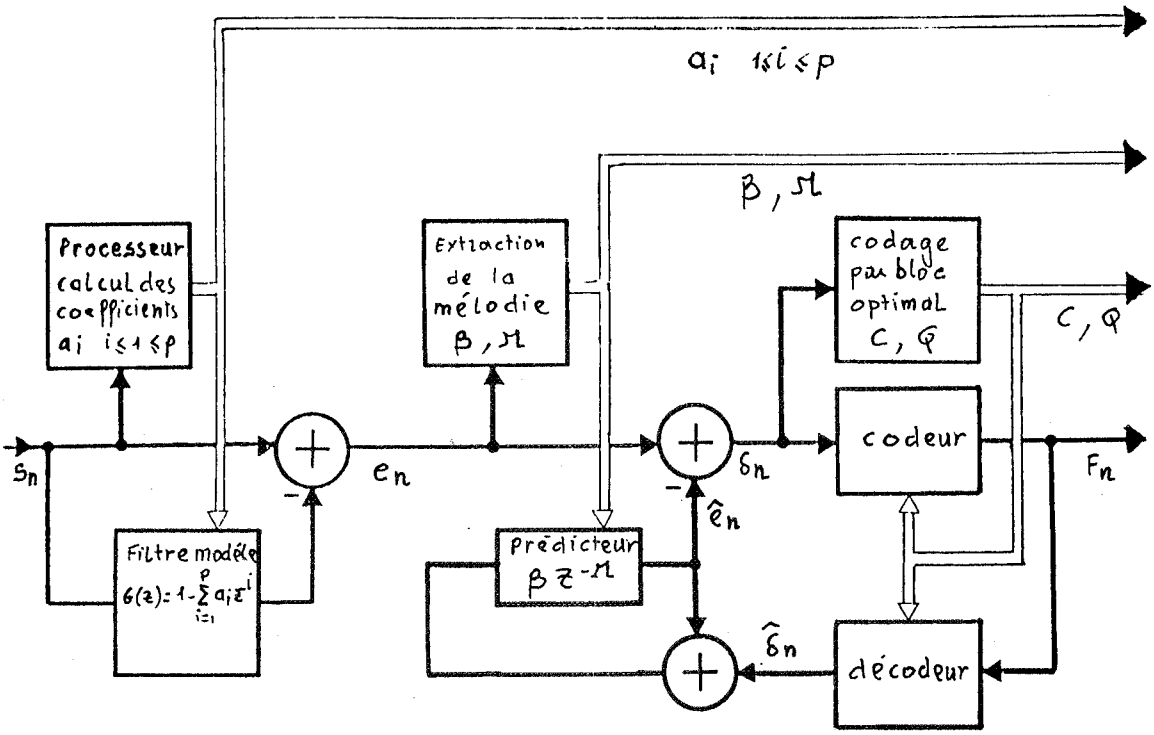


FIGURE 4 : ANALYSEUR DU CODEUR A STRUCTURE TRANSVERSALE

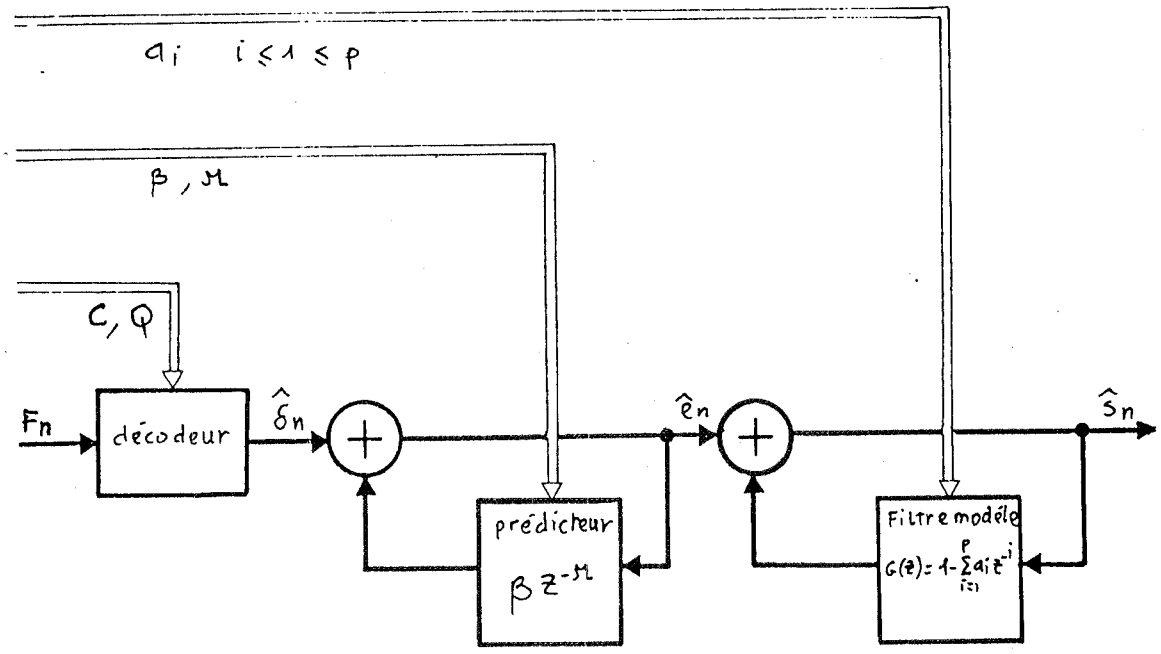
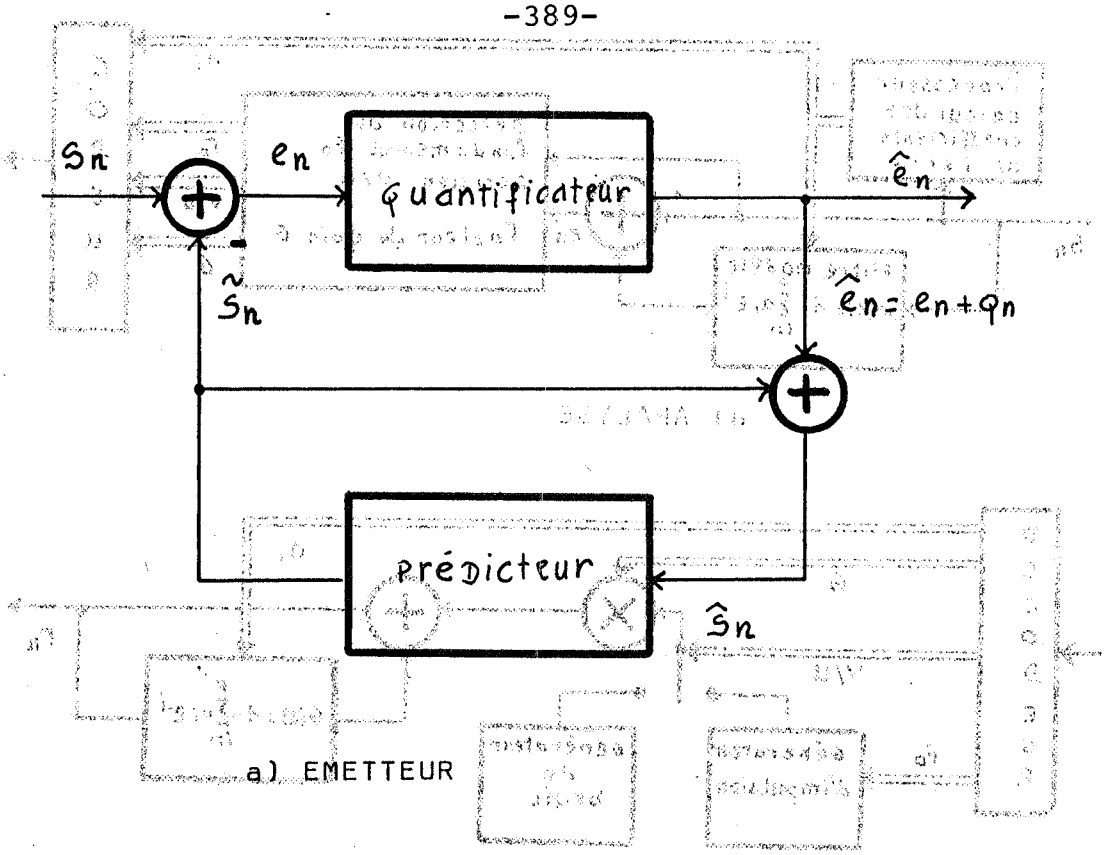
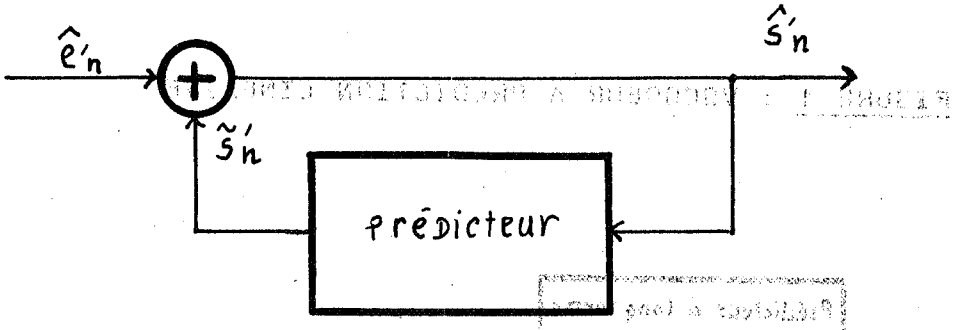


FIGURE 5 : SYNTHESISEUR DU CODEUR A STRUCTURE TRANSVERSALE

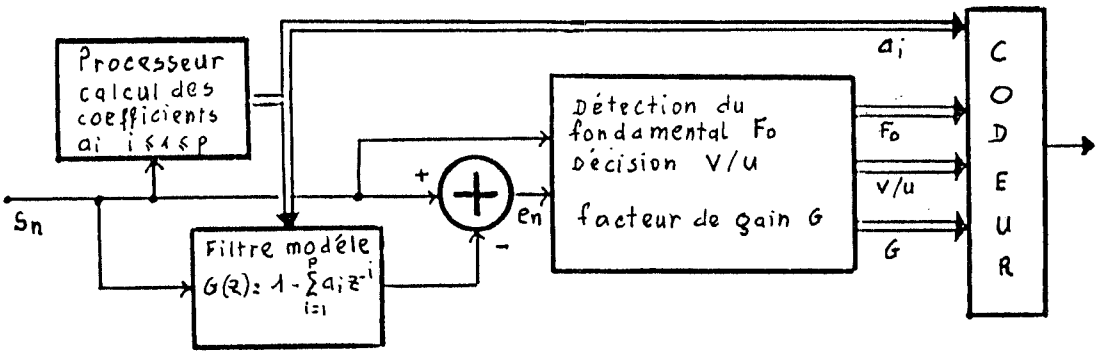


a) EMETTEUR

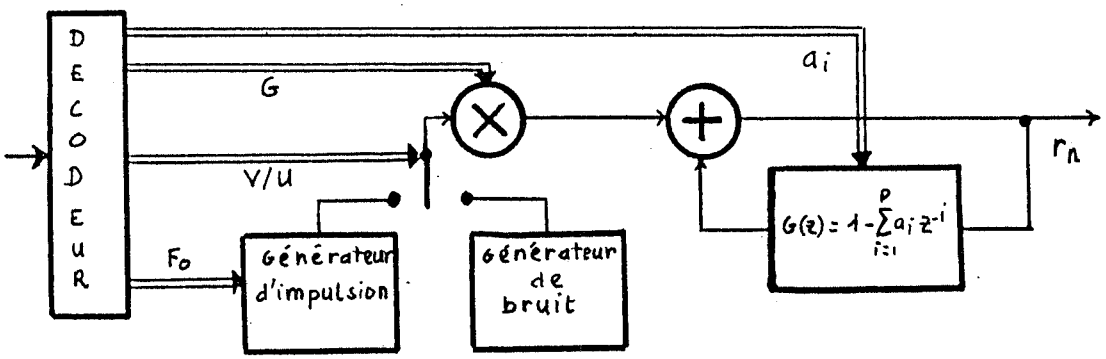


b) RECEPTEUR

FIGURE 3 : CODEUR PREDICTIF DPCM



a) ANALYSE



b) SYNTHESE

FIGURE 1 : VOCODEUR A PREDICTION LINEAIRE

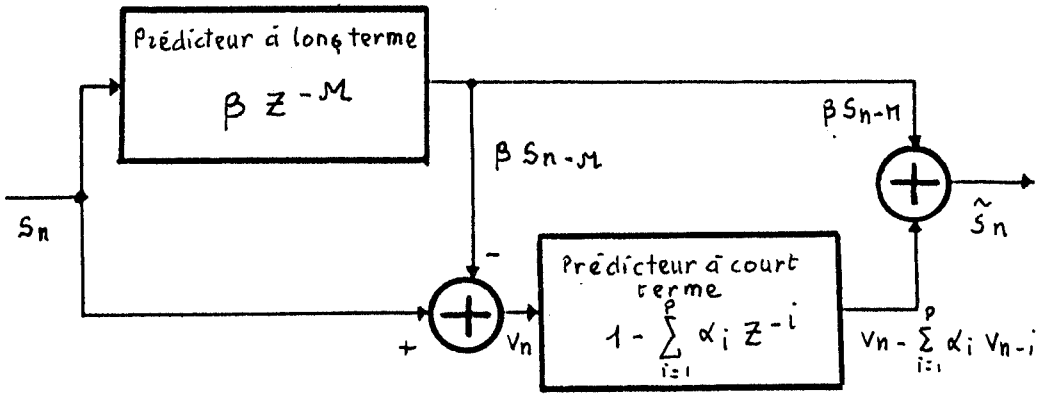


FIGURE 2 : PREDICTEUR DU CODEUR APC

7èmes JOURNEES D'ETUDE SUR LA PAROLE

NANCY 19 au 21 mai 1976

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE PAR L'EMPLOI DE LA TRANSFORMATION DE WALSH-HADAMARD

A.MURE-RAVAUD : E.E.S.P. Tananarive (Madag.)

C.BERGER-VACHON : Université de LYON I

RESUME :

On considère un ensemble de formes à distinguer composé de quatorze voyelles prononcées par sept personnes différentes. On propose un système de reconnaissance utilisant la transformée de Walsh-Hadamard du signal vocal.

Nous examinons les performances obtenues suivant les différentes sélections de caractéristiques et montrons la possibilité d'implantation sur microprocesseur en estimant le temps de calcul et le nombre de mémoires nécessaires au calcul de la transformée. Le système conviendrait à un interface homme-machine dans le cadre d'un vocabulaire limité.

SUMMARY :

We consider a group of patterns to be distinguished composed of fourteen vowels pronounced by seven different persons. A pattern recognition system using the Walsh-Hadamard transformation of the voice signal is proposed.

We examine the performances obtained according to different selections of features and show the implementation possibility on a microprocessor by estimating the computation time and the memory requirement of the transformation. The system would fit a man-machine communication system using a limited vocabulary.

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE PAR L'EMPLOI DE LA TRANSFORMATION DE WALSH-HADAMARD

A.MURE-RAVAUD - C.BERGER-VACHON

I - TRANSFORMATION DE WALSH-HADAMARD

La transformation de Hadamard est utilisée pour la transmission d'images digitales dans un canal ainsi que pour la synthèse et la transmission de la parole. On constate que les déformations subies au cours de la transmission sont moins importantes quand elle est effectuée à partir des formes transformées.

1-1 Forme matricielle de la transformation

La transformation de Walsh-Hadamard [1] fait partie de la théorie plus générale qui consiste à remplacer le système de fonctions orthogonales sinus et cosinus de la transformation de Fourier par un autre système de fonctions qui sont les fonctions de Walsh [2].

La matrice de Hadamard, appelée matrice H, est une matrice carrée d'ordre N ($N=2^n$, n entier) contenant des éléments qui ne peuvent prendre que les deux valeurs +1 et -1 ; elle est telle que :

$$H \cdot H^T = I \qquad \text{(I matrice unité d'ordre N)}$$

Soit une matrice $A = [a_{ij}]$ de N^2 points représentant une figure où chaque élément a_{ij} représente l'intensité échantillonnée de la figure. La transformée en H bidimensionnelle $\mathcal{H}(A)$ de la matrice A est donnée par le produit de matrices

$$\mathcal{H}(A) = H \cdot A \cdot H$$

$\mathcal{H}(A)$ est une matrice carrée d'ordre 2^n.

Il est donc nécessaire que la figure à transformer soit mise sous forme de matrice carrée dont la dimension est une puissance entière de 2.

1-2 Propriétés du point de vue de la reconnaissance de formes

Cette transformation de données est linéaire et orthogonale ; elle conserve l'énergie et l'entropie. Ces propriétés permettent d'effectuer des traitements partiels avant la transformation sans que soient remises en cause par l'application de celle-ci les principales caractéristiques, par exemple la normalisation de l'énergie de la forme. D'autre part, l'essentiel de l'information, dans la suite de données hautement corrélées de la forme d'entrée, est compacté en un petit nombre de composantes beaucoup moins corrélées dans la forme transformée. La transformation accroît ainsi la discrimination entre les classes pour un nombre donné de caractéristiques [3].

L'effet des bruits sur certaines composantes de chaque forme est grandement atténué parce que chaque point de la suite transformée contient une contribution de tous les points de la forme originale.

Les composantes intéressantes du point de vue de la reconnaissance devront être cherchées dans trois zones répondant à des objectifs différents :

a) La première colonne de la matrice transformée est insensible aux translations horizontales de la forme originale. De même, la première ligne est insensible aux translations verticales.

b) Les composantes de faible séquence (la séquence dans la transformation de Hadamard est analogue à la fréquence dans celle de Fourier) qui résultent des sommes et différences de grandes portions de formes sont moins sensibles aux bruits, petites distorsions de formes et petites variations de position.

c) Quelques informations de séquence élevée sont cependant utiles pour distinguer des formes qui ne diffèrent que par de petits détails.

Il sera donc intéressant de chercher ces éléments, en nombre limité, afin de réduire les temps de calcul et la place mémoire.

1-3 Théorème de décomposition

La transformée en $\mathcal{H}$ d'une figure A binaire, composée de lignes verticales de 1 (les 1 sont consécutifs), dans les colonnes i, j, k, ..., l peut s'écrire :

$$\mathcal{H}(A) = \underline{F}_a^b \cdot \underline{H}^i + \underline{F}_c^d \cdot \underline{H}^j + \underline{F}_e^f \cdot \underline{H}^k + \ldots + \underline{F}_m^n \cdot \underline{H}^l$$

où $\underline{H}^i$ est le i^e vecteur ligne de la matrice de Hadamard.

$\underline{F}_a^b$ est un vecteur colonne dont les indices a et b désignent les numéros des lignes commençant et finissant par une suite de 1 ; il a pour expression :

$$\underline{F}_a^b = \begin{bmatrix} \sum_{j=a}^{b} h_{1j} \\ \vdots \\ \sum_{j=a}^{b} h_{Nj} \end{bmatrix}$$

Dans le cas qui nous intéresse, a est égal à b, et il n'est pas nécessaire d'effectuer des sommations.

Chaque quantité $\underline{F}_a^b \cdot \underline{H}^i$ est une matrice d'ordre 2^n ; la transformée de Hadamard est donc obtenue par une somme de matrices très simple.

II - RECONNAISSANCE DES PRINCIPALES VOYELLES PARLEES

Le travail réalisé porte sur la reconnaissance de quatorze voyelles parlées couramment utilisées. Nous les définissons par leur signe phonétique et le numéro de classe correspondant utilisé dans la suite.

1	-	/ɶ/ (pâte)	;	8	-	/ɛ/	(taie)
2	-	/a/ (patte)	;	9	-	/u/	(cou)
3	-	/ə/ (je)	;	10	-	/ɔ̃/	(don)
4	-	/i/ (lit)	;	11	-	/ɛ̃/	(fin)
5	-	/o/ (beau)	;	12	-	/œ̃/	(un)
6	-	/ʏ/ (dur)	;	13	-	/ø/	(bleu)
7	-	/e/ (blé)	;	14	-	/ɑ̃/	(dans)

Ces voyelles ont été prononcées par sept personnes différentes.

2-1 - Description de la forme

La description est effectuée à partir du signal électrique, représentant le signal vocal qui subit une mise en forme classique réalisée par le système de la figure 1.

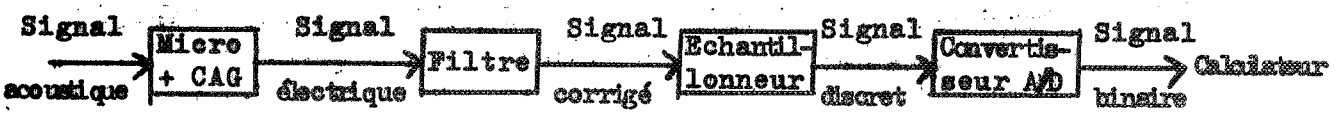


Figure 1 - Prétraitement du signal

* * * *

Le filtre a pour but d'éliminer les fréquences supérieures à 4 kHz et de relever le niveau énergétique des moyennes fréquences à partir de 1500 Hz [4]. Il est réalisé à l'aide de deux circuits passe-bas du deuxième ordre.

La transformation de Hadamard nous oblige à mettre le signal sous forme matricielle ; le nombre de points d'échantillonnage dans le temps doit être égal au nombre de points d'échantillonnage de la valeur de la tension(qui doit être égal à une puissance entière de deux) Nous devons traiter au minimum une pseudo-période et pour être logique, échantillonner à une valeur correspondant à 4 kHz ; pour ce faire, nous avons considéré des "segments de parole" de 10 ms environ échantillonnés en 64 points. La valeur de la tension est donc donnée par six bits, ce qui est faible par rapport à la valeur dix souvent considérée comme nécessaire.

La particularité du signal vocal (fonction bi-univoque du temps) mène donc à une matrice 64 x 64 dont chaque colonne comporte un élément égal à 1, tous les autres étant nuls. A la sortie du convertisseur, nous obtenons, à chaque instant d'échantillonnage, la valeur binaire du numéro de la ligne non nulle du vecteur colonne de la forme matricielle à transformer. Par exemple,si l'énergie au j^e instant (correspondant à la j^e colonne de la matrice A) vaut 37, le 37^e élément de la j^e colonne vaudra 1 et tous les autres éléments seront nuls.

2-2 Méthode de reconnaissance

Les méthodes de reconnaissance ne sont pas très élaborées puisqu'il s'agit d'un travail exploratoire dans une direction particulière.

La première méthode est basée sur deux hypothèses :

- Caractéristiques statistiquement indépendantes
- Densités de probabilités normales.

L'application de la règle de décision de Bayes mène au calcul des distances euclidiennes séparant la forme transformée examinée de la forme moyenne de chaque classe, et consiste à affecter la forme observée à la classe pour laquelle cette distance est minimale (méthode du plus proche voisin).

L'abandon de la seconde hypothèse nous fait considérer les variances relatives à chaque variable et calculer les distances normalisées (encore appelées distances probabilistes ou distances de Mahalanobis). Ceci constitue la seconde méthode.

Dans les deux cas, nous avons supposé des probabilités d'apparition des différents phonèmes égales. Ceci est rarement vérifié puisque la fréquence des phonèmes dépend de la langue utilisée, de l'emplacement dans le mot et de la probabilité d'apparition des diphonèmes [5].

Le système de reconnaissance a donc la structure représentée sur la figure 2 :

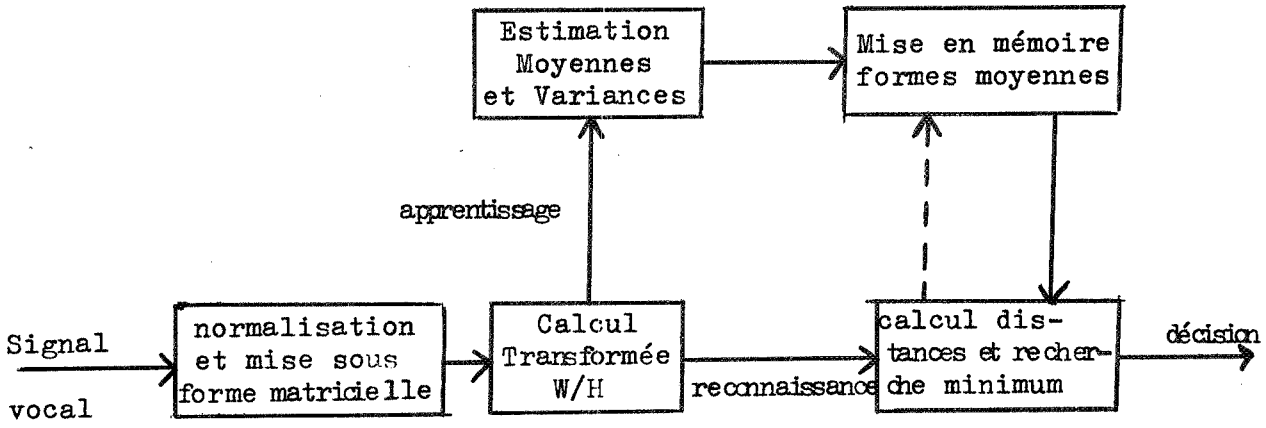


Figure 2 - Structure du système de reconnaissance.

* * * *

2-3 Recherche des cellules caractéristiques

Afin d'obtenir un système performant au point de vue temps de reconnaissance, nous cherchons d'une part à diminuer le nombre de caractéristiques et d'autre part à définir une méthode pouvant convenir à un traitement parallèle plus rapide que le traitement séquentiel. C'est pourquoi nous cherchons pour chaque classe les cellules statistiquement représentatives de cette classe et suffisamment différentes, quant à la moyenne correspondante de celles des autres classes.

Nous sommes donc amenés à calculer pour chacune des 4096 cellules de la matrice $\mathcal{H}(A)$ un facteur d'importance noté f_{ik} (i et k étant les numéros des classes étudiées par leurs moyennes et leurs variances).

En choisissant $$f_{ik} = \frac{|m_i - m_k|}{\sigma_i + \sigma_k}$$

où f_{ik} est le coefficient de séparation entre les classes i et k obtenus pour une cellule

m_i et m_k sont les moyennes des valeurs obtenues dans la cellule pour les classes i et k

σ_i et σ_k sont les écarts-type correspondants,

on obtient une séparation, entre les zones de présence à 95%, lorsque f_{ik} est supérieur à 1,6 [6]. Cette condition est trop restrictive pour un nombre de classe élevé car nous ne trouvons pas de cellules pour lesquelles

$$f_{ik} \geqslant 1{,}6 \qquad \left| \begin{array}{l} k = 1,14 \\ i \neq k \end{array} \right.$$

Le facteur d'importance défini par la relation

$$f_{ik} = \frac{p(m_i+\sigma_i/\omega_i) \cdot p(m_i-\sigma_i/\omega_i)}{p(m_i+\sigma_i/\omega_k) \cdot p(m_i-\sigma_i/\omega_k)}$$ est $\gg 1$ comme on peut le voir ci-dessous.

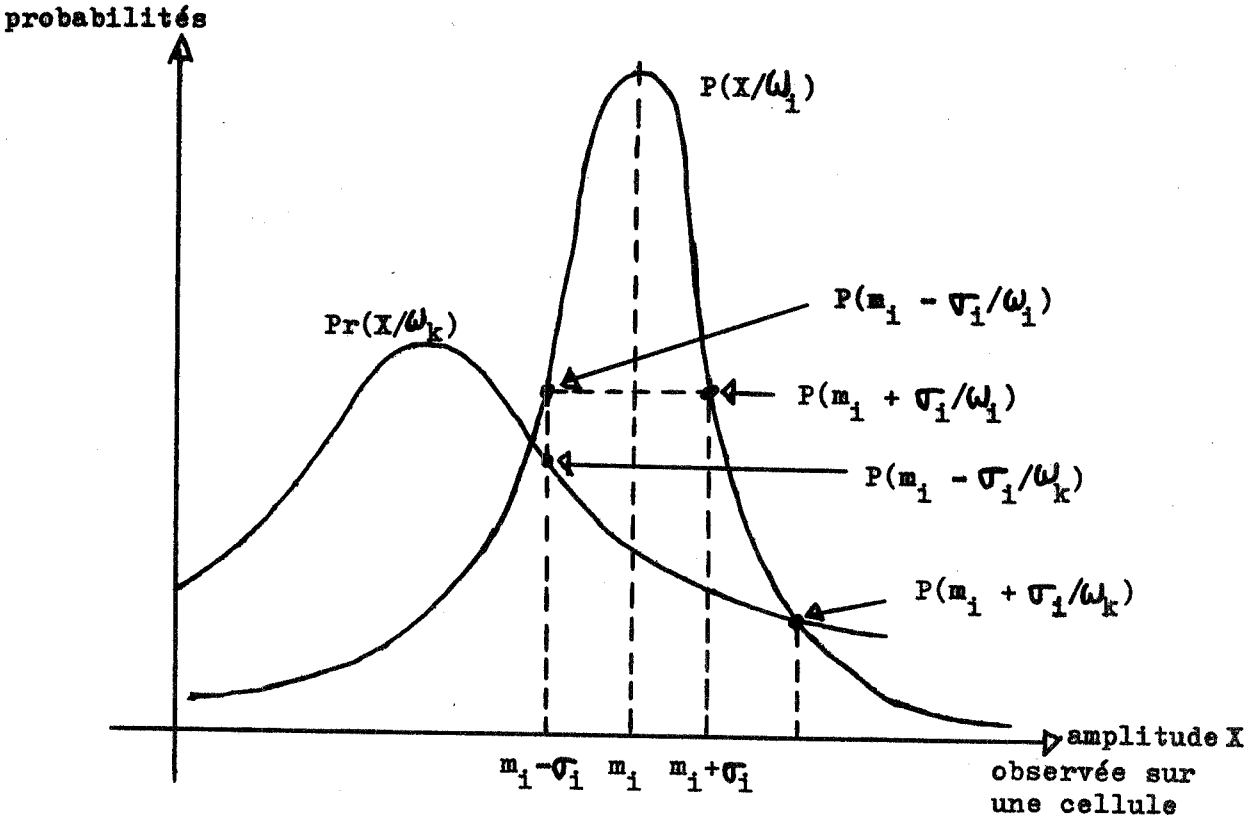


Figure 3 - Densités de probabilité de l'énergie observée pour un caractère (ou dans une cellule de la matrice $\mathcal{H}(A)$) lorsque les classes ω_i et ω_k sont analysées par le système.

* * * *

Cette expression se met sous forme simplifiée suivant les hypothèses faites et conduit à la mise en évidence de cellules caractéristiques qui sont approximativement en nombre de 50 par classe.

Il est évidemment intéressant de trouver des cellules identiques pour la séparation de plusieurs classes car le calcul de la transformée est alors effectué une seule fois pour plusieurs classes.

III - RESULTATS DE LA RECONNAISSANCE

Nous avons vu, dans les propriétés de la transformation, que la première colonne est insensible aux translations horizontales (donc au temps) et qu'elle contient de nombreux éléments de faible séquence ; il apparaît donc utile d'effectuer une reconnaissance à partir de ces caractéristiques en nombre limité (64 sur 4096) afin de comparer avec les résultats obtenus à partir des cellules choisies par des méthodes statistiques.

3-1 Classement utilisant les éléments de la première colonne

La distance euclidienne d_e^k relative à la classe k est donnée par l'expression

$$d_e^k = \sum_{i=1}^{64} (x_{ij} - m_{ij}^k)^2 \qquad j = 1$$

où : x_{ij} est la valeur observée dans la cellule, ou caractère, (i,j) de $\mathcal{X}(A)$

m_{ij} est la moyenne des valeurs correspondant à la classe ω_k dans cette cellule.

La recherche de la distance minimale pour chaque échantillon donne le tableau de classement suivant. La première ligne et la première colonne indiquent respectivement le numéro de classe du phonème observé et le numéro du locuteur. A l'intérieur est noté le numéro de classe décidé par le système de reconnaissance.

Nº de voyelle / Nº de locuteur	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	7	2	3	13	-	-	1	8	-	9	14	6	13	4
2	1	-	3	9	5	6	3	8	9	12	11	-	13	14
3	1	2	3	10	5	6	7	8	9	14	11	-	13	14
4	1	2	8	4	5	6	7	12	9	-	11	12	13	14
5	1	8	3	4	-	5	7	8	9	-	-	2	13	14
6	12	2	-	-	-	6	7	8	9	10	11	3	3	14
7	1	2	-	7	5	6	7	8	9	10	14	12	13	14

3-2 Classement utilisant dix cellules caractéristiques par classe

Les essais ont montré qu'il était nécessaire d'utiliser au minimum dix cellules caractéristiques par classe pour obtenir un niveau de performance identique aux résultats précédents. Le classement obtenu par calcul de la distance euclidienne est donné ci-dessous.

Nº de locuteur \ Nº de voyelle	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	3	4	-	-	10	10	-	10	11	12	5	3
2	1	-	4	4	5	6	7	11	7	10	11	-	13	14
3	1	2	11	4	5	4	7	8	11	10	11	-	11	14
4	1	2	3	14	10	6	7	8	9	-	11	12	11	4
5	4	2	3	4	-	6	7	5	7	-	11	12	4	14
6	1	2	3	4	5	1	7	8	10	10	11	12	3	14
7	1	2	-	4	5	5	10	3	9	10	13	12	13	5

Le remplacement de la distance euclidienne par la distance normalisée (probabiliste) améliore les résultats comme le montre le tableau de classement suivant ; la distance normalisée est :

$$d_e^k = \sum_{i=1}^{10} \left(\frac{x_{ij} - m_{ij}}{\sigma_{ij}} \right)^2$$

Nº de locuteur \ Nº de voyelle	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	3	4	-	-	7	8	-	10	11	12	13	14
2	1	-	3	4	5	6	1	8	9	2	11	-	13	14
3	1	2	11	4	5	6	7	8	11	10	11	-	13	14
4	1	2	3	14	10	6	7	8	9	-	11	12	13	4
5	1	2	3	4	-	6	7	8	9	-	11	12	13	14
6	1	2	3	4	5	5	7	8	9	10	11	12	13	14
7	1	2	-	4	5	5	8	2	9	10	13	12	13	14

Les deux premiers tableaux font apparaître des résultats globaux respectifs de 62 et 60 décisions correctes sur les 84 formes présentées. L'utilisation de la distance normalisée porte ce nombre à 75 voyelles reconnues exactement.

IV - IMPLANTATION SUR MICROPROCESSEUR

La méthode de classification utilisée s'adapte mal aux micro-ordinateurs et d'autres méthodes sont actuellement testées [7]. Par contre, la transformation de Hadamard a de nombreux avantages de ce point de vue. La matrice H représentant la transformation a des éléments réels à valeur binaire (+1 ou -1) que nous avons remplacés par 0 et 1 [8] ; elle occupe ainsi 512 mots de 8 bits de mémoire au lieu de 4096. D'autre part, l'utilisation de la méthode de décomposition se trouve encore simplifiée par le fait que le signal vocal mène à une forme matricielle ayant un seul élément non nul par colonne. Le double produit matriciel se trouve ainsi remplacé par une somme de résultats partiels obtenus par des "ou exclusifs". Le nombre de dix cellules caractéristiques par classe (14 classes) nécessite le calcul d'environ 60 éléments transformés seulement du fait des coïncidences de certains éléments. Le temps de calcul de ces 60 éléments transformés de Hadamard du signal vocal a été estimé à 125 ms. Le temps de calcul pour les 64 éléments de la première colonne est beaucoup plus court du fait qu'on peut les calculer en chaîne ; mais cet avantage est largement annulé par l'augmentation du temps qui en découle dans le système de classification. Ces temps de calcul permettent de prévoir la réalisation d'un système de reconnaissance d'un vocabulaire limité en temps réel.

Les auteurs remercient les personnes qui les ont aidées dans la réalisation de ce travail : MM G.MESNARD (Université de Lyon) et J.P.HATON (Université de Nancy).

BIBLIOGRAPHIE

[1] - N.A.ALEXANDRIDIS - "Walsh-Hadamard transformations in image processing"

[2] - N.J.FINE - "The generalized Walsh functions". Trans. Amer. Math. Soc., Vol 69, 1950, pp 66-77

[3] - H.C.ANDREWS - "Multi-dimensional rotations in feature selection". IEEE Trans. Comput., Vol C-20, pp 1045-1051, sept 1971

[4] - R.W.SCHAFFER and L.R.RABINER - "System for Automatic Analysis of voiced speech". The Journal of the acoustical society of America - Vol 47, n°2, pt 2, pp 634-648, feb 1970

[5] - J.P.HATON - "Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole". Thèse d'Etat, 1974, Nancy

[6] - C.BERGER-VACHON - "Conception d'une entrée vocale automatique". Thèse d'Etat, 1975, Lyon

[7] - L.J.ULMAN - "Computation of the Hadamard transform and the R-transform in ordered form". IEEE Trans. Comp. Vol C-19, pp 359, 1970

[8] - A.MURE-RAVAUD - "Reconnaissance automatique des chiffres manuscrits et des principales voyelles parlées - Emploi de la transformation de Walsh-Hadamard". Thèse de spécialité, 1976, Lyon.

GROUPEMENT DES ACOUSTICIENS DE LANGUE FRANCAISE

Groupe de la Communication Parlée

7èmes JOURNEES D' ETUDE SUR LA PAROLE

NANCY - 19 - 21 MAI 1976

Avec la participation

de l' IRIA, du CNRS et de l' AFCET

V O L U M E 2

Exposés de synthèse, Discussions et Tables Rondes

Organisées à l' UNIVERSITE de NANCY 1 par

le LABORATOIRE
D' ELECTRICITE ET D' AUTOMATIQUE

le LABORATOIRE
D' INFORMATIQUE

I.M.P. - C.U.C.E.S.

GROUPEMENT DES ACOUSTICIENS DE LANGUE FRANCAISE

Groupe de la Communication Parlée

7èmes JOURNEES D' ETUDE SUR LA PAROLE

NANCY - 19 - 21 MAI 1976

Avec la participation
de l' IRIA, du CNRS et de l' AFCET

V O L U M E 2

Exposés de synthèse, Discussions et Tables Rondes

Organisées à l' UNIVERSITE de NANCY 1 par

le LABORATOIRE
D' ELECTRICITE ET D' AUTOMATIQUE

le LABORATOIRE
D' INFORMATIQUE

I.M.P. - C.U.C.E.S.

Ce 2ème volume complète les Actes des 7èmes journées d'Etude sur la Parole, organisées sous le patronage du Groupement des Acousticiens de Langue Française, avec la participation de l'AFCET, de l'IRIA et du CNRS, par le Laboratoire d'Informatique et le Laboratoire d'Electricité et d'Automatique à l'Université de NANCY I, les 19, 20 et 21 Mai 1976.

Ce volume contient la conférence plénière du Pr REDDY sur la reconnaissance de la parole, les exposés de synthèse sur les trois thèmes des journées:

- reconnaissance de la parole continue,
- synthèse de la parole,
- analyse du signal vocal.

ainsi que les discussions sur ces thèmes et les comptes-rendus des tables rondes (aide aux handicapés et conversion graphème-phonème).

Jean-Paul HATON Michel LAMOTTE

Co-organisateurs

CONFERENCE PLEINIERE DU PROF. REDDY

"A Review on Speech Understanding Research"

REVUE DES RECHERCHES EN RECONNAISSANCE AUTOMATIQUE DU DISCOURS CONTINU (*)

D.-R. REDDY
Carnegie - Mellon University
Pittsburgh, USA

I - INTRODUCTION.

L'utilisation de la parole comme moyen de communication avec une machine présente un certain nombre d'avantages et de caractères spécifiques :

1) *Vitesse de communication* : la parole est en moyenne 4 fois plus rapide qu'une entrée par clavier.

2) *Temps de réponse total* : l'entrée directe d'informations à partir d'un microphone supprime des étapes intermédiaires.

3) *Fiabilité du système*.

4) *Mode de communication simultané* avec d'autres modes : utilisable par exemple quand les mains sont occupées.

5) *Liberté de mouvement* : il n'est pas nécessaire d'être assis près d'un terminal pour communiquer.

6) *Caractère inné de la parole*, ne nécessitant pas des utilisateurs entrainés. Par contre, la mise au point d'un système de compréhension de la parole est plus aisée avec des locuteurs coopératifs.

7) *Spontanéité de la parole* : elle permet des communications non prévues à l'avance. En contrepartie, un système de reconnaissance doit pouvoir accepter des phases spontanées, comportant éventuellement des fautes de syntaxe ou autres.

(*) Texte rédigé par J.-P. HATON à partir de l'exposé fait par le Professeur REDDY.

8) *Identification possible du locuteur* : ceci introduit une sûreté supplémentaire dans certaines communications.

9) *Fiabilité dans le temps* : un système automatique peut acquérir indéfiniment des informations orales, tandis qu'un opérateur humain se fatigue au bout d'un certain temps.

10) *Faible coût d'installation*.

Cependant, malgré ces avantages, les systèmes de reconnaissance de la parole sont encore peu avancés car les problèmes à résoudre restent énormes. Malgré tout, ces problèmes ne sont pas insurmontables et on peut constater des progrès lents mais continus. On peut donc être raisonnablement optimistes en ce qui concerne la solution future de ces problèmes.

II - DIFFERENTS ASPECTS DE LA RECONNAISSANCE VOCALE.

Le tableau I donne une idée des différents types de systèmes de reconnaissance, suivant leur complexité et les applications envisagées.

III - EVALUATION DES PERFORMANCES DES SYSTEMES DE RECONNAISSANCE DE LA PAROLE.

La taille du vocabulaire reconnu est un paramètre insuffisant pour comparer les performances de différents systèmes. Par exemple, les 3 vocabulaires suivants, composés chacun de 3 mots :

V1 : "B" , "D" et "V"
V2 : "1" , "2" et "3"
V3 : "A" , "B" et "C"

sont de difficultés très différentes.

Les résultats obtenus par Itakura (tableau II) sur 2 vocabulaires de taille très différente sont une autre preuve : dans les mêmes conditions expérimentales, les résultats les plus mauvais sont obtenus sur le vocabulaire le plus petit.

	Mode of Speech	Vocabulary Size	Task Specific Information	Language	Speaker	Environment
Word recognition-isolated (WR)	isolated words	10-300	limited use	-	cooperative	-
Connected speech recognition-restricted (CSR)	connected speech	30-500	limited use	restricted command language	cooperative	quiet room
Speech understanding-restricted (SU)	connected speech	100-2000	full use	English-like	not uncooperative	-
Dictation machine-restricted (DM)	connected speech	1000-10000	limited use	English-like	cooperative	quiet room
Unrestricted speech understanding (USU)	connected speech	unlimited	full use	English	not uncooperative	-
Unrestricted connected speech recognition (UCSR)	connected speech	unlimited	none	English	not uncooperative	quiet room

Different types of speech recognition systems ordered according to their intrinsic difficulty, and the dimensions along which they are usually constrained. Vocabulary sizes given are for some typical systems and can vary from system to system. It is assumed that a cooperative speaker would speak clearly and would be willing to repeat or spell a word. A not uncooperative speaker does not try to confuse the system but does not want to go out of his way to help it either. In particular, the system would have to handle "uhms" and "ahs" and other speech-like noise. The "-" indicates an "unspecified" entry variable from system to system.

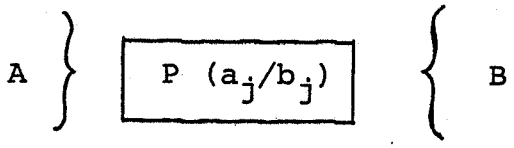
Tableau I

ISOLATED WORDS (Itakura, 1975)

Vocabulary	Recognition Rate (%)	Rejection Rate (%)	Error Rate (%)
Alpha-Digit	88,6	0	11,4
Japanese Geographical Names	97,3	1,7	1,0

Tableau II

NOISY CHANNEL MODEL



EQUIVOCATION = Information Transmitted
- Information Received

= Information Lost

= H(A/B)

EQUIVALENT VOCABULARY SIZE = $2^{H(A/B)}$

["A" , "B" , "C"]	1.138
["B" , "D" , "V"]	2.628
["1" , "2" , "3"]	1.023

Tableau III

Il est donc nécessaire de connaître, en plus de la taille, la "complexité" d'un vocabulaire. Cette connaissance nécessite la définition de modèles rendant compte des ambiguités acoustiques, phonétiques, lexicales, syntaxiques, pouvant apparaître. M. GOODMAN (Carnegie-Mellon University) a mis au point une méthode de calcul de l'ambiguité lexicale d'un vocabulaire. Cette méthode fait appel à des résultats de la théorie de l'information, comme résumé sur le tableau III. Elle permet de mesurer une "taille équivalente" d'un vocabulaire donné, d'autant plus grande que le vocabulaire comporte plus de confusions possibles. Ce coefficient ne constitue pas une mesure absolue, mais il donne une idée relative de la difficulté de reconnaissance d'un vocabulaire. En particulier, il n'est pas forcément lié à la taille du vocabulaire, comme on le voit sur le tableau IV.

Mais cette mesure n'est pas valable pour la parole continue, car la position du mot dans la phrase intervient. On définit alors la notion de taux de branchement moyen dans la grammaire. On peut alors évaluer un "taux équivalent de branchement" en y incorporant la notion précédente valable pour les mots. Quelques valeurs sont indiquées sur le tableau V.

Le but d'un système de reconnaissance de la parole est d'utiliser un maximum d'informations de source différente de façon à faire tendre le "taux équivalent de branchement" vers 1.

IV - MODELES DE SYSTEMES DE RECONNAISSANCE DE LA PAROLE.

1) Le modèle HEARSAY :

La figure 1 présente un des modèles développés par le Professeur REDDY, le modèle "blackboard" HEARSAY. C'est un modèle de traitement parallèle dans lequel un certain nombre de processus, guidés par les données, travaillent indépendamment à lever les ambiguités d'une phrase et communiquent entre eux par un "tableau" central. Ce modèle pose actuellement surtout des problèmes purement informatiques (traitements parallèles, coopération de processus), encore mal résolus. C'est un modèle très puissant, sur lequel les travaux vont se poursuivre dans l'avenir ; pour les travaux actuels concernant la reconnaissance de langages limités, il s'avère finalement trop sophistiqué.

Le problème général de la reconnaissance de la parole est de définir une hiérarchie de niveaux de traitements (acoustique, phonétique, etc...) -chaque niveau introduisant de nouvelles erreurs- capable de réduire l'explosion du nom-

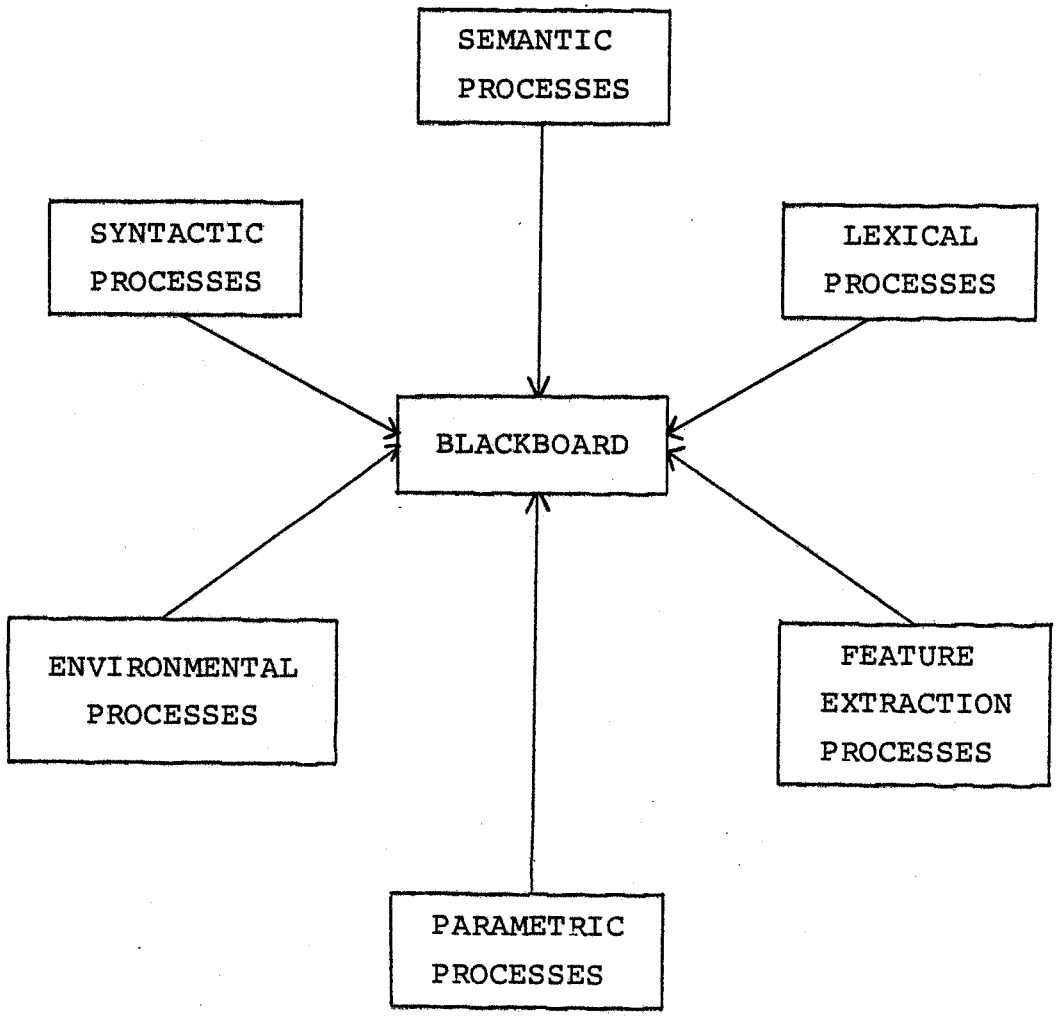
TASK	VOC SIZE	EQUIVALENT VOC SIZE
DIGITS	10	1.18
ALPHA	26	4.18
ALPHA-DIGIT	36	3.58
CHESS	31	1.98
LINCOLN		
Basic	236	3.16
Extended	410	4.63
IBM	250	2.78
Voice Programming	37	2.17

Tableau IV

SYNTAX and LANGUAGE ANALYSIS

	Average Branching Factor	Equivalent Branching Factor
CHESS	7.29	1.30
LINCOLN		
Basic	9.15	1.26
Extended	20.28	1.45
IBM	7.32	1.12
Voice Programming	10.82	1.33

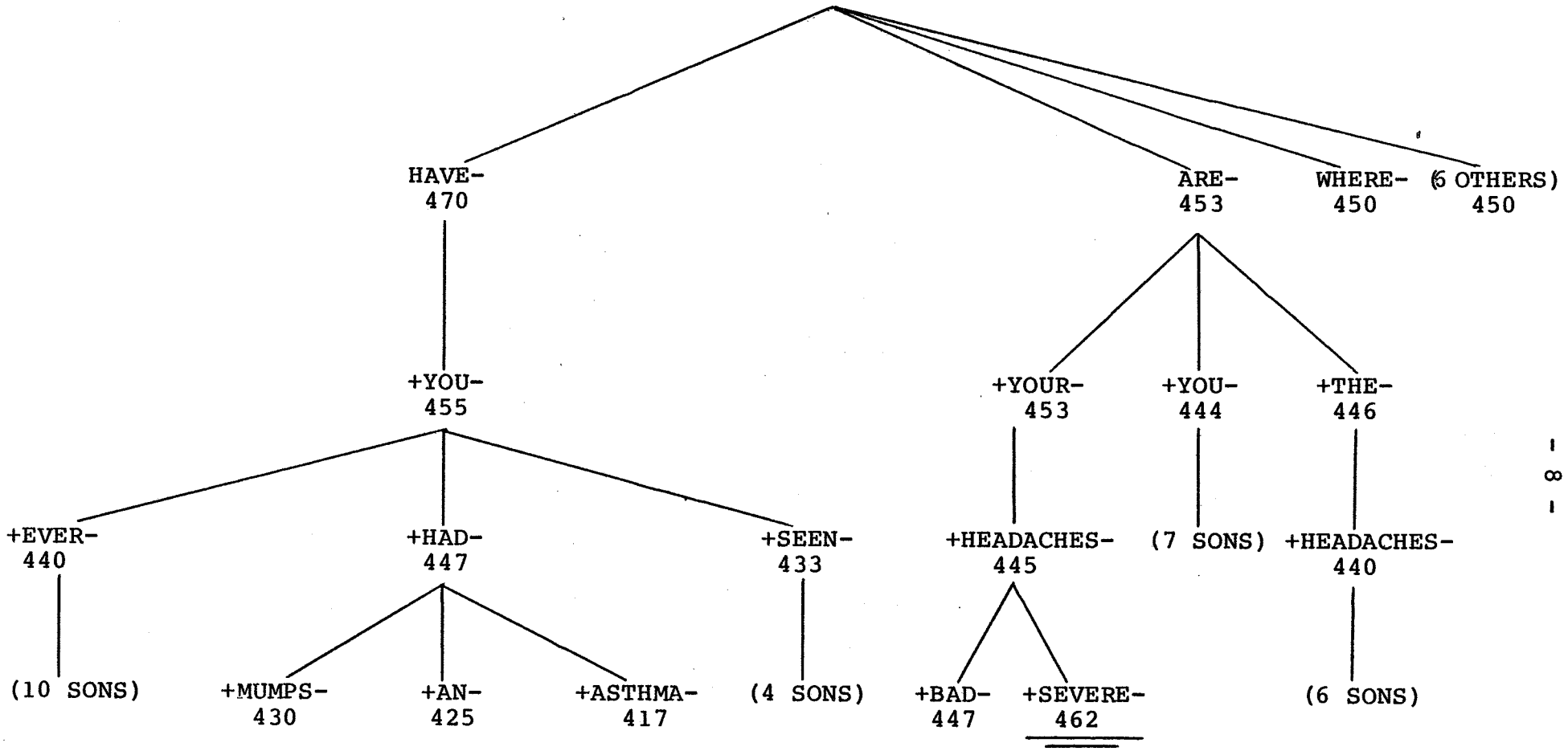
Tableau V



THE BLACKBOARD MODEL

Data Directed
Information Gathering
Hypothesize and Test
Parallel and Independent
Deactivation Simple

Figure 1



ARE YOUR HEADACHES SEVERE ?

Figure 2

bre possible de choix dans l'interprétation d'une phrase.

Dans HEARSAYI , cette tâche est réalisée par un algorithme de recherche descendante, arborescente de type "best-first". Le principe consiste à suivre le chemin ayant, à tout instant, le score le plus élevé (i.e. le plus probable).

Un exemple de traitement est donné figure 2. On constate que le nombre total de chemins qui ont été examinés puis abandonnés avant d'arriver à la fin de la phrase est assez réduit (environ une vingtaine), et très inférieur au nombre de combinaisons possibles.

2) Les systèmes DRAGON et HARPY :

Plusieurs systèmes, de conception très différente, ont été réalisés à la suite de HEARSAY, par l'équipe du Professeur REDDY. En particulier, le système DRAGON de J. BAKER. Ce système utilise un réseau unique pour représenter toutes les sources d'information ; il met en oeuvre un modèle markovien et utilise la programmation dynamique pour la reconnaissance. Ce système, plus lent que HEARSAY, a des performances un peu supérieures.

Enfin, plus récemment, le système HARPY a été réalisé par LOWERRE. Ce système est à la fois plus rapide et plus performant que les précédents.

V - ETAT ACTUEL DES SYSTEMES DE COMPREHENSION DE LA PAROLE CONTINUE.

Le système HARPY figure parmi les systèmes actuels les plus performants. Les performances sur une application de programmation à la voix sont données sur le tableau VI. Ces performances sont nettement supérieures à celles des systèmes précédents, ce qui prouve une évolution positive au cours des trois dernières années.

Le système HARPY fonctionne environ en 13 fois le temps réel. A ce sujet, on peut remarquer que cette mesure est très subjective puisqu'elle dépend du type de calculateur utilisé. Une autre mesure, meilleure mais pas encore idéale, consiste en le nombre d'instructions exécutées par le système par seconde de parole traitée. Cette dernière mesure présenterait l'avantage d'être moins dépendante du matériel utilisé.

HARPY PERFORMANCE
ON FOUR SPEAKERS

	IMMEDIATE TEST DATA	FIVE MONTH TEST DATA
CORRECT SENTENCES (OUT OF 79)	91.1%	77.2%
CORRECT WORDS (OUT OF 436)	97.7%	95.4%
TIME (X REAL-TIME (PDP KA-10)	13.1	13.2
INSTRUCTIONS EXECUTED PER SECOND OF SPEECH (x MILLION)	4.5	4.5
BRANCHING FACTOR	10.8	10.8

Application : Voice-Programming
37 word vocabulary.

Experiment : 79 sentences for training
158 sentences for testing.

Tableau VI

DESIGN CHOICES AFFECTING ACCURACY :

	HEARSAY I	DRAGON	HARPY
WORD REPRESENTATION	BASEFORM DICTIONARY	BASEFORM DICTIONARY	MACHINE DERIVED
JUNCTURE RULES	PROCEDURAL	NONE CAN BE ADDED	YES
PROSODIE RULES	PROCEDURAL	NONE	DURATION CONTOUR (no pitch)
PARAMETRIC REPRESENTATION	OCTAVE FILTERS	OCTAVE FILTERS	LPC (not spectra)

Tableau VII

DESIGN CHOICES AFFECTING SPEED :

	HEARSAY I	DRAGON	HARPY
SEARCH STRATEGY	BEST FIRST AND BACKTRACKING	ALL PATHS IN PARALLEL	BEST FEW FIRST AND BACKTRACKING
SEGMENTATION	YES	NO	YES
PHONETIC MATCH	MATCH ALL PHONES	MATCH ALL PHONES	SELECTION MATCH

Tableau VIII

VI - CHOIX LORS DE LA CONCEPTION D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE.

Lorsque l'on conçoit un système de reconnaissance de la parole, on est amené à effectuer un nombre important de choix (type d'analyse, structure de données, ...) qui vont conditionner les performances. Le tableau VII compare les choix fondamentaux des systèmes successifs décrits par le Professeur REDDY, qui influent sur les performances obtenues.

On peut noter, en particulier, que, dans le système HARPY, les mots sont représentés par un réseau obtenu directement par la machine.

Le tableau VIII indique les choix qui affectent principalement la vitesse du système. Un facteur important est la stratégie d'analyse utilisée. HARPY utilise une technique "best few" qui consiste à mener en parallèle, par une technique de programmation dynamique, les quelques meilleurs chemins, sans retours-arrière.

A côté de ces facteurs, il en existe d'autres dont l'effet est pour l'instant encore mal compris. Le principal résultat est qu'il faut être très prudent dans l'expérimentation des systèmes de reconnaissance. Il est nécessaire de comprendre les raisons de la supériorité d'un système sur un autre de façon à profiter de l'expérience dans l'avenir.

VII - CONCLUSION.

L'évolution des systèmes de reconnaissance de la parole permet d'être optimiste pour l'avenir de ces recherches. Actuellement, l'accent est mis sur la mise en oeuvre de contraintes syntaxiques et sémantiques. Cependant, un système complet ne pourra fonctionner qu'en utilisant toutes les sources d'information et, en particulier, qu'avec un système acoustique et phonétique très performant.

La fin de l'année 1976, correspondant à la fin de l'actuel projet ARPA sur la reconnaissance vocale, devrait voir la mise au point de systèmes performants de compréhension de phrases dans un univers limité. Mais, il ne s'agit que d'une étape et il reste bien d'autres domaines d'investigation en reconnaissance de la parole.

THEME 1 :

RECONNAISSANCE DE LA PAROLE CONTINUE

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE
INTRODUCTION, THEMES ET TENDANCES EN 1976

Jean-Yves GRESSER *
Centre National d'Etudes des Télécommunications

115 rue du Bac 75007 Paris

" H.J.WENSLEY des Laboratoires Westinghouse crée l'automate "Televox" (qui réagit) au son de la voix et d'un instrument musical... les vibrations déclenchent dans l'automate un moteur électrique qui le fait agir selon les indications reçues "

VU. Journal de la semaine 21 Mars 1928 (Couverture).

Résumé

Le texte essaie de remettre en perspective les exposés présentés à ces Journées d'Etudes sur le thème reconnaissance automatique de la parole continue, et de dégager des thèmes de coopération pour les prochaines années.

Summary

This paper is a short synthesis of studies presented in Nancy on recognition of (connected) speech, it contains a general discussion on ASR studies in France and results of a quizz on recent trends.

* également Université Paris XIII (MSTT)

7 ème J E sur la parole . Nancy, Mai 1976

1- Perspective, résultats.

Cette introduction emprunte une partie de son plan à l'exposé de D.R. REDDY sur les travaux américains récents.

I-1 Intérêt de la reconnaissance vocale Science ou technique ?

A la recherche d'un financement des industriels, des utilisateurs possibles ou des organismes de tutelle les laboratoires justifient les travaux actuels par les besoins immédiats d'un certain nombre de secteurs d'activité :

- amélioration du travail manuel,
- simplification, automatisation des taches,
- automatisation des services.

La justification par retombées se pratique de nouveau, après un certain déclin :

- transmission, compression du signal vocal analogique ou numérisé,
- conception automatisée des machines.

Dans ces justifications comme dans leurs travaux les chercheurs sortent rarement des mathématiques ou de la physique appliquée dont ils sont issus.
C'est une constatation curieuse si l'on examine les emprunts faits, au niveau du langage plus qu'à celui des idées, aux sciences humaines.

Les emprunts constituent le choc en retour de la cybernétique, de l'automatisme (par exemple de la théorie des systèmes) partiellement assimilés par les sciences humaines (communication). Malgré le décalage des emprunts des chercheurs d'horizons divers, que l'on pourrait caricaturer en "techniciens" et "humanistes" , se sont rapprochés.

L'utilitarisme des travaux va à l'encontre de ces rapprochements. Les solutions immédiates et totales se trouvent rarement chez les autres, qui ne s'intéressent d'ailleurs qu'à leurs propres problèmes. L'intérêt commun ne se manifeste que sur des questions limitées, dont les réponses auront une signification différente dans les disciplines concernées. Au moins faut-il pouvoir poser les questions, c'est à dire disposer d'un langage et d'une problématique réduits communs.

Aujourd'hui les équipes francophones disposent d'un embryon de langage qui rend possible un début de communication.

Les exposés sur la reconnaissance vocale se font souvent dans un contexte utilitaire mais les préocupations des chercheurs, mises en évidence à la fin de ce texte, sont plus fondamentales. La reconnaissance vocale remet en cause notre conception de la chaine linguistique, de la perception des sons dits élémentaires au processus de compréhension mutuelle de deux interlocuteurs. La conception des phonéticiens, des linguistes, des psychologues etc... doit se compliquer au niveau des modèles qu'ils manipulent. En revanche les ingénieurs, les mathématiciens doivent se garder des abstractions pures et des contresens, sur les phénomènes traités comme sur la finalité de leurs travaux.

Les préoccupations actuelles sont fondamentales. Les questions essentielles à la solution ambitieuse des problèmes pratiques resteront posées pendant plusieurs décennies : l'accélération technique est limitée ; les sciences humaines progressent lentement et avec peu de moyens. Il est important que les "décideurs" de tous niveaux le réalisent et ne sclérosent pas une discipline naissante.

Cela doit être d'autant plus facile que certains avantages de la reconnaissance vocale sont évidents et que des résultats intéressants sont déjà disponibles.

Avantages

Dans son exposé REDDY mentionne 10 avantages de la reconnaissance vocale. Certains snnt évidents comme :

- la vitesse de communication. Comparée à l'utilisation de leviers, du clavier, à l'écriture manuscripte l'énoncé par mots, par phrases est nettement plus rapide. La linéarité de la parole la rend incomplète dans la communication visuelle. Mais l'utilisation de pseudo-langages linéaires couvre un grand nombre d'applications,

- la fiabilité. Des comparaisons clavier/parole ont montré la supériorité de la dictée vocale sur la frappe. Cette supériorité est probablement dûe à la régulation instantannée de l'élocution. Cette fiabilité est évidemment limitée dans la pratique par les performances des "reconnaisseurs", non par le taux d'erreurs nouvelles introduites (en général d'un ordre de grandeur inférieur aux erreurs naturelles, humaines) mais par l'étendue du vocabulaire traité.

- le parallélisme. La parole fournit un canal de commande supplémentaire, alors que les autres canaux sont déjà occupés. Il reste à déterminer comment le travailleur (ouvrier spécialisé ou pilote de véhicule cosmique) peut supporter un processus où l'ensemble de ses canaux de communication est occupé. La parole est souvent dans le cas des tâches répétitives, un dérivatif à l'ennui, la fatigue. Elle évite des déréglements plus graves. Une utilisation continue est probablement à déconseiller.

- la simplification des chaînes de travaux. Dans la communication verbale il est classique de relever le nombre de fois où un même message est recopié avant d'être traité ou même simplement transmis. Il est aussi classique de rappeler que la recopie est une source d'erreurs multiples et difficilement contrôlables .
La dictée directe supprime au moins une à deux recopies. Elle permet aussi, effectuée en temps réel, une vérification immédiate du texte obtenu. Elle simplifie la chaîne des travaux par suppression de tâches, comme de cycles entiers lecture -vérification habituellement espacés et longs.
Peut-être n'est-elle pas indispensable à l'ensemble des simplifications. La lecture vérification immédiate peut se faire sur le seul mode écrit. Mais la communication écrite nous a habitués à des réactions différées alors que la communication verbale est habituellement instantannée. L'idée lecture-vérification immédiate vient de la parole.
La simplification des tâches peut aller juaqu'à l'automatisation complète ou partielle d'un processus, y compris un processus de communication où l'interlocuteur humain est remplacé par un robot (cela est envisagé pour les centres de renseignements) .

Vitesse, fiabilité, parallélisme, simplification sont déjà mis en évidence par l'utilisation des reconnaisseurs actuels, qui sont pourtant limités à quelques dizaines à quelques centaines de mots pour un seul ou un petit nombre de locuteurs.

Les autres avantages mentionnés (communication naturelle, pour grand public, l'économie) restant à prouver.

Dans ce contexte anormal de la communication homme-machine la parole est-elle le moyen approprié ?

Dans un contexte d'échanges parlés entre humains un robot d'intelligence limitée peut-il se substituer à l'un des interlocuteurs ?

L'ambition de la reconnaissance automatique dépend des réponses à ces questions. Ces réponses seraient le seul point de départ sérieux d'un courant "utilitaire".

Certaines études, qui ont l'avantage de pouvoir se passer d'un reconnaisseur, cherchent à apporter des réponses.

C'est un signe de maturité encourageant.

Quant à l'économie apportée, il faudra juger au moment où les reconnaisseurs seront moins limités.

I-2 Classement des recherches.

Le lecteur pourra se reporter à plusieurs articles antérieurs de Cessel, Haton ou l'auteur, sur l'histoire des études, et des tentatives de classifications plus complètes.

I-2-1 Le langage reconnu

Vocabulaires "mots isolés, 10-300, usage limité, déjà résolu"
Des systèmes expérimentaux existent déjà dans cette gamme à Lannion (CNET), Orsay, Toulouse (Université,IUT), Marcoussis (Centre de Recherches de la CGE), Saclay (CEA)

A ce jour il n'y a aucun projet commercial sérieux fondé sur les systèmes développés, contrairement à ce qui se passe aux Etats-Unis ou en URSS, où les débouchés sont encore lents à apparaître.

"Parole continue, langage artificiel 30-300 mots, usage limité commande".

Il existe des systèmes expérimentaux réels ou simulés, selon l'étendue du vocabulaire, à Lannion et Nancy. Le langage de commande est une étape des approches utilitaires.

"Compréhension de la parole, 100 -2000 mots, usage illimité, pseudo (français) machine à dicter, 1000 -10 000 mots, usage limité, pseudo ... "

Ici les recherches menées en France se démarquent des recherches menées aux Etats-Unis malgré les travaux de Lannion et Nancy. L'étude de la parole continue accuse un retard d'environ 2 à 3 ans. Par contre l'étude des vocabulaires de mots isolés a été poussée plus loin dès 1973 (Lannion) à la suite des travaux soviétiques, jusqu'aux environs de 1000à 1500 mots.

Le projet SUR, américain, aurait récemment rattrapé le retard sur les grands vocabulaires.

Originalité La notion de dialogue homme-machine est pratiquement absente des travaux étrangers, alors qu'elle est au centre des recherches avancées menées en France, depuis les débuts modestes du calculateur Vocal jusqu'au projet ambitieux d'automatisation des centres de renseignements. Elle implique dans le processus de compréhension de la machine, non seulement la compréhension du langage utilisé mais encore celle du comportement de son interlocuteur. Le dialogue se déroule sur deux plans : concret par échange de signaux, entre l'homme et la machine ; abstrait à l'intérieur de la machine entre elle-même et l'image de son interlocuteur.
La formalisation du dialogue est en cours à Lannion.

Le dialogue avec des machines d'intelligence limitée recourt aux langages artificiels. Il va à l'encontre du mythe communication naturelle.

1-2-2 Les performances

Complexité des langages (voir 1-2-1)

Taux de reconnaissance. Dès qu'un auteur présente des résultats ceux-ci sont uniformément bons. L'échantillon expérimental, les limites du reconnaisseur sont choisies en conséquence. Il est difficile de classer les travaux d'après les publications.

Ce n'est qu'à partir des études sur la parole continue que l'on a vu annoncer sans rougir des taux de 60% de reconnaissance. Les taux mesurés sur des échantillons infimes, quelques phrases alors qu'il faudrait des milliers, ne compromettent personne. Ni Nancy, ni Lannion, les seuls à s'être aventurés sérieusement dans le domaine, n'échappent à ce travers.

Donc les limites sont ailleurs :

- dans le nombre de locuteur qu'une machine peut comprendre, simultanément sans adaptation fastidieuse,

- dans l'ensemble des phonèmes transcrits aves exactitude. Mais on ne sait même pas dire en phonétique automatique si les problèmes ne sont pas déjà résolus. Le reconnaisseur automatique est peut-être supérieur à la machine humaine. Les expériences limitées sur les voyelles le suggèrent.

2 - Thèmes

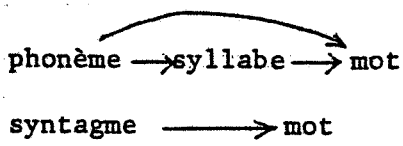
2-1 Description des reconnaisseurs

La plupart des travaux se place dans un cadre théorique ou pratique élargi où le but final est la mise en oeuvre d'un système complet. Dans un même système les processus élémentaires ne sont pas forcément homogènes (Lannion, Nancy, Toulouse). Seule l'ENST à Paris poursuit une approche globale où les mêmes types de processus s'emboitent d'un niveau à l'autre du reconnaisseur.

Les reconnaisseurs peuvent être décrits sur deux plans :
celui des traitements, celui des entités linguistiques manipulées. Dans chaque plan la représentation en "cadran" est commode. Elle évite une hierarchisation abusive des processus ou des écritures (listes d'entités linguistiques) intermédiaires. Le cadran a déjà été utilisé pour engendrer l'enchainement logique, et l'enchainement réel des traitements de certains reconnaisseurs ( ).

Il y a une certaine redondance d'un cadran à l'autre. Un traitement peut se définir comme une ou plusieurs manipulations d'entités, représentées par un graphe construit sur le cadran inférieur.

exemple recherche lexicale

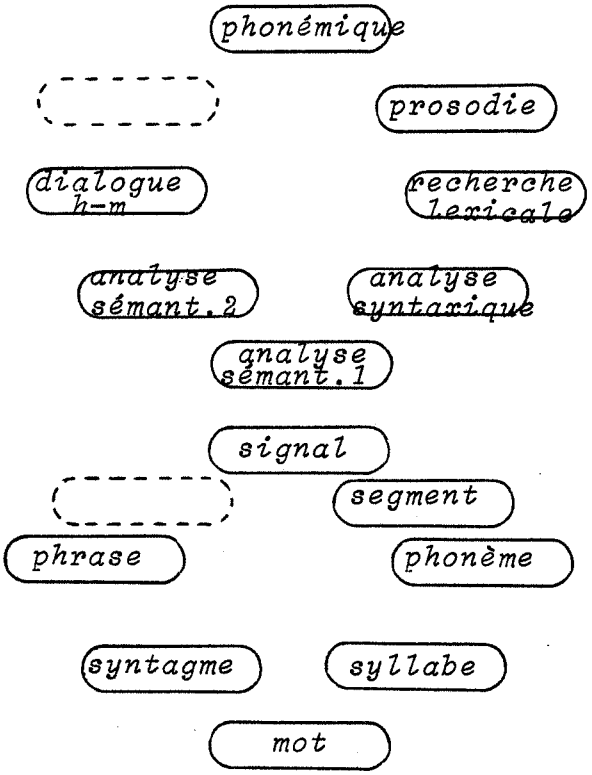


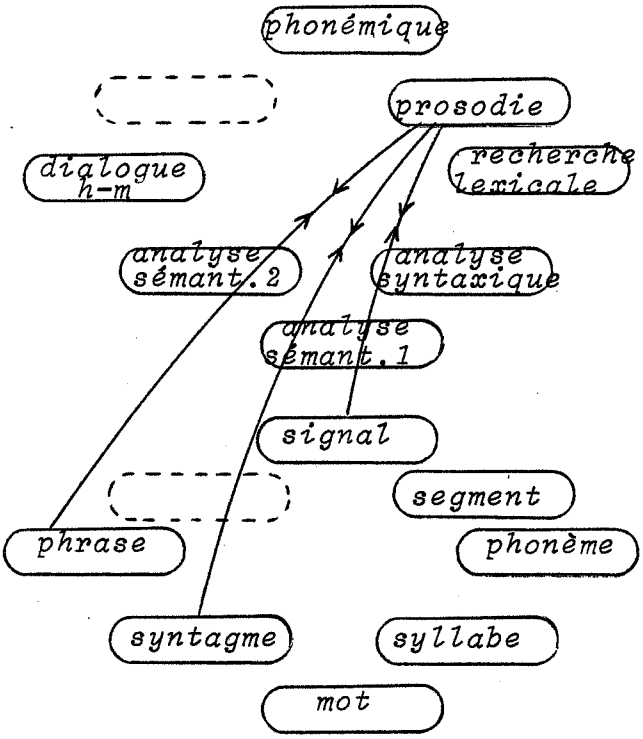
Le cadran des traitements est nécessaire pour une description simple du programme du reconnaisseur.

Le schéma proposé est encore incomplet. il ne décrit pas les éléments extérieurs au reconnaisseur : la qualité du signal initial, l'environnement, l'adaptation aux locuteurs, etc...

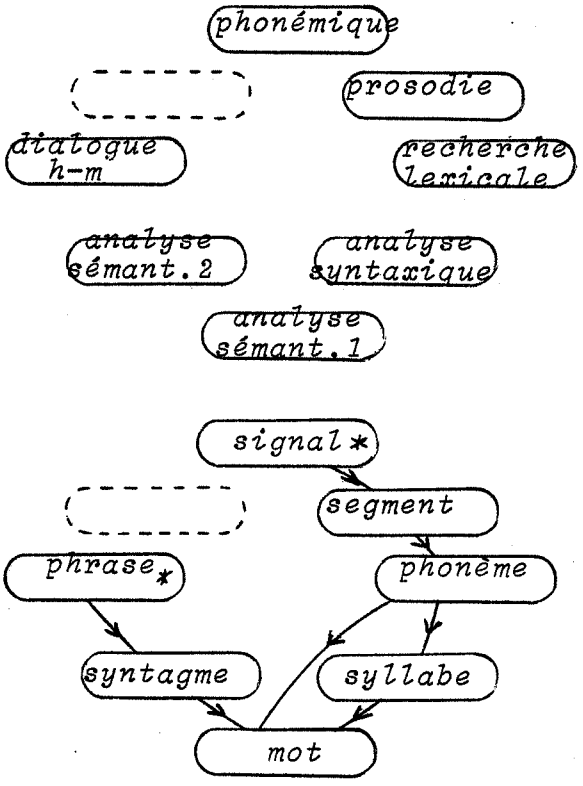
Dans les travaux présentés à ces journées la sémantique et le dialogue, homme-machine sont absents.

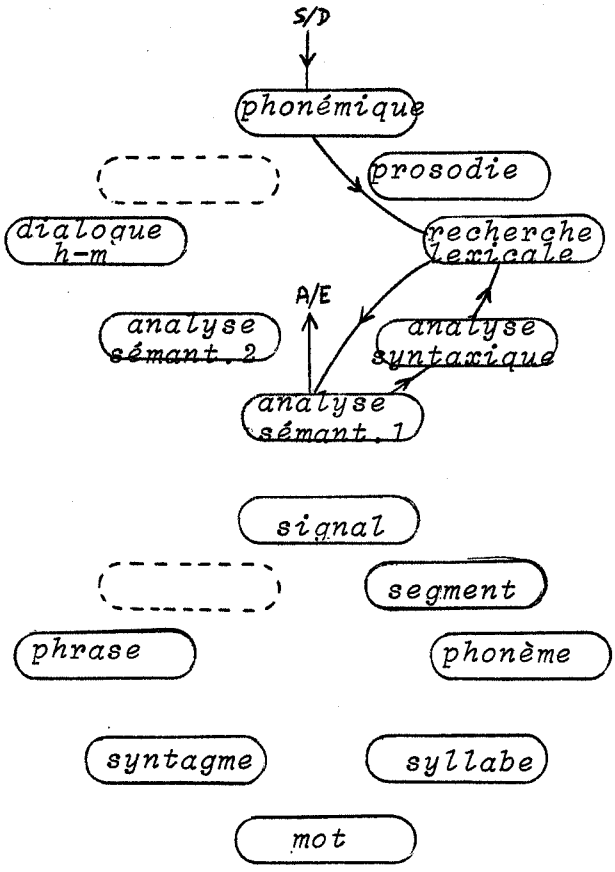
Le domaine couvert est moins étendu qu'aux J.E. de Toulouse où l'accent était mis sur l'interpénétration des niveaux linguistiques.





La représentation en cadran et son utilisation





Les textes sont des fragments de développements plus importants que les auteurs ont bien voulu exposer ou rappeler dans leur présentation en se référant au besoin schéma descriptif proposé.

2- 2 Parole continue, 4 analyses

Le terme parole continue emprunté à l'anglais (connected speech) n'a de sens que dans le jargon des chercheurs. Je lui préfère ceux de parole naturelle ou élocution naturelle qui expriment mieux l'énonciation sans pause artificielle d'une parole "liée".

Les recherches présentées ne sont pas toutes spécifiques de la parole naturelle. La distinction entre "mots isolés" et parole naturelle n'est pas toujours très nette même dans l'analyse syntaxique.

La prosodie n'a guère de sens dans un cadre artificiel.

L'étude objective des sons articulés (baptiséé phonique) est presque indépendante de l'introduction des pauses (cela ne contredit pas le fait qu'une pause puisse être interprétée comme un son).

Lexique (voir exposés de J. BREMONTetc., B;CAUSSE etc., R.VIVES)

En reconnaissance automatique un mot est un élément d'une liste appelée lexique ou dictionnaire. Le dictionnaire est défini par son contenu, le nom du mot, et par sa ou ses méthodes d'accès, la ou les descriptions des mots. Car si un mot a un nom il peut avoir plusieurs interprétations ou composantes en écritures. Certaines écritures sont entachées d'erreur.
C'est la particularité de la reconnaissance automatique de chercher à reconnaitre des mots mal écrits.

Autre particularité de la parole naturelle : absence de pause forcée signifie souvent absence de pause tout écrit (insensible au reconnaisseur).

La recherche lexicale doit comporter ou être précédée d'une phase de cadrage du mot dans une écriture continue.

La consultation des listes longues, nécessaire pour les grands vocabulaires est un problème classique facilité par les descriptions multiples qui permettent des accès accélérés, parfois hiérarchisés.

BREMONT etc. décrivent une méthode séquentielle ascendante sur des mots phonétiques à 2 dimensions, complétée (comment ?) d'une analyse syntaxique simple. Cette méthode parait utilisable sur des petits vocabulaires.

CAUSSE etc. fondent leur description sur la syllabe l'accès ascendant au lexique est corrigé par des règles de réécriture syllabiques ou phonétiques utilisant des marqueurs prosodiques ou phonologiques particuliers.
On attend ici aussi beaucoup de l'analyse syntaxique pour limiter le nombre de consultations à effectuer.

VIVES décrit certaines propriétés d'une méthode non séquentielle, déjà connue : stabilité à certains seuils (similitude entre phonèmes, détection) ; hiérarchisation des accès.
Il est intéressant d'accoler cet exposé à l'exposé de QUINTON sur l'analyse syntaxique. La même méthode non séquentielle est appliquée à la détection des mots dans la phrase.
Ce type de méthode est à mon avis mieux adapté à l'étude des grands vocabulaires.

Syntaxe (voir exposés de HATON etc., QUINTON).

La syntaxe est l'enchaînement des mots (éléments du lexique).

La syntaxe joue-t-elle un rôle priviligié dans l'architecture du reconnaisseur ? L'affirmative est défendue par HATON etc.

Quels mécanismes mettre en oeuvre pour s'accomoder d'une écriture erronnée ?

A la suite d'autres travaux menés à Lannion le texte de QUINTON est probablement à ces J.E. le plus complet sur les mécanismes de choix des hypothèses et de reprise. La maitrise de l'étendue du contexte manipulé est encore incomplète.

La solution est peut-être ailleurs dans des méthodes progressives (ou latérales, au contraire de bas en haut) et parallèles.

Le texte de QUINTON contient les seuls résultats globaux présentés sur le thème. Ils sont modestes mais valent : la peine d'être relevés (p.102 vol.I). (Voir feuille jointe).

Prosodie (voir exposé de J. VAISSIERE).

Ce thème avait été traité en détail à Toulouse. VAISSIERE présente certaines idées sur la place de la prosodie, autour d'une méthode simple de segmentation en mots intonatifs. Il faut bien reconnaître que jusqu'à Nancy, la prosodie est restée au niveau infralexical.

Phonique (voir les textes de BELLISSANT, BERGER-VACHON, DREYFUS-GRAF, GRENIER).

Les travaux effectués à ce niveau sont les moins typiques de la parole naturelle. Les problèmes spécifiques (attaques, fins, plosives en contexte, pauses n'ont pas été traités dans les textes présentés à ces J.E.

La majorité des développements montre que les vieux thèmes ne sont pas épuisés.

GRENIER etc. présente une version "automatiste" d'un phonémographe, à l'opposé de la démarche expérimentale de DREYFUS-GRAF, continuée dans le phonétographe V. On attend avec impatience le jour où ces machines seront intégrées à des systèmes plus complets tel KEAL, comme entrée principale ou auxiliaire.

Comme déjà écrit en 1 les résultats obtenus à ce niveau manquent de références perceptives.

L'expérience de KLATT et STEVENS est la seule, à mon avis, à placer l'être humain dans une situation semblable à celle du reconnaisseur doté comme c'est souvent le cas d'un médiocre codeur, les résultats obtenus que l'on peut rapprocher de ceux fournis par MERCIER (tableau joint) sont comparables.

Il reste que l'on peut toujours espérer des progrès à ce niveau. Ils sont lents mais continus depuis trente à quarante ans. Les progrès décisifs ne se feront pas indépendemment des autres niveaux.

3.Tendances actuelles.

Les textes remis n'étant pas jugés représentatifs des travaux actuels, j'ai demandé aux équipes participantes de répondre rapidement en fin de séance (à raison d'une réponse par laboratoire) aux 3 questions générales suivantes :

- Quels sont les thèmes de recherche importants ?
- Quel est l'interêt des travaux francophones ?
- Pouvez-vous suggérer des thèmes de recherches interdisciplinaires ?

Il n'a pas été possible de classer les thèmes par ordre d'importance. Cela nécessiterait sans doute l'envoi d'un nouveau questionnaire détaillé conçu à partir du tableau.

Les thèmes cités une seule fois n'ont pas été porté sauf exception (entre parenthèses).

On peut noter sur l'ensemble du tableau une assez bonne coïncidence entre les thèmes importants et les thèmes interdisciplinaires (avec parfois des noms différents entre 1 et 3) l'interêt pour la phonétique et la phonologie est évident.

Les études sur le dialogue homme- machine ne suggèrent pas de thème commun à l'intérieur du groupe mais à l'extérieur

exemple : psychocommunication -(sic)

intelligence artificielle.

L'interêt pour les études perceptives parait mitigé. Cette attitude est nettement démarquée de l'attitude des "synthétiseurs".

Une modestie sincère se dégage des réponses à la deuxième question. Il est heureux de constater qu'une équipe n'a pas oublié qu'une des caractéristiques des travaux était de traiter le français.

4

Il n'y a rien à conclure après ces J.E.. Comme je l'ai répété les travaux présentés sont fragmentaires. La multiplicité des thèmes suggérés traduit la jeunes du domaine.

Il serait vain de vouloir en éliminer aucun, mais il serait superflu de les appuyer tous. 2 à 3 poles seraient bien suffisants pour les moyens disponibles en francophonie.

Quant à la parole continue ou naturelle on en est encore au niveau des intentions sauf à Lannion et à Nancy. Le retard estimé sur les recherches avancées dans ce domaine n'a pas diminué. Mais au fond que représentent 2 à 3 ans par rapport à quelques décennies. Il faudra que le groupe, dont il n'existe pas d'équivalent à l'étranger, et des groupes semblables issus du croisement avec d'autres disciplines vivent jusqu'en 2000 et peut-être au delà.

Est-ce possible ?

TABLEAU I : Détection automatique des phonèmes
(tableau fourni par G. MERCIER)

NOMBRE DE PHONEMES	757	I00 %	539	I00 %
CORRECTEMENT IDENTIFIE A	305	38 %	235	43.6%
DETECTION PARTIELLE (NON IDENTIFIE B EN PREMIERE POSITION)	278	40.3%	20I	37 %
DETECTION TOTALE A + B	583	78.3%	436	80 %
NON RECONNU	I50	I9.8%	74	I3.7%
OMISSION	26	3.4%	25	4.6%
INSERTION	34	4.5%	74	I3.7%
APPLICATION	C.D.R		C.A.O	

NOMBRE DE LOCUTEURS : 3

NOMBRE DE PHRASES PAR LOCUTEUR I3 + IO = 23

I. Thèmes importants

Organisation des données et traitements (couplages entre niveaux)

Phonétique : paramètres, détection des traits (simple, temps réel, etc.)
syllabes
prosodie

Linguistique : règles phonologiques
analyses sémantique, syntaxique (formalistation, definition opérationnelle)

Dialogue

Reconnaissance multilocuteur

(modélisation phonation
audition, perception)

2. Thèmes originaux (vus par les chercheurs eux-mêmes)

Dialogue homme-machine

Phonétique

Adaptation et simplification des analyses lexicale et syntaxique.

(français)

3. Thèmes interdisciplinaires

Modèles articulatoires, auditifs

(physiologie)

Perception, psychoacoustique

Phonétique : paramètres, temps réel
prosodie

Règles phonologiques

Intelligence artificielle

Tableau 2 : réponses à "l'interro" fournies par les équipes ayant assisté à la séance sur la reconnaissance vocale.

RECONNAISSANCE DE LA PAROLE CONTINUE

DISCUSSION GENERALE

Remarque de M. J.P. HATON :

Je suis d'accord avec l'introduction d'un certain ordonnancement dans le modèle de système de reconnaissance, par rapport au "blackboard model". Une raison supplémentaire pour introduire un ordre me semble être le fait que les relations entre les différents processeurs sont très variables suivant ces processeurs. La communication que nous présentons avec J.-M. PIERREL est d'ailleurs consacrée à l'étude de ces relations interprocesseurs.

Commentaire de M. J.P. HATON (suite à la communication de M. QUINTON, vol. 1, p. 89) :

Je voudrais remercier M. QUINTON pour la conclusion de son exposé, concernant l'importance des niveaux de traitement en amont du niveau syntaxico-sémantique. Je suis également convaincu de celà et je pense que cette conclusion réjouira un certain nombre de participants à nos Journées d'Etude.

Commentaire de M. J.S. LIENARD (suite à la communication de M. DREYFUS-GRAF , Vol. 1, p. 55) :

Je suis d'accord avec M. DREYFUS-GRAF en ce qui concerne l'importance du niveau acoustique-phonétique et, à ce niveau, en ce qui concerne la nécessité d'une analyse temporelle très fine.

M. MERCIER à MM. MICLET et GRENIER :

Avez-vous une idée assez précise de la complexité des transitions entre phonèmes ?

Réponse de M. MICLET :

"Précise", non. Mais, au moins dans le corpus examiné pour le moment, une transition entre phonèmes est bien marquée par une chute de confiance du premier, corrélativement avec un gain pour le suivant. La modélisation plus précise (automates finis) est l'objet du travail d'apprentissage que nous menons actuellement. La qualité des données issues à cadence fine (10ms) du vocoder à canaux adaptés nous fait espérer de bons résultats pour le décodage et la segmentation en phonèmes.

M. VIVES à M. MICLET :

Il sera très intéressant d'avoir une évaluation des performances de la méthode d'accès aléatoire aux mots du dictionnaire. Ne peut-on pas dire tout de suite que les "points d'ancrage" risquent d'être beaucoup moins efficaces dans une application utilisant la parole continue que dans celles qui ne nécessitent que la reconnaissance de mots isolés ?

Réponse de M. MICLET :

La méthode d'accès décrite est plus une organisation informatique qu'un algorithme proprement dit ; il s'agit de représenter le lexique en machine de façon à permettre un accès par un phonème quelconque et son contexte (point d'ouvrage), et ceci de façon directe, sans comparaison. Le problème de l'utilisation de cette technique (nombre de points d'ancrage, chevauchements, etc...) est un peu différent, et bien sûr essentiel. D'une façon plus générale se pose le problème de l'organisation informatique des données en machine, et de leur accès par plusieurs modules, de façon différentes. Je pense qu'il y a beaucoup à faire dans ce domaine, si l'on veut avoir assez de souplesse dans l'utilisation des différents modules d'un système de reconnaissance.

M. MICLET à M. VIVES :

Les problèmes d'organisation du lexique sous forme informatique sont essentiels pour permettre d'éviter une croissance exponentielle du nombre de résultats et du temps de calcul avec les erreurs de la chaine phonétique. Nous essayons à l'E.N.S.T. de privilégier une organisation à accès aléatoire, où les mots peuvent être retirés à partir des points forts de la reconnaissance ("points d'ancrage"), ceci quelle que soit la place de ce phonème dans le mot. Nous voulons aussi, dans cette organisation privilégier les mots longs, où les contraintes sont les plus fortes. Je pense qu'un débat pourrait être organisé spécialement sur le problème de la recherche lexicale, et que la confrontation des points de vue (et des données) serait très utile.

Réponse de M. VIVES :

Entièrement d'accord pour discuter du problème de la recherche lexicale.

M. QUINTON à MM. PIERREL - HATON :

1) Comment ont été simulées les chaines phonétiques sur lesquelles a été faite la reconnaissance ?

2) Comment ont été extraites les formes phonétiques représentant les mots du lexique ?

Réponses de M. PIERREL :

1) Pour ce qui est de la simulation des chaines phonétiques sur lesquelles a été testée la reconnaissance, elle a été effectuée à partir des résultats obtenus en 1973, par J.-P. HATON, et plus spécialement à l'aide de la matrice de confusion qu'il avait obtenu à cette époque. Actuellement, nous obtenons les premières chaines de phonèmes non simulées et elles permettent aussi au système de fournir de bons résultats.

2) Quant aux formes phonétiques représentant les mots du lexique, elles ont été extraites à la main, compte-tenu des représentations phonémiques classiques, des performances du système acoustico-phonétique de J.-P. HATON (datant de 1973) et de la matrice de confusion de phonème citée plus haut... Nous envisageons aussi de modifier ces formes phonétiques par apprentissage.

M. MICLET à MM. PIERREL - HATON :

Quelles "retombées" sur l'analyse syntaxique pour le traitement de la parole peut-on, à ton avis, attendre des développements de la reconnaissance structurelle des images ?

Réponse de M. PIERREL :

Actuellement, à Nancy, un algorithme d'analyse syntaxique pour la reconnaissance structurale des images est mis en oeuvre. Cet algorithme me permet de démarrer la reconnaissance à n'importe quel point en effectuant ensuite une analyse ascendante "du milieu vers les côtés" (cet algorithme fut présenté à San Diego en novembre 1976, par J.-P. HATON et R. MOHR.

Parallèlement à cette mise en oeuvre pour les images, nous travaillons à un algorithme du même type pour la reconnaissance de la parole. Un tel algorithme se fonde au départ sur des ilots de confiances (ou mots clés), préalablement bien reconnu.

M. PERENNOU à MMe VAISSIERE :

Les résultats que vous indiquez vont tout à fait dans le sens que nous souhaitons pour l'analyse lexicale.

Il serait donc ainsi possible de donner une appréciation de chaque découpage possible (lexicalement) ?

(Nous avons également abordé ce problème, mais bien plus modestement, au cours des 6ièmes Journées de Toulouse).

Réponse de MMe VAISSIERE :

Oui, c'est possible. Il faut que la segmentation en mots proposée par le programme de reconnaissance soit compatible avec les traits prosodiques. Il ne faut cependant pas perdre de vue que cette appréciation sera plus efficace si le signal comporte un grand nombre d'informations prosodiques. De meilleurs résultats sont à attendre avec les locuteurs dont vous trouvez subjectivement que la voix est sûre, bien marquée, claire qu'avec ceux qui donnent l'impression de ne pas ouvrir la bouche pour parler ou de parler "d'un seul trait".

Nous avons noté la forte tendance de tous nos locuteurs (non sélectionnés) à faire largement usage de la possibilité qu'ils ont de faire varier le fondamental au cours de la phrase et nous avons voulu démontrer qu'ils le font selon certaines règles très précises, tout au moins lorsqu'ils lisent des textes, et que, en conséquence, ces variations peuvent être utilisées comme aide à la reconnaissance. Cependant, une phrase prononcée sur un ton monotone (par un humain ou une machine) peut à la rigueur être comprise (donc segmentée en mots) par un auditeur, grâce à la syntaxe et au sens possible de cette phrase. En conséquence, la segmentation de la phrase n'est pas seulement assurée par les variations du fondamental (on rejoint le problème de la redondance dans la parole), et elle sera plus ou moins indiquée par celles-ci selon les locuteurs et le rythme avec lequel ils parlent. Par exemple, en ce moment je compte beaucoup sur la syntaxe et la sémantique pour vous aider à segmenter mes flots de parole, vu que pour respecter les horaires, j'ai accéléré mon débit qui est déjà naturellement assez rapide et je n'ai pas le temps de bien marquer mes débuts et fins de mots prosodiques.

M. RODET à MMe VAISSIERE :

Est-ce que les variations de durée des phonèmes suivent aussi le découpage en mots littéraux, ou seulement le découpage en mots prosodiques ?

Réponse de MMe VAISSIERE :

Il semble que les variations de durée suivent plutôt le découpage en mots prosodiques et on peut constater une élongation des voyelles et des consonnes durant lesquelles il se produit un mouvement du fondamental, en particulier une montée. Quand deux mots littéraux sont groupés pour ne former qu'un seul mot prosodique, l'élongation sur la dernière syllabe du premier mot est moins importante que dans le cas où les deux mots sont indépendants et où le premier mot correspond à lui tout seul à un mot prosodique. Nous manquons de données précises sur les variations de durée des voyelles et des consonnes françaises intégrées dans les phrases et c'est là un champ d'étude très intéressant : nos études sur la voix d'un locuteur professionnel nous ont montré que la longueur relative des syllabes, consonnes et voyelles en séquence est également porteuse d'informations concernant la segmentation, mais le problème semble plus compliqué que dans le cas du fondamental, dû à une importance plus grande de la durée intrinsèque des phonèmes et de l'influence des phonèmes les uns sur les autres.

M. DREYFUS-GRAF à MMe VAISSIERE :

L'analyse prosodique de l'anglais donne-t-elle des résultats très différents de celle du français, puisque le français accentue la dernière syllabe, contrairement à l'anglais ?

De même : "Comment se présente l'analyse prosodique du français féttéral" ?

Réponse de MMe VAISSIERE :

Contrairement à ce que peuvent affirmer certains livres, nous ne croyons pas à un accent fixe sur la dernière syllabe des mots en français contemporain : c'est justement pour cette raison que nous décrivons la prosodie comme une organisation de la fréquence du fondamental, de la durée et de l'intensité au niveau des groupes et des mots prosodiques et nos patterns doivent être considérés comme des touts indivisibles. Toutes les syllabes jouent un rôle également important dans un pattern, ne serait-ce pour certaines qu'un rôle de bouche-trou, de tremplin pour aller d'une hauteur à une autre, ou encore il y en a qui s'écrasent pour mieux mettre en valeur la voisine de droite ou de gauche, et ce rôle n'est pas négligeable puisque sans elles le pattern ne serait pas possible. Il y a bien quelques français qui trainent un peu sur la dernière syllabe de leurs mots, mais même dans ce cas peut-on parler d'accent ? (Il s'agit en général de français appartenant à la catégorie P2).

Du point de vue auditif, on ne peut pas comparer l'accent anglais (principal et secondaire) à ce qui se passe en français. Une description de la fréquence du fondamental utilisant approximativement le même système d'attributs intonatifs a été réalisée pour le hollandais, par Cohen et t'Hart, et pour l'anglais par Maeda, et il semblerait que chaque langue utilise les mêmes moyens pour les mêmes fins (démarcation et indicateur syntaxique), mais chacune a sa propre organisation. Sans doute que le français fédéral a apporté quelques variantes originales au français parisien, variantes que nous ignorons puisque nous n'avons jamais eu de locuteur suisse.

M. MICLET à MMe VAISSIERE :

Trouve-t-on parfois des marqueurs de segmentation en milieu de mots, et si oui, peut-on les éliminer algorithmiquement ?

Réponse de MMe VAISSIERE :

Oui, on peut trouver des marqueurs de segmentation (montée rapide ou descente rapide du fondamental) au milieu des mots. Nous avons vu qu'une montée peut à la fois marquer le début d'un mot prosodique (attribut Ri), ou encore la dernière syllabe d'un mot (attribut Rc dans P1 et Rp dans P2). La prise en considération du mouvement du fondamental conséquent à cette montée permet de lever l'ambiguité sur son interprétation (Ri, Rc ou Rp) dans de nombreux cas (voir texte) : si la montée est suivie d'une partie stationnaire (attribut S), ou légèrement descendante (attribut F), elle est considérée comme marquant le début d'un mot ; si elle est suivie d'une chute rapide (attribut L), elle est automatiquement considérée comme une montée sur la dernière syllabe d'un mot :

montée brusque + descente brusque	=	Rp + L (P2)	(1)
(Données acoustiques)		(Interprétation)	

En fait, dans les mots prosodiques courts, on peut observer une neutralisation (due à des contraintes temporelles) entre les mouvements L, F, (S+L) et, en conséquence, entre les patterns P2, P3 et P4 : ces trois patterns sont réalisés par une montée brusque suivie d'une descente brusque, qui sont interprétées par le programme comme (Rp+L) (1)

Données acoustiques :	Montée brusque	+	descente brusque	
Interprétation en attributs	P2 : (Ri + (S) + Rp)	+	L	
	P3 : (Ri + (S))	+	L	
	P4 : Ri	+	F	(2)
Interprétation par le programme	Rp	+	L	(P2)

En conséquence, un mot prosodique court, dont les frontières prosodiques auront pourtant été bien marquées par le locuteur, sera considéré par le programme comme la dernière syllabe d'un mot prosodique avec le pattern P2. Une référence au nombre de syllabes sousjacentes à la montée et à la descente permettrait d'améliorer les scores présentés dans cette communication :

montée brusque + descente brusque sur une syllabe seulement (données acoustiques) = dernière syllabe d'un mot ou monosyllabique (interprétation) (3)

montée brusque + descente brusque sur plus d'une syllabe (données acoustiques) = mot prosodique court (interprétation).

La règle (3) n'a pas été introduite à la place de la règle (1) parce qu'elle sous-entend que les sommets vocaliques ont été préalablement détectés et dénombrés, ce qui n'est pas prévu par le programme actuel. Nous voyons donc qu'il est possible dans une certaine mesure de sélectionner parmi les mouvements du fondamental ceux qui correspondent effectivement à des marqueurs de segmentation et les autres.

M. HATON à M. PERENNOU :

1) Pouvez-vous préciser le type de la correction de F2 par F3 dans la reconnaissance de voyelles ?

2) Ne pensez-vous pas que les transformations dues aux erreurs de reconnaissance phonémique vont donner une explosion du nombre possible de versions pour une phrase, particulièrement importante avec votre approche ?

Réponses de M. PERENNOU :

1) La correction du formant F2 par F3 permet des améliorations considérables dans la distinction /i/y/.

Une légère amélioration de la distinction /u/o/ peut être aussi obtenue par cette voie.

Voir figures 1 et 2 ci-jointes.

Il est à noter qu'une bonne mesure de F1 et de F2 apporte déjà de bons résultats. Les figures 1 et 2 montrent aussi qu'une erreur supplémentaire de 40 hz dégraderait considérablement les résultats.

Pour l'examen détaillé de nos méthodes, nous renvoyons à notre rapport de contrat SESORI IV/137 n° 74-94, dont la parution est prévue pour septembre 1976.

2) La méthode que nous avons choisie suppose évidemment un filtrage des versions de découpage. Une première mesure consiste à travailler sur des intervalles de parole aussi courts que possible.

Si des indices prosodiques fournissent des délimitations objectives, il faut les prendre. Sinon, il faut statuer sur l'ensemble des versions au bout de quelques syllabes (disons 5 environ).

Le rôle des filtres syntaxiques est d'écarter les suites aberrantes de termes fonctionnels (puisque l'élision peut être postulée, beaucoup de mots courts sont en effet générés).

Ex. : /i la mõ te/ ne peut s'interpréter par :

y la montée (y la est une suite impossible, sauf s'il y a une pose entre y et la : vas-y : la ...)

M. ALINA à M. PERENNOU :

Au sujet de la syllabation :

Vous avez parlé du fait que la suite des syllabes peut comporter des éléments plus ou moins erronnés. C'est vrai. Mais, envisagez-vous la possibilité d'attribuer un degré d'accentuation à chaque syllabe ?

Réponse de M. PERENNOU :

Votre question souligne un point primordial pour la reconnaissance de la parole continue. A ce sujet, l'intéressant exposé de Rossi M., aux 6ièmes Journées d'Etude sur la Parole, suggère des voies de recherches intéressantes.

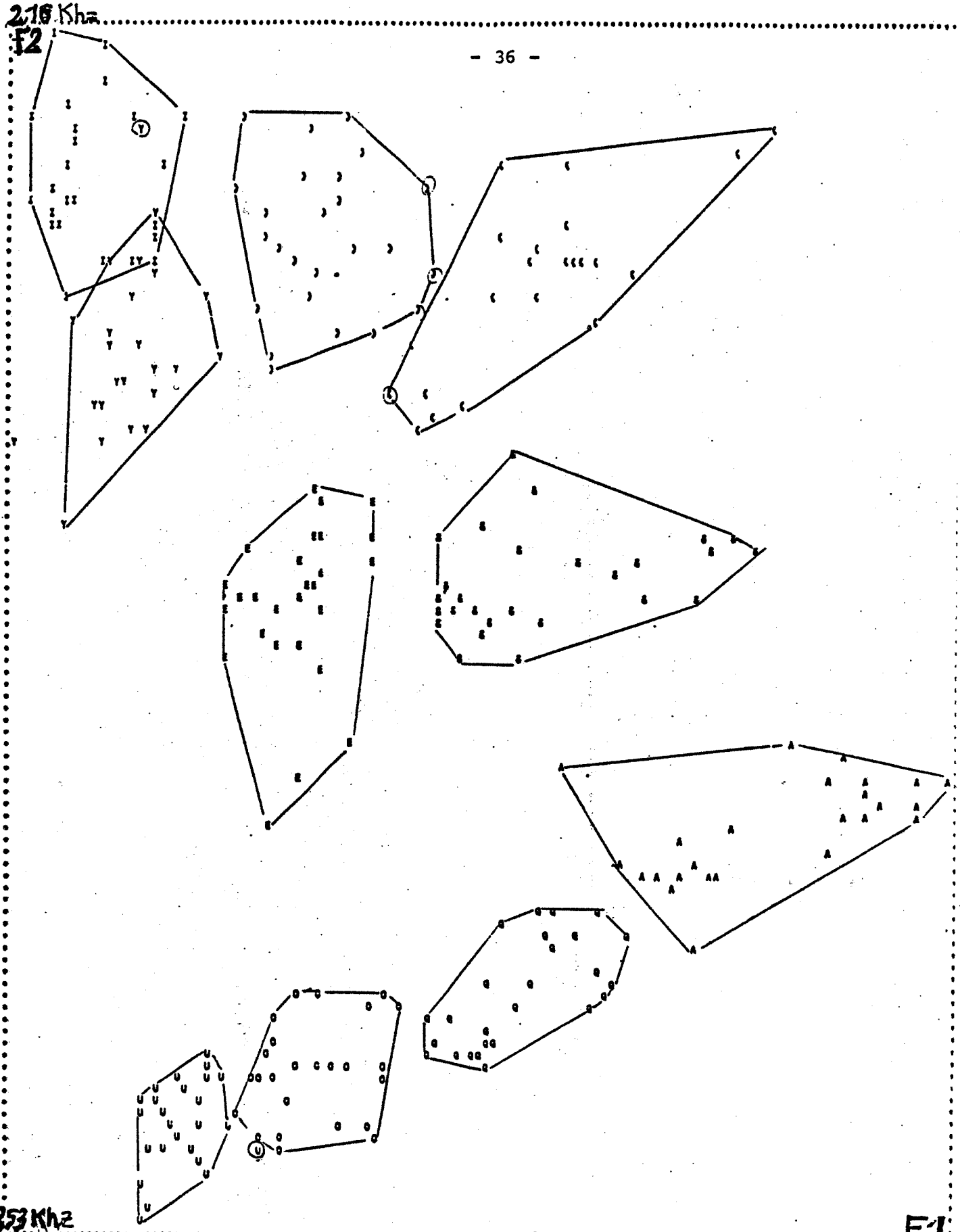


Fig. 1. REPRESENTATION D'UN ECHANTILLON DE VOYELLES DANS LE PLAN F1. F2

$F'2 = \Delta_2 + F2$, où :

$\Delta_2 = 2. [F3 - B]_+ [0,45 - F1]_+$ avec :

$B = 2,4 - 0,4 (F1 - 0,2)$, si $F1 \leq 0,45$,

$B = 2,3$, si $F1 \geq 0,45$.

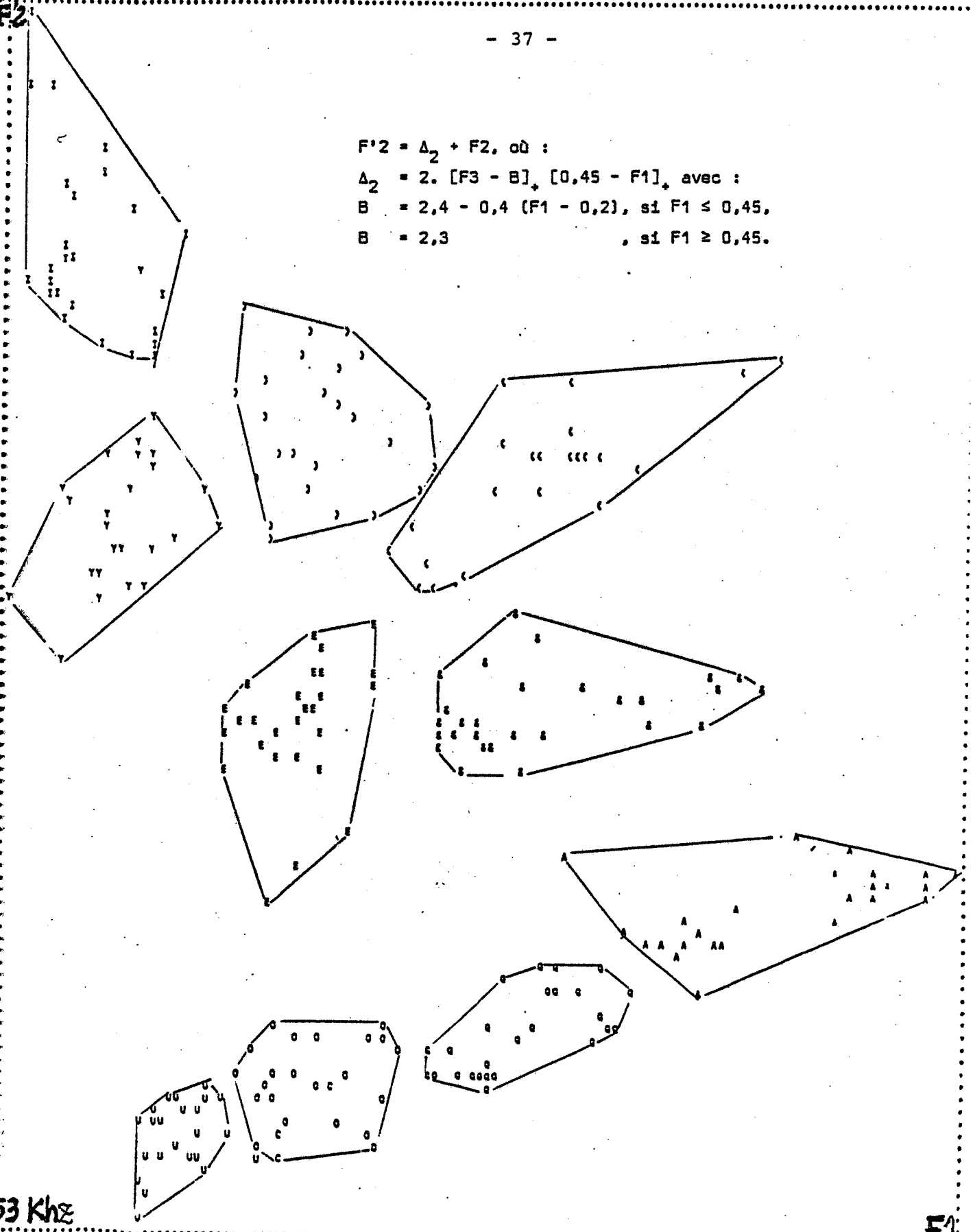


Fig. 2. REPRESENTATION D'UN ECHANTILLON DE VOYELLES DANS LE PLAN F1 - F2

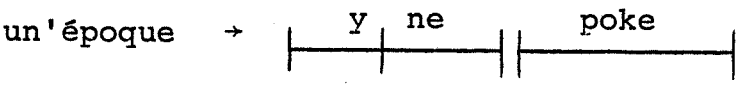
Pour notre part, nous observons que certaines syllabes "sous-accentuées" poseront des problèmes de reconnaissance phonémique.

Il est probable que la sous-accentuation, alliée parfois au contraste mélodique, contient en elle-même des informations qu'il faut utiliser :

Ex. : la première d'une suite de syllabes brèves est peut-être un déterminant ou, plus généralement, un terme fonctionnel.

Une syllabe brève est peut être la première d'un mot si cette première syllabe est faible.

Ces deux points se retrouvent dans l'exemple suivant où la longueur de la syllabe est figurée par l'intervalle qui la contient :



M. MICLET à M. PERENNOU :

1) Quelle est la "stabilité" d'un tel système à des chaines phonétiques erronnées, ou composées d'hypothèses multiples ?

2) La syntaxe peut-elle être représentée simplement par des contraintes d'enchainements entre catégories grammaticales, ou ne faut-il pas utiliser des techniques d'analyse syntaxique, et alors piloter la recherche lexicale par la syntaxe ?

(Cf. Actes des Journées de Bruxelles, Communication Gresser et al.).

Réponse de M. PERENNOU :

1) Voir réponse à Monsieur Haton.

2) Nous ne disons pas que les contraintes d'enchainement suffisent pour décrire la syntaxe du français.

Ces contraintes permettent simplement de détecter des suites lexicales aberrantes par des procédés simples.

Nous ne pensons pas que la recherche lexicale doit être gouvernée par la syntaxe globale, mais qu'il doit y avoir des modules autonomes pour le lexique, la syntaxe, la sémantique, la pragmatique (cf. Ready). Ces modules entretiennent entre eux des échanges.

Le module d'analyse lexicale fournit des suites de mots. Pour les produire, il a ses méthodes propres (qui peuvent du reste s'exprimer par une syntaxe).

En retour, les autres modules peuvent contribuer au filtrage en indiquant que telle suite est mal construite syntaxiquement (syntaxe d'ensemble de la phrase) ou est aberrante sémantiquement.

Ex. :

/i la mõ te/ → y la (montée) : "impossible au niveau du module d'analyse lexicale",

/i la mõ te/ → ile a mon té : "improbable au niveau sémantique",

/i la mõ te/ → il amont ... : "rejet au niveau syntaxique".

THEME 2 :

SYNTHESE DE LA PAROLE

LA SYNTHESE DE LA PAROLE

Rapport de Présentation

R. CARRE

Laboratoire* de la Communication Parlée

E.N.S.E.R.G. 23, rue des Martyrs GRENOBLE

INTRODUCTION

L'objet de cette présentation sur la synthèse de la parole est de faire le point dans ce domaine de recherche, de situer les différents travaux dans une perspective générale.

Nous essayerons de voir la progression des recherches, en particulier depuis les premières journées d'études de Grenoble où un exposé sur la synthèse avait été donné.

Rappelons aussi qu'une rencontre avait été organisée à l'IRIA en 1973 pour étudier les applications possibles des synthétiseurs.

Ici, sans négliger l'aspect application, j'aimerai plutôt m'arrêter sur certains problèmes fondamentaux qui ont fait l'objet d'études ces dernières années ou bien qui mériteraient un plus grand développement.

Des résultats intéressants ont été obtenus qui conduisent à des applications. Mais, il serait très dangereux de laisser croire qu'il n'y a plus ou presque plus de problèmes dans ce domaine de recherche.

Ce serait un danger pour la recherche fondamentale en ce qui concerne l'analyse acoustique du signal de parole car une amélioration de la synthèse passe par des travaux d'analyse particuliers, travaux qui, par ailleurs, faciliteront la reconnaissance.

Ce serait un danger vis à vis des organismes financiers qui ont effectivement tendance à penser que les problèmes en synthèse sont réglés et qui ne tiennent plus à financer des travaux en synthèse.

Or, on assiste actuellement à un grand développement des travaux dans le domaine de la reconnaissance de la parole sans doute au détriment de la recherche en synthèse. C'est vrai en particulier aux U.S.A.

J'ai voulu citer dans cet exposé certains travaux Américains présentés à la rencontre de l'A.S.A. à Washington et à celle des I.E.E.E. à Philadelphie. Or, il y a peu de travaux. Dans le rapport de 800 pages de la rencontre de IEEE, 40 à 50 pages sont consacrées à la synthèse.

* Equipe Associée au C.N.R.S.

Et pourtant, FANT signalait au Symposium de PITTSBURG sur la reconnaissance, qu'il manquait encore beaucoup de données sur le signal acoustique, sur ses caractéristiques physiques déduites du fonctionnement de l'appareil vocal. En effectuant des travaux de synthèse, on peut améliorer l'analyse en comparant l'analyse du signal original et celle du signal de synthèse.

Rappelons aussi que les meilleurs résultats en reconnaissance de parole ont été obtenus au Lincoln lab. par une équipe ayant au préalable longuement travaillé en analyse-synthèse.

Au cours de cette présentation, nous évoquerons certains faits importants qui se sont produits depuis 1970 et nous essayerons de situer certains travaux Français, en particulier ceux qui sont décrits dans le rapport de ces Journées.

Nous essayerons alors de proposer quelques perspectives dans le domaine des recherches et des applications.

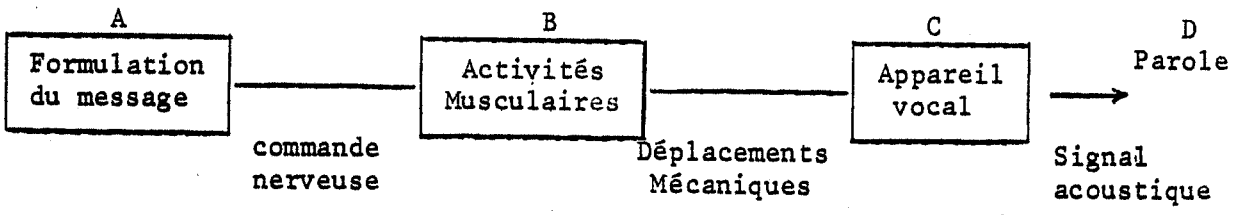
Après discussion générale et pause, chacun des auteurs des communications sur la synthèse aura 10 minutes pour présenter certains points de son rapport. Mais, nous aimerions aussi qu'il situe très brièvement son travail dans le cadre du programme de recherche de son laboratoire avec les perspectives envisagées.

A - *LE DEVELOPPEMENT DES SYNTHETISEURS DE PAROLE*

Le développement des travaux sur la synthèse de la parole est très lié au développement des connaissances sur le fonctionnement de notre appareil vocal.

Cette liaison existait assez peu au moment de la conception des vocoders à canaux. Puis, des améliorations de l'intelligibilité et de la qualité de la parole de synthèse se sont faites progressivement à partir d'une connaissance de plus en plus fine des mécanismes de production de la parole, les améliorations se faisant à deux niveaux différents : au niveau du synthétiseur et au niveau de la commande.

On peut représenter le processus de production de la parole selon le schéma suivant :



On a d'assez bonnes connaissances en D. En particulier, l'analyse spectrale du signal de parole nous a révélé de très nombreuses caractéristiques de ce signal.

Le lien appareil vocal - signal acoustique commence à être assez bien connu. De récents travaux, ceux de FANT et de MRAYATI par exemple, sur la contribution des différents types de pertes à l'intérieur du conduit vocal, sur la sensibilité de petites modifications en un point du conduit vocal selon la configuration générale de ce conduit, sur le couplage conduit vocal- conduit nasal, conduit vocal - source vocale sont très précieux et peuvent entraîner des améliorations au niveau du synthétiseur.

On peut entrevoir à l'horizon des prochaines années un terme à ces travaux.

Le lien B-C est plus difficile à préciser. On commence seulement à proposer des modèles articulatoires qui sont déduits à partir de résultats cinéradiographiques et spectraux.

Rappelons simplement les travaux principaux, ceux de COKER, MERMELSTEIN et ceux de LINDBLOM qui ont été présentés à LANNION en 1972.

Les modèles articulatoires élaborés devraient correspondre au niveau B et naturellement on cherche à aller au delà de ce niveau et à étudier les commandes nerveuses qui devraient correspondre aux commandes d'un modèle articulatoire.

Citons ici les travaux de OHMAN.

De tels modèles sont aussi étudiés pour représenter le fonctionnement de la source vocale et la génération automatique de la source de bruit.

Bien que limitées, les mesures électromyographiques et neurophysiologiques paraissent très prometteuses.

Au niveau A, on ne connait rien de manière sûre.

Certains pensent toujours à une liaison production-perception de la parole. Aucune réponse définitive ne peut être proposée sur ce point.

Maintenant, sur le schéma, peut-on dire à quel niveau est effectué le passage signaux discrets, signaux continus ?

En B, C et D les signaux semblent continus ou bien on ne connait pas les règles complexes de discrétisation.

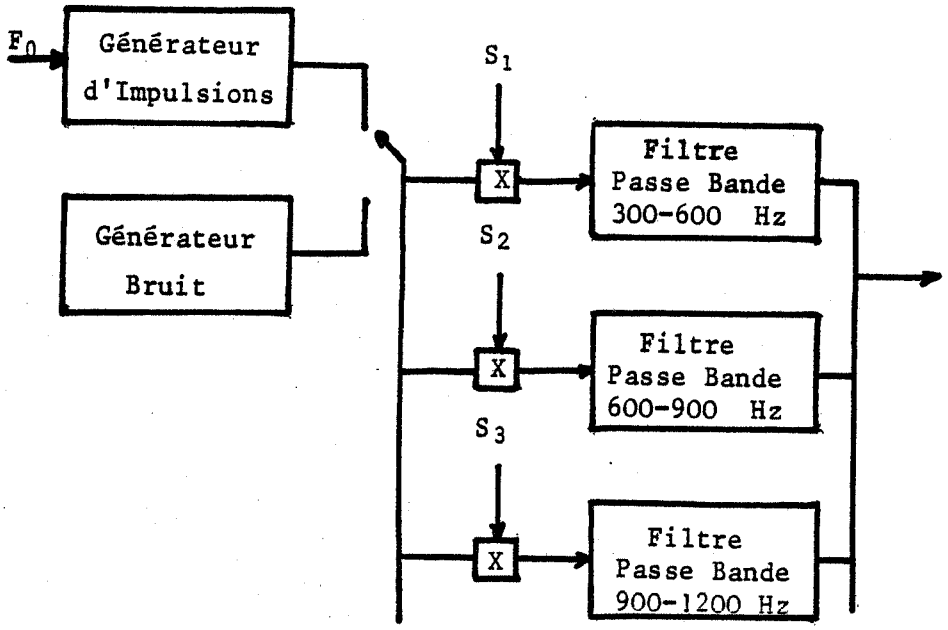
Une telle connaissance de l'endroit de passage discret-continu permettrait d'étudier de manière sérieuse les problèmes de segmentation.

Ces problèmes sont brûlants et concernent d'ailleurs aussi bien la synthèse que la reconnaissance.

Maintenant, nous allons passer en revue les perfectionnements apportés aux synthétiseurs et aux commandes de ces appareils.

B - LES SYNTHETISEURS

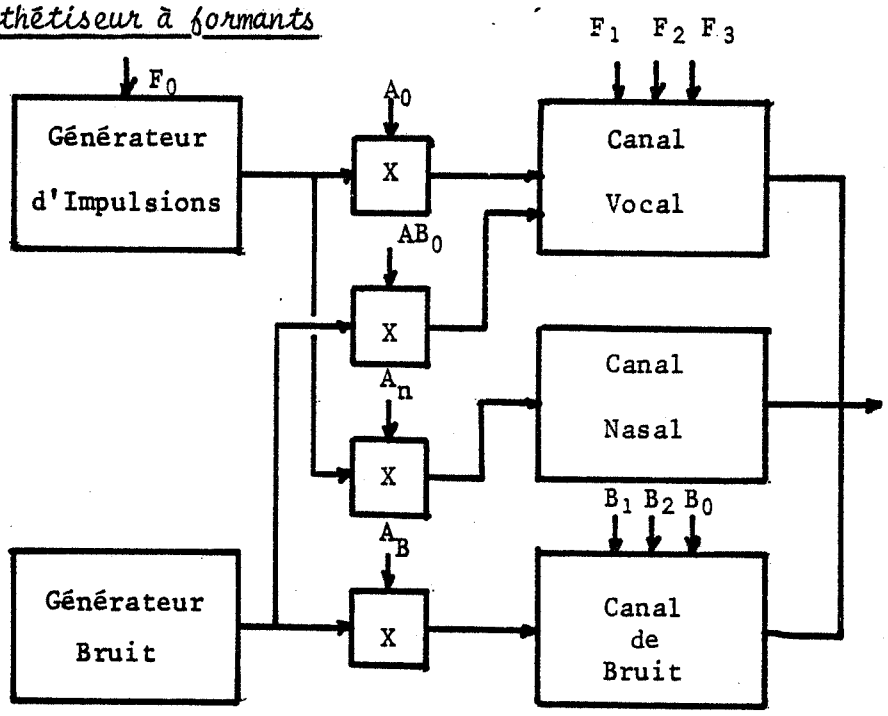
Le vocoder à canaux



Les améliorations concernant les vocoders à canaux, vocoders à bande de base etc... sont pour l'essentiel de caractère technologiques. Ils sont plus sûrs, plus stables.

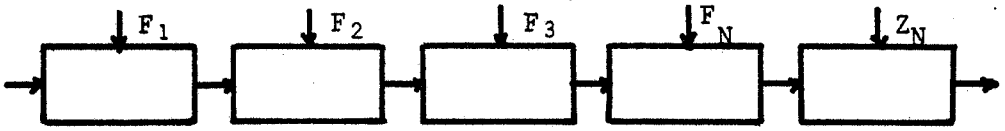
On classe l'icophone dans la catégorie des vocoders à canaux.

Le synthétiseur à formants



Plusieurs améliorations ont été apportées, tout au moins en simulation pour rendre compte de certains phénomènes.

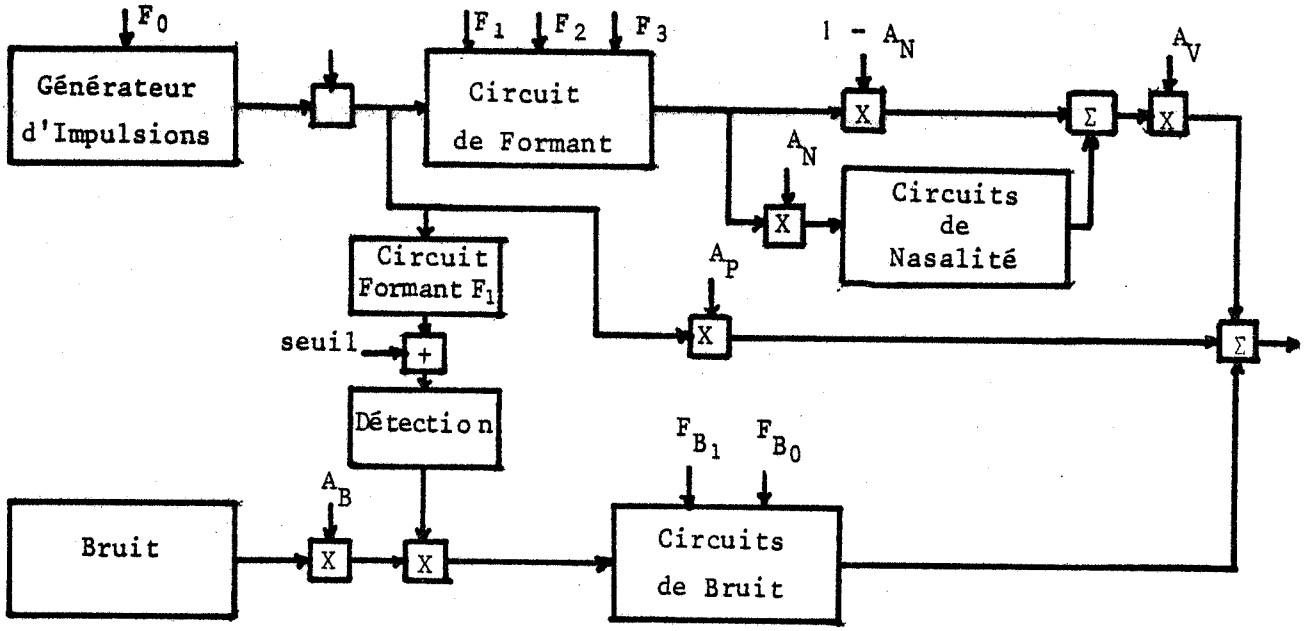
Citons simplement le synthétiseur de KLATT qui permet d'effectuer plus simplement des synthèses de consonnes et de voyelles nasales par inclusion dans la chaine des formants d'une paire de pôle-zéro.



Pour les voyelles orales $F_N = Z_N$.

En fait, il semble que de meilleurs résultats puissent être obtenus avec 2 paires de pôle et zéro, en particulier pour faire la synthèse des voyelles nasales françaises.

RABINER a proposé une structure intéressante pour la synthèse des occlusives sonores et les fricatives voisées.



La commande par ordinateur de ce type d'appareil a permis de développer considérablement la qualité et la quantité des travaux sur la synthèse des sons.

Ces systèmes sont, soit simulés sur ordinateur, soit réalisés selon des techniques analogiques ou digitales. Citons ici la réalisation à fonctionnement numérique de RABINER et celle de la CGE à Marcoussis.

Le système siphon très original réalisé par la THOMSON C.S.F. s'apparente à un système à formants.

La synthèse est effectuée à partir de 3 sinusoïdes de fréquence correspondant aux 3 premiers formants. Une remise en phase des signaux est effectuée à la fréquence fondamentale ce qui permet d'introduire l'effet de mélodie.

C'est un appareil simple, commandé avec le minimum de paramètres.

Associé à un analyseur particulier, il délivre une parole intelligible quoique peu naturelle.

Est-ce qu'un synthétiseur à formants de type parallèle n'aurait pas donner de meilleures performances sur le plan de la qualité ?

Analogues du conduit vocal

Les synthétiseurs analogues du conduit vocal ont fait l'objet de développements importants en France.

Beaucoup espèrent obtenir une meilleure qualité avec ce type d'appareil et une synthèse par règles plus facile à effectuer.

On pense pouvoir simuler plus facilement certains phénomènes qui n'ont pas encore été étudié avec un synthétiseur à formants.

Certaines caractéristiques dites intrinsèques devraient pouvoir être automatiquement reproduites.

Dans des modèles statiques, on commence à tenir compte de tous les types de pertes : par viscosité, par chaleur, par vibrations des parois et par rayonnement.

Les bandes passantes des formants sont ainsi correctement respectées (MRAYATI).

Il y a plusieurs méthodes de simulation, certaines partent directement des équations de propagation, d'autres correspondent à une analogie acoustico-électrique.

Les simulations des sources sont proches des phénomènes réels et les paramètres de commande sont la pression subglottique et la tension des cordes vocales. Le modèle de fonctionnement à 2 masses des cordes vocales proposé par ISHISAKA et FLANAGAN parait être très proche de la réalité.

La source de bruit peut être simulée automatiquement en tenant compte des sections des constrictions et de la différence de pression au niveau de ces constrictions.

L'association judicieuse d'un modèle de conduit vocal et d'un modèle de source vocale peut rendre compte automatiquement des couplages acoustiques source-conduit.

Des réalisations, à fonctionnement dynamique, sont en cours de développement au CNET - LANNION, à l'ENSERG Grenoble et au Laboratoire d'Automatique de BESANCON.

Pour ces journées, 3 rapports nous ont été adressés concernant la synthèse à l'aide d'un modèle du conduit vocal.

Un premier rapport, celui du CNET décrit une réalisation complète simulant le modèle de source à 2 masses proposé initialement par FLANAGAN, le conduit vocal avec 12 sections, et la création automatique du bruit. Un programme permet la commande articulatoire de l'ensemble qui commence à balbutier.

Un deuxième rapport, celui de l'ENSERG de Grenoble, décrit les effets de couplage conduit vocal - source vocale sur la forme et le spectre de l'onde glottique et sur la fréquence de vibration des cordes vocales.

Malheureusement, d'après les résultats obtenus, un tel couplage ne peut expliquer les fréquences fondamentales intrinsèques des voyelles. Ces résultats doivent être examiner de près, car ils sont opposés à ceux obtenus par FLANAGAN.

Le rapport du Laboratoire d'Automatique de BESANCON décrit une méthode très originale de simulation à partir des équations de propagation.

Le conduit vocal et les sources ne font alors qu'un tout. Les sources sont assimilées à certaines portions du conduit vocal douées de caractéristiques particulières.

Ainsi la source de bruit pourrait être obtenue par vibrations particulières des parois du conduit vocal. Mais je ne comprends pas comment une source de bruit peut prendre naissance, les parois ayant une fréquence de vibration propre qui n'a rien d'aléatoire.

Synthétiseur à codage prédictif

Le développement des techniques de codage de parole par prédiction linéaire a été considérable ces dernières années et constitue l'un des apports les plus importants pour tout ce qui concerne la parole.

La parole synthétique est intelligible et surtout de bonne qualité.

Le système est intéressant sur le plan de la production de la parole car il cherche à identifier la fonction de transfert de l'appareil vocal produisant le signal à analyser.

En fait, on calcule des coefficients de prédiction qui correspondent à la réponse impulsionnelle d'un modèle inverse de l'appareil vocal.

A partir de ces coefficients, on peut obtenir les fréquences des formants ou bien les coefficients de réflexion au niveau des sections d'un modèle de conduit vocal.

WAKITA et EL MALLAWANY ont étudié particulièrement le calcul de ces coefficients. Disons qu'il est difficile de faire ce calcul en régime dynamique.

Donc, pour l'analyse, le codage prédictif est un moyen très puissant donnant de nombreuses informations qui sont actuellement parmi les plus sûres et les plus rapidement obtenues. Ces informations permettent une commande du synthétiseur qu'il soit à formants, analogue au conduit vocal ou bien qu'il soit un filtre numérique récursif.

Le filtre dont les caractéristiques sont contrôlées par les coefficients de prédiction ou bien par les coefficients de réflexion est de réalisation simple. Il peut ainsi être facilement utilisé comme sortie parlée d'ordinateur.

OLIVE au Bell Labs. vient d'en réaliser un, où le calcul du gain a été particulièrement étudié en tenant compte de sa valeur aux périodes précédentes. Ceci pour éliminer une certaine rugosité à la synthèse.

EL MALLAWANY propose aussi, pour ces journées la description d'un tel appareil. Les conditions optimales de fonctionnement d'un tel synthétiseur ont été bien étudiées. De bons résultats peuvent être obtenus avec un débit de 4800 bits/seconde.

Commande des synthétiseurs, synthèse par règles

On peut commander les synthétiseurs décrits par des paramètres : spectres, formants, fonction d'aire, coefficients de prédiction, etc....

On peut ainsi faire la synthèse de phrases, de mots ou d'élements du type CVC, VCV, CV, VC, C, V.

On a alors naturellement cherché à faire une commande à partir des éléments les plus simples, dont le nombre est le plus réduit, c'est-à-dire les "segments phonétiques" ou "phones".

Malheureusement un "phone" a des caractéristiques acoustiques qui dépendent du contexte.

On a des problèmes de coarticulation, de réduction des voyelles, etc....

On a donc cherché à établir des règles de transition d'un segment à un autre, des règles de transition des formants, des règles de passage d'une configuration articulatoire à une autre.

L'étude de ces règles est complexe mais elle permet de mieux connaître l'aspect évolutif de la parole. Elle permet de préciser des contraintes.

Naturellement, ces études ont été faites au début sur les formants à cause des nombreux travaux effectués sur ces paramètres.

Puis, devant la complexité des règles de transition, on a pensé qu'un modèle articulatoire pourrait rendre plus naturellement les effets désirés dans la mesure où les contraintes articulatoires sont bien simulées.

Un tel modèle est alors commandé par des paramètres correspondant aux commandes nerveuses émises par le cerveau.

En synthèse par règles, citons plus particulièrement les travaux de MATTINGLY et de KLATT avec un synthétiseur à formants.

Les transitions des formants peuvent être calculées à partir d'une table de valeurs en tenant compte des éléments adjacents. La durée des éléments est aussi indiquée dans la table.

LINDBLOM, en particulier, a proposé un modèle articulatoire et COKER a effectué des synthèses par règles à l'aide d'un modèle articulatoire. COKER dispose d'un modèle qui peut représenter toutes les formes du conduit vocal que l'on peut rencontrer au cours de la

production de la parole, mais qui présente des contraintes : continuité de la surface de la langue, courbure du conduit vocal, discontinuités du conduit au niveau des dents, du vélum. Le modèle tient compte aussi de contraintes temporelles déterminant l'évolution possible de la forme du conduit. Le modèle de LINDBLOM paraît plus complet et plus précis mais, il n'a pas, à ma connaissance, été utilisé en synthèse.

On peut éviter le calcul des transitions et décomposer la parole en diphones. Cette solution particulièrement intéressante dans de nombreuses applications, n'ouvre pas, me semble-t-il, de perspectives en ce qui concerne la recherche fondamentale.

Ce procédé développé tout d'abord au Laboratoire de Mécanique de PARIS a été repris par le CNET avec un vocoder et par GRENOBLE avec un synthétiseur à formants.

Disons que cette solution, proposée à l'origine par PETERSON est maintenant typiquement française et intrigue souvent les étrangers.

En parallèle avec ces travaux sur les transitions, des recherches sur la prosodie ont été entreprises et les résultats obtenus améliorent considérablement la qualité de la parole de synthèse.

Trois rapports sur ce thème ont été proposés pour cette rencontre.

Celui du CNET décrit différentes règles utilisées pour le calcul de l'intensité, de la durée et de la fréquence fondamentale. Il n'y a pas d'analyse syntaxique mais on introduit des marques de début et de fin de syntagme. Le modèle proposé est appliqué à un ensemble de synthèse par diphones.

Le travail de Philippe MARTIN est original. Il suppose au départ une mise en mémoire des syntagmes minimaux de la totalité des phrases pouvant être synthétisées. La méthode repose sur le parallélisme des structures mélodique et syntaxique de la phrase.

L'analyse syntaxique est facilitée puisque les marques de présupposition sont repérées lors de la mise en mémoire des syntagmes minimaux.

L'étude présentée par l'Institut de Phonétique de GRENOBLE permet de définir les différents paramètres permettant de générer une phrase interrogative. Cette étude a été complétée par des tests de perception pour préciser les degrés de liberté possibles pour l'évolution de la fréquence fondamentale.

Utilisations des synthétiseurs. Applications

Comme sortie parlée d'ordinateur, le synthétiseur de parole a certainement un bel avenir à cause du téléphone qui restera toujours un moyen très simple de communication ou bien simplement parce que l'ouïe est l'organe sensoriel le plus disponible comme récepteur dans un contexte déterminé.

Mais alors de nombreuses questions se posent ? Quel est le synthétiseur le plus intéressant en tenant compte de son prix, de la qualité de parole obtenue, du nombre de paramètres de commande.

Comment le commander ? Par mots, par phones, par diphones ?

Il semble difficile de formuler des réponses, ces réponses dépendant du type d'application.

Parmi les matériels disponibles, on trouve aussi bien des systèmes sans codage particulier (le synthétiseur est simplement un haut-parleur) pouvant produire jusqu'à 10 000 mots différents et certaines phrases complètes les plus courantes.

Les mots prononcés sont "naturels" évidemment mais les phrases provenant de l'assemblage de ces mots le sont beaucoup moins.

L'utilisateur peut entrer de nouveaux mots mais ils sont alors épelés. C'est le Nucleus 4000 fabriqué par la firme AMERICAN SYSTEM.

On trouve aussi des synthétiseurs commandés par les phones comme celui de KLATT ou bien comme OVE III.

Ou bien on trouve l'icophone commandé par des diphones ce qui permet aussi de fabriquer n'importe quel mot.

Le CNET utilise une sortie parlée d'ordinateur pour donner des renseignements téléphoniques dans le cas de changement de numéro d'un correspondant.

Le LIMSI a proposé une communication concernant une utilisation possible de la sortie parlée dans l'aide à la programmation.

En télécommunications, on retrouve toujours le problème du temps réel au niveau de l'analyse. Mais, lorsqu'un ensemble fonctionne, on s'intéressera alors souvent aux possibilités de codage à caractère secret, à la sécurité de la communication, plutôt qu'à une diminution du nombre d'informations à transmettre qui va de 50 000 à 1000 bits/s.

Par ailleurs, la Bell Telephone, IBM La Gaude, développent des techniques de compression d'informations basées sur l'utilisation de certaines propriétés du signal de parole. Ils arrivent à des débits de 10 à 20 k bits/s.

De son côté, le LINCOLN Lab. a construit un système autonome à codage prédictif fonctionnant en temps réel.

Dans le domaine des recherches sur la parole, il n'est pas question de citer tout ce qui peut être fait mais simplement d'insister sur le processus : analyse puis synthèse pour vérifier la qualité de l'analyse puis analyse à nouveau pour raffiner certaines interprétations.

Ce type de travail peut être fait avec des synthétiseurs à formants ou articulatoires.

On peut utiliser un analogue du conduit vocal pour mieux connaître le fonctionnement de l'appareil vocal, sa fonction de transfert, ses bandes passantes, la distribution de l'énergie le long du conduit, la distribution de la sensibilité à des variations de l'aire des sections, etc...

La synthèse peut être utilisée dans des travaux de perception

Ainsi BOE a étudié l'importance de la variation de la fréquence fondamentale pour les phrases interrogatives.

Mais on utilise surtout les sons de synthèse pour proposer un modèle fonctionnel du système auditif.

A partir de simples tests de perception, par exemple, on peut retrouver les courbes de sélectivité de la cochlée comme l'a fait CHISTOVICH.

PERSPECTIVES

Les travaux sur la synthèse ont encore un bel avenir.

Si la parole synthétique est intelligible on peut améliorer considérablement la qualité. On peut chercher à reproduire certains accents, on peut chercher à reproduire les caractéristiques individuelles.

Pour arriver à ces résultats l'analyse devra être raffinée.

En synthèse avec un analogue du conduit vocal, on peut encore améliorer la simulation des sources, tenir compte de la transmission du son à travers les parois des joues, du volume ou du palais.

Les modèles articulatoires ne sont pas au point. On peut peut-être utiliser la distribution de la sensibilité à des variations de l'aire des sections pour étudier l'aspect "quantique" de la parole avec les zones de stabilité acoustique du conduit vocal.

. Pour faire ces travaux sur le conduit vocal, nous manquons cruellement de données. On peut par des méthodes indirectes obtenir la fonction d'aire du conduit vocal mais de nombreuses données cinéradiographiques seraient les bienvenues. Aucune étude électromyographique importante n'a été entreprise, à ma connaissance en France.
Citons les Japonais qui développent considérablement ce type de travaux.

. Les sons du Français doivent être étudiés systématiquement à partir de diverses réalisations et dans différents contextes. Les règles phonologiques doivent être vérifiées par la synthèse. Ces études sont importantes pour la synthèse par règles mais aussi pour la reconnaissance. Enfin, trop peu de travaux utilisant les sons de synthèse ont été effectués en France pour des tests de perception. Ces travaux doivent être faits d'abord pour vérifier la synthèse et en parallèle pour étudier notre appareil auditif.

BIBLIOGRAPHIE

[1] *COKER C. (1967)*
Synthesis by rule from articulatory parameters
Proc. 1967 Conf. Speech Com. Process, paper A9.

[2] *EL MALLAWANY (1975)*
Contribution aux recherches sur la communication parlée. Etude de vocoders à prédiction linéaire. Détermination de l'intervalle de fermeture de la glotte. Détection de la mélodie. Extraction de la fonction d'aire du conduit vocal.
Thèse de Docteur Ingénieur -Grenoble-.

[3] *FANT G. (1974)*
Key note address
I.E.E.E. symposium on speech recognition. Pittsburg.

[4] *FANT G. (1976)*
Vocal-tract determinants of resonance frequencies and band widths. J.A.S.A. *59*, S 70

[5] *HOLMES J., MATTINGLY I., SHEARME J. (1964)*
Speech synthesis by rule
Language and Speech, *7*, 127-143

[6] *ISHISAKA K., FLANAGAN J.L. (1972)*
Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell System Tech. J. *51*, 1233-1268

[7] *LINDBLOM BE, SUNBERG J. (1972)*
Approaches to articulatory modeling.
3e journée d'étude sur la parole. G.A.L.F. Lannion

[8] *MER MELSTEIN P., MAEDA S., FUJIMURA O. (1971)*
Description of tongue and lip movement in a jaw-lased coordinate system
J. Acoust. Soc. Am., *49*, 104 (A)

[9] *MRAYATI M. (1976)*
Contribution aux études sur la parole. Modèles électriques du conduit vocal avec pertes, du conduit nasal et de la source vocale. Etude de leurs interactions. Relations entre disposition articulatoire et caractéristiques acoustiques.
Thèse d'Etat -Grenoble-.

[10] *OHMAN S.E.G. (1967)*
Numerical model of coarticulation
J.A.S.A. *41*, 310-320

[11] *OLIVE J.P. (1976)*
Hard-wired real-time LPC synthesizer
J.A.S.A., *59*, S 55

[12] *PETERSON G., WANG W., SIVERTSEN E. (1958)*
Segmentation techniques in speech synthesis
J. Acoust. Soc. Amer., *30*, 739-742

[13] *RABINER L. (1968)*
Digital Formant synthesizer for speech-synthesis studies
J. Acoust. Soc. Amer., *43*, 822-828

[14] *SMITH C. (1967)*
Vocal response synthesizer
J.A.S.A., *37*, 170-171

SYNTHESE DE LA PAROLE

DISCUSSION

Question posée par M. CARRE :

Il reste encore beaucoup à faire en synthèse de la parole. Que peuvent apporter aujourd'hui les phonéticiens, à ce sujet, sur le français ?

Réponse de M. CARTON :

C'est la première année qu'aux J.E.P. les thèmes proposés n'ont pas incité les phonéticiens à présenter leurs recherches fondamentales les plus susceptibles d'être utiles aux ingénieurs. Le Laboratoire de Phonétique de Nancy a soumis à approbation un exposé sur la validation de classes intonatives, mais le thème où il s'inscrivait a été supprimé car il n'avait attiré que deux envois. Je peux encore signaler que notre Laboratoire a travaillé aussi cette année, à partir de films radiologiques réalisés à Nancy, sur les transitions et la coarticulation en français.

M. POTAGE à M. CARRE :

1) Y a-t-il eu des travaux effectués sur l'importance de la largeur des formants en synthèse par formants ?

2) Qu'entend M. Carré par "curieusement intelligible" à propos du vocoder CIPHON ?

Réponses de M. CARRE :

1) A ma connaissance, il n'existe pas de travaux sur le rôle des bandes passantes des formants. Seul, Flanagan a étudié les "différences juste perceptibles" lors de modifications des bandes passantes.

2) Par "curieusement intelligible", je veux dire qu'il est curieux, sur le plan perceptif, de constater qu'un signal aussi simple par rapport au signal réel puisse être aussi intelligible.

MMe KONOPCZYNSKI, réponse à M. BOURJAULT :

Réponse à la question de M. BOURJAULT : intérêt (futur) des analogues vocaux.

La mise au point de tels analogues me semble fort intéressante pour la recherche fondamentale, étant donnée la foule de renseignements apportés par la radiocinématographie. Actuellement, ces données servent dans un but essentiel descriptif, mais quel est le rôle linguistique exact de chacun des paramètres relevés ? Les analogues devraient apporter des réponses, au moins sur certains points (cf. par ex. : l'importance des variations de hauteur dans la position du larynx démontrée par Lindblom, à l'aide d'analogues vocaux).

Le problème actuel est en fait le travail commun entre ingénieurs et phonéticiens. Les travaux de cinéradiographie faits en France depuis longtemps déjà (Institut de Phonétique de Strasbourg), n'ont guère servi de base aux recherches des ingénieurs, probablement en partie parce que les bits des uns et des autres sont forts différents. En outre, la cinéradiographie classique -coupe sagittale- ne peut évidemment donner aucun renseignement sur les problèmes de section, fonctions d'aire, ...

Remarque de M. LIENARD :

Je souhaite que l'on clarifie l'emploi des termes : pourquoi utiliser le terme de "diphone" qui fait référence aux travaux de l'équipe IBM (1965), alors que le terme français de "diphonème" me semble convenir parfaitement ?

Remarque de M. DESCOUT :

Il me semble que les problèmes de simulation du conduit vocal, en tant que fonction de transfert, ne posent pas beaucoup de difficultés actuellement, quel que soit le type de simulation adopté (analogie électro-acoustique, calcul analogique ou numérique des équations différentielles, etc...), de

même en ce qui concerne les sources vocales ou de bruit. Mais il ne faut pas que les phonéticiens pensent qu'ils pourront utiliser aussi simplement ce modèle, et il ne faut pas que les ingénieurs pensent utiliser ce conduit vocal simplement. Tout reste à faire au niveau de la modélisation articulatoire et c'est là que toute collaboration reste ouverte.

Remarque de M. WAJSKOP :

En ce qui concerne l'utilisation à des fins fondamentales de la synthèse, il serait important et urgent de veiller à fournir aux phonéticiens et aux psychologues des appareillages sur lesquels ils puissent manipuler avec aisance des paramètres, tels que la durée. A l'heure actuelle, il est difficile au phonéticien d'obtenir des stimulis de parole aussi simples qu'une syllabe CV dans laquelle la douzaine d'indices acoustiques, qui manifestent l'opposition de mouvement, puissent être contrôlés avec précision et de manière indépendante.

M. LIENARD à MM. GENIN, BOURJAULT et GUERIN :

Il y a presque deux cents ans, Kelpelen, puis Faber, construisaient des machines parlantes qui faisaient beaucoup mieux que dire "papa". Je crois que l'idée d'analogue mécanique du conduit vocal est toujours actuelle, et qu'un tel appareil permettrait d'étudier de manière simple et réaliste certains mécanismes fondamentaux de la phonation. Zagorujko, à Novossibirsk, travaille sur un appareil de ce genre, commandé par un calculateur.

Réponse de M. GENIN :

Zagorujko : un modèle purement mécanique et acoustique n'a pas d'avenir direct pour une réalisation et une implantation opérationnelle.

Sur le plan fondamental, il est évident que ce dernier est d'un grand intérêt, mais pas sur un plan application technologique rapide.

Nous envisageons dans un avenir que nous espérons proche pouvoir essager notre modèle de simulation de source glottique, avec d'autres simulateurs de conduit vocal.

M. GRESSER à M. TEIL :

Le LIMSI a-t-il fait des mesures sérieuses sur le temps de programmation.

Réponse de M. TEIL :

Nous n'avons pas fait de mesures pour comparer les temps de programmation mis en utilisant le système cartes perforées et le système décrit. Mais nous pouvons affirmer que le gain de temps global (construction du programme, nombre de compilations en cas d'erreurs syntaxiques, etc...) est très important.

M. BENBASSAT à M. RENARD (remarque) :

Je tenais simplement à ajouter à la liste des applications de la réponse vocale citée par M. Renard, son utilisation pour l'enseignement assisté par ordinateur où le stimuli vocal est utilisé comme complément des stimulis visuels. Je me permets de mentionner au passage l'existence d'un tel système, opérationnel depuis janvier 1976, à l'IMSSS de Stanford University.

THEME 3 :

ANALYSE DU SIGNAL VOCAL

INTRODUCTION A L'ANALYSE DE LA PAROLE

- - -

C. GUEGUEN - Prof. à l'E.N.S.T.
Laboratoire de Théorie des Systèmes

- - -

Résumé : Cet exposé effectue un rapide bilan des techniques conventionnelles d'analyse de la parole. Sur la base des communications proposées sur ce thème aux 7e journées du GALF, il présente les tendances actuelles dans le domaine. L'axe commun à ces tendances étant la modélisation du système phonatoire, une place prépondérante est laissée à la prédiction linéaire qui est analysée sous les points de vue de l'approche théorique, de l'implantation numérique et de ses retombées en analyse spectrale. On montre comment les autres approches récentes s'y rattachent par des choix différents tentant d'en combler certaines lacunes ou d'en affiner certaines des propriétés.

1. Introduction

L'analyse de la parole consiste à extraire de ce signal un faible nombre de paramètres pertinents. Les applications qui mettent en jeu le signal de parole sont à la fois très riches et très diverses : Transmission numérique de la parole, synthèse dans le cadre d'une unité de réponse vocale d'ordinateur, reconnaissance et compréhension de la parole en communication homme-machine, identification ou vérification des locuteurs, aide aux handicapés...

Il ne fait alors pas de doute que le qualificatif de pertinent est variable d'un contexte à l'autre, voire même opposé. Les paramètres extraits en vue de la reconnaissance se devraient d'être attachés au sens du message émis, en même temps que d'être relativement insensibles à la personnalité du locuteur. Les exigences sont évidemment inverses dans un système d'identification du locuteur où l'on aura tendance à diminuer la variabilité due aux significations des messages (en prononçant une phrase code) de façon à ne laisser surgir que les caractères spécifiques du locuteur.

Le nombre de paramètres extraits doit cependant rester relativement faible pour ne pas engorger inutilement les étapes ultérieures du traitement. Dans le cadre d'un système de reconnaissance par exemple, on pourrait être tenté par précaution pour ne pas laisser échapper l'information pertinente, de laisser subsister un nombre important de paramètres pour représenter la forme. Un tel choix serait catastrophique pour les algorithmes suivants qui auraient à manipuler et à mémoriser des monceaux de données parfois inutiles. C'est d'ailleurs dans cet esprit que de nombreux schémas de reconnaissance incluent une étape de réduction de donnée suivant immédiatement la paramétrisation. Cette nécessité de parcimonie est peut-être encore plus claire en transmission de la parole où, à égalité de qualité, la minimalité du taux de transmission (en bits par secondes) est le but final cherché.

Il en résulte que tout processus d'analyse de la parole doit établir, s'il prétend à quelque généralité, un équilibre judicieux entre la richesse de l'information extraite et son volume. Le processus en question doit à la fois respecter une certaine neutralité vis-à-vis du signal et faire usage du maximum d'information à priori sur celui-ci. A ce titre et se cantonant aux techniques à caractère général, nous pensons que les méthodes de modélisation, qui seront étudiées dans la suite, sont d'un intérêt certain.

2. Analyse du signal et modélisation

Le signal de parole, tel qu'il apparaît à la sortie d'un microphone, est une courbe temporelle très complexe puisqu'elle fait figurer des parties quasi-périodiques (sons sonores) et des segments fortement bruités (sons sourds). Face à cette complexité, l'approche typique de la théorie du signal consiste à considérer cette occurrence comme un élément d'un espace fonctionnel et à la projeter sur les fonctions de base de l'espace. La valeur des coordonnées (projections) correspondantes représentera alors le signal en question. C'est ainsi que procède l'analyse spectrale qui dans le cas de fonctions périodiques projette la fonction sur des sinusoïdes de fréquences multiples (développement en série de Fourier) ou plus généralement dans le cas de distributions quelconques, sur un continuum de sinusoïdes (transformée de Fourier). Cette approche considère le signal d'une façon très neutre et répond en général aux deux impératifs de l'analyse : ne pas trop altérer le contenu informationnel du signal, tout en réduisant le nombre de paramètres utilisé à le représenter (contributions des diverses fréquences de base).

Mais l'on peut se demander si, dans le cas de la parole en particulier, cette approche ne manque pas d'une certaine spécificité. Une autre façon d'aborder le problème est de sacrifier l'aspect général de l'approche, pour se questionner sur la manière dont a été engendré ce signal précis. La parole est, en effet, le produit de vibrations acoustiques de l'air dans le canal vocal dont la forme est délimitée par la position des organes phonatoires et le canal nasal, en réponse à la vibration incidente des cordes vocales ou sous l'influence de turbulences dues à des constrictions variées. Il est alors clair que la parole n'est pas un signal banal et qu'une bonne façon de l'analyser consisterait à modéliser le processus qui lui a donné naissance.

La première idée dans cette voie consiste à s'inspirer fortement des données physiologiques et physiques pour aboutir à cette modélisation. Le système phonatoire est, en effet, d'un fonctionnement relativement compréhensible pour le scientifique /1/. Il peut être séparé en trois parties interactives : le canal vocal (éminemment variable), le canal nasal (quasiment fixe) et la source vocale. Moyennant des hypothèses simples qui peuvent être confrontées avec la réalité, il est possible de décrire les vibrations acoustiques de l'air par des équations de propagation (aux dérivées partielles) liant pression et débit volumique en un point du canal. Pour des raisons de simplicité, on supposera, à juste titre, le canal linéaire à section circulaire, divisé en sections de taille quasi constante.... Il n'en reste pas moins qu'un tel modèle vise à une compréhension intime des phénomènes de l'élocution et peut être qualifié de "modèle de connaissance".

Les retombées d'une telle modélisation sont extrêmement importantes et ce surtout au niveau de la recherche. Sur les données issues d'une analyse de l'élocution par une suite d'instantanés (cinématographie aux rayons X par exemple) de la géométrie du canal vocal et de l'ouverture de la glotte, il sera possible de calculer théoriquement le signal de parole. L'analyse pourra alors s'orienter vers le test d'hypothèses pour mesurer l'influence de paramètres négligés (vibration des parois, couplage source-canal, ...) ou pour explorer les conséquences de variations fines de la position des organes phonatoires ou de leur dynamique. Il s'agit donc d'une analyse par synthèse très sophistiquée qui est d'un apport incontestable en phonologie. La difficulté, inhérente à ce type d'approche réside d'une manière évidente dans l'acquisition des différentes grandeurs physiologiques impliquées (aires du canal mais aussi "tension" des cordes vocales....) c'est pourquoi des développements récents d'un grand intérêt visent à mettre au point des techniques d'exploration de ces paramètres physiologiques /2/ /3/. Il convient aussi de dire que, par le biais d'une analogie formelle entre les équations acoustiques et électriques, un moyen de simulation du canal vocal est ainsi disponible donnant accès à la synthèse de la parole. Les paramètres issus de ce type de modélisation bénéficient d'une interprétation physique claire et sont donc bien adaptés à l'implantation de règles de synthèse ou de contraintes phonatoires calquées sur la réalité.

Dans le même esprit, l'idée de s'inspirer des propriétés du système d'audition a été retenue. Dans la mesure où l'on peut analyser le fonctionnement de l'organe de captation (cochlée en particulier) et où l'on peut se faire une idée du traitement ultérieur par le système nerveux, l'approche est, en effet, séduisante. Elle revêt, d'ailleurs, les mêmes nuances dans le caractère plus ou moins physiologique de la modélisation que dans le cas de la phonation. A ce titre, les tentatives où l'élaboration du modèle mathématique se laisse guider par les hypothèses biologiques sont d'un intérêt particulier /4/. Les rapports apparaissent alors étroits entre analyse et reconnaissance /5/.

Mais quand le souci est uniquement d'ordre technique, on est tenté de remettre en question le type de modélisation précédent pour se tourner vers un modèle de représentation. La mise au point d'un modèle de connaissance est toujours délicate malgré les hypothèses simplificatrices. Il doit être remis en question pour chaque nouveau type de signal. Sa forme, elle-même, est le plus souvent complexe du fait de la nécessité de cerner étroitement le phénomène réel (souvent non-linéaire) et son exploitation délicate (nombre

de paramètres assez grand). On peut donc se demander si dans le cadre d'un problème donné, l'établissement d'un modèle simple traduisant le seul comportement externe du système (boîte noire) ne serait pas suffisant. On peut, en effet, questionner dans la parole l'unicité de la position des organes engendrant un son fixé puisqu'il s'agit de résonateurs en cascade. La nature ne semble pas toujours économe de ses moyens et ne doit pas être invitée trop étroitement surtout si l'on tient compte des moyens technologiques disponibles actuellement. Il apparaît donc qu'un modèle simple sous la forme d'une équation de récurrence linéaire (ARMA) joint à un critère d'ajustement externe pourrait fournir certains des avantages précédents sans entraîner une complication excessive.

Les moyens techniques pour l'établissement de tels modèles existent sous la forme du processus d'identification. Pour cela, conformément à la figure 1, le signal y_t réel à analyser est confronté avec un signal synthétique $\hat{y}_t$ issu d'un modèle paramétrique convenablement excité. Un critère mesure la dissemblance entre le signal et son estimé, et un algorithme d'optimalisation tend alors à ajuster les paramètres du modèle pour minimiser le critère retenu. Il s'agit donc typiquement d'une analyse par synthèse où l'on cherche à identifier le processus générateur du signal à analyser. Le résultat fondamental de l'analyse est contenu dans les paramètres du modèle ajusté sur une période stationnaire du signal. L'analyse est réversible puisqu'il est toujours possible de synthétiser à postériori le son correspondant au modèle pour vérifier "à l'oreille" la pertinence de l'analyse.

On a donc ainsi accès à un compromis entre l'aspect trop spécifique d'une modélisation visant à une certaine exactitude physiologique et le caractère standard des procédures usuelles d'analyse qui ignorent délibérément l'origine du signal.

3. L'approche spectrale conventionnelle

Il serait, à coup sûr, inexact de déduire de l'analyse précédente que l'application systématique des techniques spectrales est inadéquate pour le signal de parole. Ce serait, de plus, ignorer que l'essentiel de nos connaissances sur ce signal ont été acquises par ce biais. C'est pourquoi, dans la suite, nous présenterons la séquence d'opérations qui conduisent à une estimation satisfaisante du spectre de la parole et attirerons l'attention sur certaines de ses limitations.

Tel que sa production a été décrite, le signal de parole apparaît susceptible d'une description fréquencielle. Les cavités délimitées par le canal vocal semblent devoir jouer le rôle de résonateurs vis-à-vis du spectre de raies émanant des vibrations périodiques des cordes vocales. Certaines fréquences vont donc se trouver avantagées et dessiner une densité spectrale de puissance dotée de maxima accentués. Les fréquences caractéristiques de ces maxima, appelés formants, sont clairement en relation avec la position des organes phonatoires, donc avec le son émis. Les formants seront donc des paramètres naturels pour le signal vocal, mais leur extraction réaliste suppose un enchaînement d'opérations complexes :

i - dans un premier temps, un segment de parole supposé stationnaire d'une durée de 30 à 50 ms par exemple (quelques périodes fondamentales) est choisi. On en extrait un certain nombre de coefficients d'auto-corrélation (128 par exemple) par des sommes de produits d'échantillons décalés. Pour obtenir une bonne estimation de ces valeurs en évitant les effets de bords dus à la troncation du signal, il convient de pondérer celui-ci par une fenêtre temporelle convenable (Hamming,....) (figure 2a).

ii - La transformation de Fourier discrète (FFT) est appliquée aux coefficients pour calculer la densité spectrale de puissance sur le nombre de points correspondant. Du fait du caractère convolutionnel du signal de parole (source attaquant le canal vocal) le spectre en question est modulé par la fréquence fondamentale ce qui rend délicate la localisation précise des formants. (figure 2b).

iii - Il convient alors de lisser le spectre en séparant les contributions de la source et du canal. Pour cela, suivant la méthode du cepstre /6/, on applique une transformation logarithmique permettant de transcrire sous forme de somme ces deux contributions et d'utiliser leurs zones de fréquences différentes pour les séparer. La transformation inverse après élimination de la source fournit un spectre lisse (figure 2c, d, e)

iv - La détection des maxima peut être alors réalisée pour caractériser la fréquence des formants. Cependant, dans certains cas, des formants peuvent être proches et ne se traduire que par un seul maximum ou par un épaulement de la courbe, mettant la méthode en défaut. On peut alors avoir recours à la chirp z-transform /7/ qui, utilisant un contour d'intégration différent, permet, en les accentuant, de séparer les maxima voisins.

v - Le signal de parole étant, de fait, non stationnaire, on entreprend alors l'analyse du segment suivant. Cette évolution du spectre pose le problème de la poursuite des formants au cours du temps et demande la mise en jeu de techniques souvent complexes pour assurer la continuité des trajectoires.

A la suite de ces différentes étapes, l'analyse du signal vocal est convenablement condensée dans l'évolution temporelle des formants. Mais on note que l'analyse peut se révéler laborieuse et représente un coût de calcul non négligeable (cette remarque est cependant à pondérer du fait de l'existence de matériels spécialisés effectuant corrélation et FFT). Si une certaine condensation de l'information est réalisée dans le spectre (passage 500 points à 128 par exemple) celle-ci n'est pas considérable. C'est plutôt au moment du choix des formants (6 valeurs) que la réduction essentielle est réalisée, mais c'est aussi ce choix qui est le plus critiquable puisqu'il ne tient pas compte de la forme générale du spectre. Par ailleurs, l'analyse est purement spectrale et laisse échapper tout évènement court ou non répétitif (vis-à-vis de la fenêtre considérée) qui peuvent pourtant être significatifs (début de période fondamentale, non-stationnarités diverses...) Pour pallier la lourdeur de l'analyse spectrale conventionnelle, on peut imaginer de recourir à des techniques transformationnelles plus rapides /8/ où à des analyses temporelles ayant un retentissement spectral /9/.

4. Bases de la prédiction linéaire

L'idée d'appliquer à la parole des techniques de modélisation peut être attribuée principalement à Atal et Schroeder /10/. Mais il semble que les concepts en question étaient déjà sous-jacents dans certaines formes évoluées de δ modulation et dans les travaux de certains statisticiens (appliqués à la géophysique, l'électro-encéphalogramme, ...). Une contribution importante est due à Saïto et Itakura qui ont introduit /11/ les techniques modernes de calcul en faisant appel à des résultats antérieurs quelque peu délaissés. Ceux-ci ont été considérablement affinés par Markel et Gray /12/ et une revue complète de l'approche est fournie par Makhoul /13/.

4.1. Principe

La prédiction linéaire est une technique d'identification particulière destinée à ajuster au signal un modèle linéaire autorégressif (AR) au sens d'un critère statistique. Dans cette approche, on suppose que le processus sous-jacent au signal est une équation de récurrence linéaire entre les échantillons successifs y_t du signal :

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

où u_t est un bruit à moyenne nulle.

Si l'on suppose les a_i connus et les y_{t-i} mesurés, l'équation (1) permet de prédire raisonnablement l'échantillon suivant du signal sous la forme :

$$\hat{y}_t = -a_1 y_{t-1} + \ldots - a_p y_{t-p} \qquad (2)$$

Lors de l'arrivée de cet échantillon, on constatera une erreur de prédiction :

$$\varepsilon_t = y_t - \hat{y}_t = \hat{u}_t \qquad (3)$$

Cette erreur de prédiction est d'après (1) et (2) une estimation du bruit u_t celle-ci intègre donc à la fois de possibles erreurs de mesures et le fait que les coefficients a_i ne seront jamais qu'approximatifs. Dans ces conditions, des mesures $\{y_t\}_{t=o,N-1}$ étant disponibles, on déterminera les coefficients a_i optimaux sur cet intervalle en minimisant, par exemple, le critère des moindres carrés de l'erreur :

$$C = \sum_{t=o}^{N} (y_t - \hat{y}_t)^2 = \sum_{t=o}^{N} \varepsilon_t^2 \qquad (4)$$

4.2. Interprétation par identification

Cette approche peut être avantageusement intégrée dans une formulation plus générale d'identification. Il s'agit, en effet, de déterminer les paramètres d'un modèle du type (1) susceptible d'engendrer, au mieux, la série de mesures y_t. Dans le cas présent, aucune information sur l'excitation u_t n'est disponible et l'on en est réduit à des conjectures la concernant. L'hypothèse communément admise ici est qu'il s'agit d'un bruit à moyenne nulle. Il va sans dire que toute information à priori supplémentaire sur cette excitation rendrait la modélisation plus pertinente. En l'absence de cette information, l'hypothèse d'un bruit blanc (à moyenne nulle) peut être retenue du fait de sa généralité et de la nature statistique des critères utilisés. Le risque principal encouru consiste à attribuer au canal vocal certaines propriétés dues à la source.

Mais pour être efficace le schéma de la figure 1 doit être aussi modifié. En effet, l'utilisation directe de cette configuration nécessite d'estimer non seulement les a_i mais aussi p valeurs initiales pour l'équation (1). Le critère ε_t se révèlerait alors très sensible à ces valeurs peu intéressantes rendant l'identification difficile. On préfèrera alors modifier le schéma général selon la figure 3 en changeant le point de calcul de l'erreur. Il s'agira alors d'une erreur d'entrée remplaçant l'erreur de sortie qui se traduira, conformément à l'équation (2), par l'utilisation de l'inverse du modèle dans lequel les vraies valeurs du signal y_t sont utilisées (et non leurs estimations $\hat{y}_{t-i}$). On réalise ainsi une prédiction $\hat{u}_t$ de l'excitation u_t.

$$\hat{u}_t = y_t + a_1 y_{t-1} + \ldots + a_p y_{t-p} \qquad (5)$$

La variance de u_t étant inconnue, on interprétera celle de $\hat{u}_t$ comme la superposition d'un terme fixe provenant de u_t et d'un terme variable indépendant résultant d'une erreur d'estimation des a_i. L'identification sera donc réalisée en minimisant ce terme variable ou, ce qui revient au même, en minimisant la variance de $\hat{u}_t$. Il y a alors coïncidence d'après (3) avec les moindres carrés.

L'essentiel de la démarche consiste donc, ayant fait des hypothèses sur l'entrée u_t, à filtrer le signal y_t par un filtre linéaire représentant l'inverse du modèle cherché, tel que l'estimation produite de l'entrée $\hat{u}_t$ satisfasse au mieux les hypothèses. On conçoit que cette formulation par filtrage inverse dépasse largement le cadre de la prédiction linéraire tel qu'elle a été décrite en premier lieu.

4.3. Formulation

Dans un processus d'identification le choix du critère est un élément important puisque c'est par son seul biais qu'est mesurée la ressemblance entre système et modèle. Parmi tous les critères statistiques qui peuvent être mis en jeu, le maximum de vraisemblance est particulièrement intéressant car il est doté de propriétés asymptotiques très satisfaisantes (quand la taille de l'échantillon d'analyse croît).

Suivant cette approche et supposant connue la densité de probabilité de u_t, on cherchera pour des mesures $\{y_t\}$ données $t = o, N-1$ l'ensemble des paramètres a_i donnant à l'estimation $\hat{u}_t$ déduite de (5) la probabilité la plus grande. Supposant u_t blanc gaussien (centré, écart type σ) et introduisant les notations :

$$u^T = \left[u_o, u_1, \ldots, u_{N+p}\right] \qquad \underline{a}^T = \left[a_o\ a_1 \ldots\ a_p\right]$$

$$Y^T = \begin{bmatrix} y_o & y_1 & \cdots\cdots & y_N & 0 \ldots\ldots 0 \\ 0 & y_o & y_1 & \cdots\cdots & y_N \ldots 0 \\ 0 & & y_o & y_1 & \cdots\cdots\ y_N \end{bmatrix} \quad \text{(p lignes)} \qquad (6)$$

On a par définition :

$$p(u) = \frac{1}{(2\,\Pi\,\sigma^2)^{N/2}} \exp - \frac{1}{2\sigma^2}(u^T u) \qquad (7)$$

Par ailleurs, l'équation (5) permet d'écrire l'entrée estimée

$$\hat{u} = Y\,\underline{a} \qquad (8)$$

Utilisant l'estimation classique de la matrice de covariance Rp de y_t selon

$$R_p = \frac{1}{N}(Y^T Y) \quad , \quad r_i = \frac{1}{N}\sum_{t=o}^{N} y_t y_{t+i} \qquad (9)$$

Il en résulte que le logarithme de (7) évalué pour $\hat{u}$ donne :

$$L = \log p(\hat{u}) = \frac{N}{2}\,\mathrm{Log}\,(2\Pi\sigma^2) - \frac{N}{\sigma^2}(\underline{a}^T R\,\underline{a}) \qquad (10)$$

En supposant σ^2 fixe (ajusté à postériori) la maximisation de la vraisemblance L, revient à minimiser la forme quadratique :

$$C = \underline{a}^T R\,\underline{a} \qquad \text{avec} \qquad a_o = 1 \qquad (11)$$

La solution de ce problème s'explicite sous la forme des équations dites de "Yule Walker" qui permettent de déterminer les a_i optimaux :

$$\begin{bmatrix} r_o & r_1 & \cdots & r_{p-1} \\ r_1 & r_o & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_1 \\ r_{p-1} & \cdots & r_1 & r_o \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix} \qquad (12)$$

D'autres critères auraient pu être retenus comme ceux des moindres carrés (4) du filtrage inverse mais, à quelques détails d'interprétation près, ils sont équivalents dans le cas gaussien et conduisent aux mêmes équations (12). En effet, le critère (4) s'exprime directement par

$$C = \hat{u}^T\hat{u} = a^T Y^T Y a = N \underline{a}^T R \underline{a}$$

mais l'on remarque (par opposition à 10) que σ^2 n'y figure pas. Par ailleurs, on note que les équations (12) sont celles vérifiées par les paramètres d'un système dont les coefficents r_i $i = o, p$ sont imposés.

5. Implantation numérique

La prédiction linéaire n'aurait pas connu son développement actuel si des techniques rapides de résolution des équations (12) n'avaient pas été proposées. L'intérêt d'une méthode d'analyse de la parole est en effet conditionnée par la possibilité de l'implanter en temps réel sans effort excessif. Nous analyserons dans la suite les différents facteurs d'une telle implantation.

5.1. Calcul des coefficients d'auto-corrélation

L'estimation des coefficients d'auto-corrélation d'une séquence discrète $\{y_t\}$ relativement courte n'est pas un problème simple. Le signal de parole étant non stationnaire, le calcul doit être effectué sur quelques périodes fondamentales pour permettre une analyse "instantanée". Les divers estimateurs classiques :

$$r_i = \frac{1}{N} \sum_{t=o}^{N-i-1} y_t y_{t+i} \qquad r'_i = \frac{1}{N-i} \sum_{t=o}^{N-i-1} y_t y_{t+i} \qquad (13)$$

ont des propriétés de biais et de variance qui dépendent notablement du signal. De plus, les estimations de deux coefficients r_i et r_j sont fortement corrélées.

Il en résulte que l'introduction de ces estimations dans R pour former la matrice d'auto-corrélation peut entraîner des difficultés (perte du caractère défini positif de R). Dans ce sens, on pourra avantageusement utiliser l'estimation directe de R sous la forme (9) qui préserve à la fois la structure (Toeplitz) et la positivité de R.

Dans le même but, on pourra utiliser une pré-accentuation des fréquences aigües (filtre passe-haut du premier ordre par exemple). Ce pré-traitement a pour conséquence une réduction de la dynamique du spectre du signal qui se traduit par un meilleur conditionnement de la matrice R (rapport des valeurs propres extrêmes).

Le calcul de ces coefficients représente l'essentiel du temps de calcul en prédiction linéaire (3/4 environ) et il existe à l'heure actuelle peu de corrélateurs adaptés à cette tâche

(faible nombre de points mais grande rapidité). On pourrait utiliser des techniques transformationnelles (FFT, Hadamard...) mais cela serait reduire la méthode à une simple modélisation spectrale (voir paragraphe 6). Une solution au problème semble être le développement pour la parole d'un corrélateur de type hybride, ou de recourir à l'approximation offerte par la corrélation non linéaire/14/.

5.2. Algorithmes de résolution

De nombreux algorithmes, tenant plus ou moins en compte la structure particulière de la matrice R, peuvent être conçus pour résoudre le système d'équations (12). Cette matrice étant symétrique définie positive, les algorithmes de Gauss et Choleski (factorisation triangulaire) peuvent être appliqués. Mais un progrès considérable a été accompli par Levinson en se basant sur la structure de Toeplitz symétrique de R, ce qui a permis à Robinson et à Durbin de fournir des techniques de résolution peu coûteuses.

Considérant les ordres successifs $j = o,p$, la solution est établie de manière itérative sur les coefficients a_i estimés à l'ordre j : a_i^j.

$$k_j = -\left(\sum_{i=o}^{j} a_i^j \, r_{j+1-i}\right) / \alpha_j \qquad (14\ a)$$

$$\alpha_{j+1} = \alpha_j \, (1-k_j^2) \qquad (14\ b)$$

$$a_i^{j+1} = a_i^j + k_j \, a_{j+1-i}^j \qquad (14\ c)$$

$$j = o,p$$

avec pour initialisation :

$$\alpha_o = r_o \quad , \quad a_o^o = 1$$

Les paramètres intermédiaires k_j dans (14 a) sont d'un intérêt considérable et sont appelés coefficients de corrélation partielle (PARCOR). Il peuvent être interprétés comme des coefficients de réflexion tels que ceux existant entre deux sections d'un modèle réaliste du canal vocal. Par ailleurs, ils interviennent de façon fondamentale dans la réalisation câblée du modèle sous la forme d'un filtre en treillis (filtre d'onde) doté de propriétés de sensibilité remarquables. De plus la stabilité du filtre est théoriquement assurée et se traduit $|k_j| < 1$.

Récemment, Le Roux /15/ a proposé un algorithme qui permet une implantation en virgule fixe en introduisant de nouvelles variables e_i^j sous forme d'intercorrélations dont la variation est bornée :

$$e_i^j = E(y_{t-i}\hat{u}_t^j) \qquad \text{avec} \qquad \hat{u}_t^j = \sum_{i=o}^{j} a_i^j \, y_{t-i} \qquad (15)$$

Ces variables peuvent s'interpréter comme des estimations des éléments i de la réponse impulsionnelle de module d'ordre j. Elles donnent lieu aux récurrences suivantes :

$$k_j = -e^j_{j+1} \; / \; e^j_o \qquad (16\ a)$$

$$e^{j+1}_o = e^j_o \; (1-k^2_j) \qquad (16\ b)$$

$$e^{j+1}_i = e^j_i + k_j e^j_{j+1-i} \qquad (16\ c)$$

où l'on remarque que les paramètres a^j_i , dont la plage de variation est inconnue, sont remplacés par les e^j_i bornés par r_o.

Cet algorithme est particulièrement adapté à une implantation sur microprocesseur. La figure 4 donne un tableau des coûts de calcul.

5.3. Détermination de l'ordre du modèle

Un problème important et délicat est l'évaluation de l'ordre p du modèle. Celui-ci peut, en effet, varier d'un son à l'autre (longueur du canal vocal...). Mais de plus, rien n'assure que certains sons complexes (nasalisés par exemple) puissent être convenablement représentés par un modèle autoregressif (AR). Il faudra, alors, compenser cette faiblesse par un ordre important.

Dans l'interprétation par prédiction linéaire, on juge de la qualité de la modélisation sur le résiduel d'erreur. Le critère correspondant est l'écart type α_j estimé du bruit sur l'entrée u_t pour un ordre j . Cette valeur est liée au déterminant de R_j. Divers tests ont ainsi pu être proposés :

$$\text{Gersch} : (\alpha_p - \alpha_{p+1}) \; / \; \alpha_p < \delta$$
$$\text{Chow} \quad : \det (R_p) < \delta' \qquad (17)$$

Mais la décroissance lente du déterminant en fonction de j (bruit, inadéquation du modèle...) ne permet pas une détermination franche de p .

Dans l'interprétation par filtrage inverse, ce sont les propriétés statistiques du résiduel $\hat{u}_t$ qui sont exploitées. Pour l'ordre correct, si le signal a été engendré par un système AR (et au moins pour un ordre infini), ce résiduel devrait être blanc. La difficulté est que les tests de blancheur sont coûteux et multiformes (définir la relative platitude du spectre, la nullité des coefficients de corrélation...)

Dans ces conditions, un critère de type entropique proposé par Akaike /16/ sera préféré pour faire le lien entre les deux approches

$$\min_p \; I(p) \; : \; I(p) = \log \frac{\alpha_p}{r_o} + \frac{2p}{N} \qquad (18)$$

Ce critère appliqué à la parole donne un ordre moyen de 12 pour une fréquence d'échantillonnage de 10kH . Il se révèle fiable et parcimonieux dans le nombre de paramètres déduit. La figure 5a montre l'évolution typique de I(p). La figure 6 représente les modèles obtenus pour différents ordres p.

6. Interprétation spectrale de la méthode

L'un des aspects les plus riches de la prédiction linéaire est de conjuguer les qualités d'une analyse temporelle (calcul de l'erreur de prédiction instantanée) et d'une analyse fréquentielle (par le biais du modèle). L'interprétation spectrale réalisée peut prendre deux formes :

i - Modélisation d'un spectre mesuré point par point $S(\omega)$ (réduction de paramètres) ;

ii - Estimation du spectre d'un signal y_t.

6.1. Critère d'estimation spectrale

Soit $S(\omega)$ le spectre y_t , la connaissance du modèle permet de déduire le spectre du résiduel (en principe blanc) à partir de la fonction de transfert $A(z)$ du filtre inverse :

$$U(\omega) = S(\omega).|A(e^{j\omega})|^2$$

D'où l'évaluation fréquentielle du critère temporel :

$$C = \sum_{t=o}^{\infty} \varepsilon_t^2 = \frac{1}{2\Pi} \int_{-\pi}^{+\pi} A(e^{j\omega})S(\omega)A(e^{-j\omega})d\omega \qquad (19)$$

Si par ailleurs on suppose la variance σ^2 du résiduel calculée à postériori (après estimation des a_i), on a l'estimation du spectre $\hat{S}(\omega)$:

$$\hat{S}(\omega) = \sigma^2 / |A(e^{j\omega})|^2$$

d'où le critère :

$$C = \frac{\sigma^2}{2\Pi} \int_{-\pi}^{+\pi} \frac{S(\omega)}{\hat{S}(\omega)} d\omega \qquad (20)$$

qui s'interprète comme la moyenne arithmétique du rapport des spectres réel et estimé (au coefficient σ^2 prés).
Cette interprétation soulève le problème de la normalisation du spectre estimé. Dans le critère des moindres carrés, la normalisation ne peut intervenir que par le calcul de σ^2 . Utilisant le maximum de vraisemblance (10) par contre, σ^2 pourrait être considéré comme une fonction des a_i et conduirait à un critère d'estimation spectrale légèrement différent /11/.

6.2. Caractères de l'estimation réalisée

Par opposition à l'approche classique de l'analyse spectrale, la détermination du spectre par modélisation présente les qualités suivantes :

i - le spectre estimé est lisse et analytiquement décrit par l'équation du modèle. Il dépend d'un faible nombre de paramètres (de l'ordre de 12) décrivant l'ensemble de la courbe. Les formants se présentent comme des maxima généralement très accentués. (Voir figure 5c).

ii - Il n'est pas nécessaire d'effectuer une détection des maximas pour calculer les formants. Ceux-ci sont donnés par les arguments des pôles de la fonction de transfert du modèle (racines de $A(z)=0$). Bien que leur présence puisse ne pas apparaître sur le spectre, l'existence de deux formants voisins se traduit le plus souvent par deux pôles permettant d'extraire les fréquences correspondantes (voir figure 5b).

A ces propriétés se superposent celles déduites du critère d'estimation spectrale (20) mis en jeu par la méthode :

iii - Propriété globale : les spectres réel et estimé étant relativement proches, la contribution de chacune des fréquences au critère est globalement uniforme. Cela n'aurait pas été le cas pour un critère portant sur la différence absolue (Analyse par synthèse)

$$C = \int_{-\pi}^{+\pi} |S(\omega) - \hat{S}(\omega)|^2 \, d\omega$$

qui du fait de la décroissance du spectre avec la fréquence a tendance à favoriser le bas du spectre.

iv - Propriété locale : Du fait de la réduction de paramètres recherchée, le spectre $\hat{S}(\omega)$ est toujours plus lisse que $S(\omega)$. En conséquence, localement autour d'un maximum de $S(\omega)$, on a

$$S(\omega) > \hat{S}(\omega)$$

d'où une contribution importante de cette zone au critère et donc un ajustement précis de spectre estimé. La prédiction linéaire a donc tendance à mieux estimer les pics du spectre que ses vallées ce qui est un élément favorable pour la parole (voir figure 7).

Utilisée comme technique d'estimation spectrale, la prédiction linéaire n'est cependant pas dépourvue de défauts. Son efficacité est toute relative quand il s'agit d'évaluer un spectre discret (raies) /17/.

Par ailleurs, si les fréquences des formants sont déterminées avec précision il n'en va pas de même des bandes passantes (amortissements) correspondantes qui sont très sensibles au bruit.

7. Autres approches et développements

La prédiction linéaire en traitement de la parole possède des propriétés intéressantes qu'il conviendrait sans doute d'affiner pour des applications particulières. Elle laisse aussi subsister quelques lacunes auxquelles il est nécessaire de remédier. C'est dans ces deux orientations que ce sont concentrées les recherches récentes dans le domaine.

7.1. Identification de la réponse impulsionnelle du canal

Une critique très directe qui peut-être valablement faite sur la prédiction linéaire est l'absence de signification physique claire aux paramètres a_i . La détermination des formants nécessite le calcul des racines du polynome $A(z)$ qui est une opération coûteuse. Si l'information formantique est requise, une approche plus directe peut-être tentée.

Considérant le canal vocal comme une suite de résonateurs en cascade, on peut modeliser /18/ le signal de parole comme une somme de réponses impulsionnelles élémentaires associées à chacun des formants :

$$\hat{y}_t = a_o + \sum_{i=1}^{p} e^{\rho_i t} \left| a_i \sin \omega_i t + b_i \cos \omega_i t \right|$$

où $\omega_i \rho_i$ représentent la pulsation et l'amortissement du formant i alors que a_i et b_i décrivent la contribution et le déphasage correspondants. Ces hypothèses sont très réalistes pour les sons sonores pendant la réponse libre du système (fermeture de la glotte). Bien que d'intérêts inégaux les paramètres ont des significations physiques claires.

Le critère des moindres carrés calculés sur l'intervalle de réponse libre de largeur N :

$$C = \sum_{t=o}^{N} (y_t - \hat{y}_t)^2$$

est alors une fonction non-linéaire des paramètres mais demeure quadratique en (a_i , b_i) . Un algorithme du gradient partitionné calcule les valeurs optimales des paramètres. L'initialisation délicate en (ρ_i, ω_i) est faite par référence au segment analysé précédent.

7.2. Modélisation linéaire factorielle

L'hypothèse fondamentale de la prédiction linéaire est que le coefficient a_o du modèle AR est égal à l'unité. Une normalisation différente est ici adoptée sur la norme du vecteur des paramètres. Le critère des moindres carrés conduit alors à la minimisation :

$$\min_{\underline{a}} \ \underline{a}^T R \, \underline{a} \qquad \text{avec} \qquad \underline{a}^T \underline{a} = 1$$

On montre alors que la solution optimale est le vecteur propre de R associé à la valeur propre minimale.

Utilisant les propriétés des matrices de Toeplitz, on déduit que /19/ les pôles du modèle sont situés sur le cercle unité et que le spectre se réduit à un certain nombre de fréquences pures dont il est possible de calculer l'amplitude (Spectre de raies). L'application à la parole /20/ montre que les fréquences correspondantes, correspondent aux maxima du spectre après filtrage par le cepstre. La méthode est donc particulièrement bien adaptée à la détection de formants voisins.

7.3. Modélisation discriminante

Dans certaines applications le modèle d'un phonème doit-être établi par référence à son environnement. C'est en particulier le cas de la segmentation automatique où l'on propose /21/ une analyse comparative de segments de parole basée sur un critère discriminant. Soient R et R' les matrices de correlation des deux segments, pour un modèle donné, la variance des résiduels correspondants est :

$$\sigma^2 = \underline{a}^T R \underline{a} \qquad \text{et} \qquad \sigma'^2 = \underline{a}^T R' \underline{a}$$

on cherchera donc le vecteur $\underline{a}$ minimisant le critère discriminant

$$C = \sigma^2/\sigma'^2$$

cette approche est analogue à la segmentation par fenêtre glissante et filtrage inverse.

L'approche peut-être appliquée à l'apprentissage d'un système de reconnaissance basé sur un banc de filtres inverses accordés sur chaque son élémentaire (vocodeur à canaux adaptés /22/). Les modèles de chacun des phonèmes sont ainsi ajustés pour accepter au mieux l'un d'entre eux et rejeter le plus possible tous les autres. Dans le même esprit, on peut songer à établir une base de modèles discriminants telle que la parole s'y projette de manière optimale /23/.

7.4. Prédiction linéaire récursive

Pour bénéficier pleinement des caractéristiques d'analyse temporelle de la modélisation, il est intéressant, au lieu d'établir un modèle moyen sur une fenêtre temporelle, de rendre l'estimation adaptative et d'actualiser l'estimation à l'arrivée de chaque donnée nouvelle.

C'est ce qui est réalisé par l'utilisation du filtre de Kalman /24/ où l'estimation à l'instant t est corrigée à l'instant suivant en tenant compte de l'erreur de prédiction constatée :

$$\underline{\hat{a}}_{t+1} = \underline{\hat{a}}_t + K_t(y_t - \hat{y}_t)$$

Cette procédure peut-être comprise comme un algorithme du gradient à pas variable. Le filtre de Kalman correspond à la détermination du pas optimal K_t assurant le minimum de variance à l'estimation.

Les calculs impliqués sont relativement lourds. Ils peuvent être réduits en considérant des versions sous-optimales (contraintes de structure sur K_t) ou en admettant de ne pas actualiser le gain à chaque instant. Dans cet esprit, on pourra appliquer après convergence un gain K asymptotique qui ne sera remis en question que lors d'une détection de non-stationnarité.

Les facultés d'adaptation du modèle peuvent être accentuées, au prix d'une variance supérieure de l'estimation, en prévenant une convergence trop rapide (gain K_t faible) dans les régions stationnaires. Une procédure du même type, mais d'une application délicate, consiste à rendre la mémoire finie en soutirant systématiquement les informations passées. Dans la pratique, on se contentera souvent, au risque de laisser échapper un évènement intéressant, de procéder à des réinitialisations périodiques ou automatiques du filtre.

Il apparaît que les techniques récursives ainsi conçues sont des outils d'analyse extrêmement puissants en ce qui concerne la parole (détection synchrone du filtre, intervalle de fermeture de la glotte). Elles restent cependant lourdes et d'une fragilité certaine au niveau du choix des initialisations.

7.5 Identification de Modèles ARMA

Le choix exclusif d'un modèle autoregressif AR (fonction de transfert ne comportant que des pôles) peut paraître quelquefois rudimentaire. En particulier si l'hypothèse de résonateurs en cascade est souvent admissible, il n'en est pas de même dans le cas des sons nasalisés où s'introduit un trajet parallèle. Dans ce cas, et dans de nombreux autres, on doit alors recourir à modèle plus complexe (ARMA) comportant à la fois des pôles et des zéros dans sa fonction de transfert :

$$y_t + a_1 y_{t-1} + \ldots + a_p y_{t-p} = b_o u_t + \ldots + n_q u_{t-q} \qquad (21)$$

ceci ne signifie pas que le modèle AR doive être systèmatiquement abandonné. L'ordre du modèle AR n'étant pas fixé, un zéro stable (intérieur au cercle unité) peut-être approximé par un nombre suffisant de pôles supplémentaires. Une telle solution peut cependant se révéler excessivement coûteuse en paramètres. D'où l'intérêt d'identifier un modèle ARMA complet .

Cette identification se heurte au fait que les u_t et b_i étant à la fois inconnus, il s'agit d'un problème d'estimation non linéaire qui n'a pas de solution connue simple. Dans ces conditions, on peut imaginer une procédure itérative condensant en un nombre réduit de zéros, un nombre élevé de pôles supplémentaires destinés à modéliser un résidu de prédiction non-blanc. Une telle technique est proposée sur la base de l'algorithme de Durbin appliqué à l'envers (réduction de l'ordre d'un modèle plus complexe) /25/.

Une telle approche est à coup sûr coûteuse en temps de calcul puisqu'elle suppose l'application répétée de la prédiction linéaire. Mais les algorithmes rapides disponibles limitent cet inconvénient.

8. Conclusions

Les techniques de modélisation et , parmi elles, la prédiction linéaire qui est la plus directe, bénéficient d'un intérêt particulier en analyse de la parole :

i- Elles réalisent une analyse de portée générale en s'appuyant sur les caractères spécifiques du sytème de phonation, d'où un nombre de paramètres restreint.

ii- La modélisation étant réversible, elles autorisent, par le biais de la synthèse, une vérification de la pertinence de l'analyse.

iii- Des algorithmes rapides permettent d'espérer l'identification, en temps réel et à coût relativement faible, des paramètres du canal vocal.

On pourrait alors croire le problème d'analyse de la parole en voie d'être résolu. Si des progrès substantiels ont été accomplis, c'est cependant encore loin d'être le cas car deux problèmes fondamentaux semblent encore résister à l'étude :

- Le signal de parole est éminemment non stationnaire et l'on peut soupçonner qu'une part importante de l'information contenue réside précisément dans ces non-stationnarités dont les processus actuels d'analyse ne rendent que faiblement compte.
- Par ailleurs, le résiduel de prédiction, dont la blancheur est le garant de la qualité de l'analyse, véhicule des informations encore mal comprises. La détection des intervalles de fermeture de la glotte est un problème qui nécessite toute la puissance de la prédiction linéaire /26/. Malgré des progrès encourageants /27/, le codage encore délicat de la source donne à penser que des aspects fondamentaux du signal vocal ont encore échappé à la modélisation du canal vocal.

C'est dans l'approfondissement de ces deux paradoxes que de nouvelles techniques d'analyse pourraient, nous semble-t-il, puiser bientôt leur sources.

REFERENCES

/1/ G. FANT : Acoustic theory of speech production,
Mouton-The Hague, Paris, 1970

/2/ P. ANDRE, M. FILLEAU, F. HAMELIN : Glottométrie en temps réel et applications,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/3/ R. DESCOUT, B. TOUSSIGNANT, M. LECOURS : Deux méthodes de détermination de la fonction d'aire du conduit vocal dans le domaine temporel,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/4/ J. CAELEN, G. PERENNOU : Un modèle d'oreille appliqué à l'analyse de la parole,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/5/ M. CARTIER, J. L. COURBON, R. DAGORNE, R. DOUCEN, J. J. LUCAS : Détection des six sons d'un vocabulaire artificiel,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/6/ A. M. NOLL : Cepstrum pitch determination,
JASA, pp 293-309, 1966

/7/ L. R. RABINER, R. W. SCHAFER, C. M. RADER : Chirp z-transform algorithm and its applications,
BSTJ, pp 1249-1291, 1969

/8/ A. MURE-RAVAUD, C. BERGER-VACHON : Reconnaissance automatique de la parole par l'emploi de la transformée de Walsh Hadamard,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/9/ M. BAUDRY, B. DUPEYRAT : Analyse du signal vocal, Utilisation des extrema du signal et de leur amplitude, Détection du fondamental et recherche des formants,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/10/ B. S. ATAL, S. L. HANAUER : Speech analysis and synthesis by linear prediction of the speech wave,
JASA, n° 50, pp 637-655, 1971

/11/ F. ITAKURA, S. SAITO : Digital filtering techniques for speech analysis and synthesis,
7ème ICA, Budapest, 1971

/12/ J. D. MARKEL, A. H. GRAY : On autocorrelation equations as applied to speech analysis,
IEEE Trans on AU, Vol AU-21, n° 2, 1973

/13/ J. MAKHOUL : Linear prediction, a tutorial review,
Proceedings of IEEE, Vol 63, n° 4, 1975

/14/ C. GALAND, D. ESTEBAN, F. DUBUS : Détection de la mélodie par auto-corrélation non-linéaire,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/15/ J. LE ROUX : Optimisation du calcul des coefficients de corrélation partielle,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/16/ H. AKAIKE : A new look at the statistical model identification,
IEEE Trans. on AC, Vol AC-19, n° 6, 1974

/17/ J. MAKHOUL : Spectral linear prediction, properties and applications,
IEEE Trans. on ASSP, Vol ASSP-23, n° 3, 1975

/18/ G. PERENNOU, D. DOURS, R. FACCA : Analyse temporelle du signal vocal comparée à l'analyse fréquentielle du point de vue de la reconnaissance,
Rapport final sur contrat SESORI 74-94, 1976

/19/ C. GUEGUEN : The modified linear prediction, a factorial approach to speech analysis,
Rapport UCLA, n° ENG 75-40, 1975

/20/ G. CARAYANNIS, C. GUEGUEN : Modélisation de la parole par diagonalisation de sa matrice d'auto-corrélation,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/21/ G. CARAYANNIS : Analyse comparative du signal de parole,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/22/ C. GUEGUEN, T. FARJAUDON, F. LE CHEVALIER : Un vocoder à canaux adaptés,
L'Onde Electrique, Vol 55, n° 7, pp 369-372, 1975

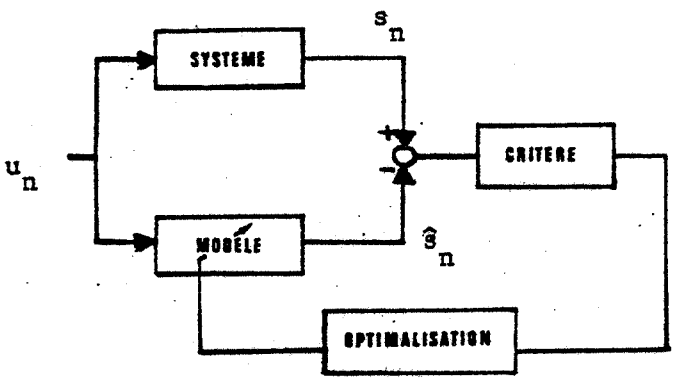
/23/ M. C. HATON, J. P. HATON : Une méthode de représentation du signal vocal en base adaptative,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/24/ C. GUEGUEN, G. CARAYANNIS : Analyse de la parole par filtrage optimal de Kalman,
4ème Journées d'Etude sur la Parole (GALF), Bruxelles, 1973

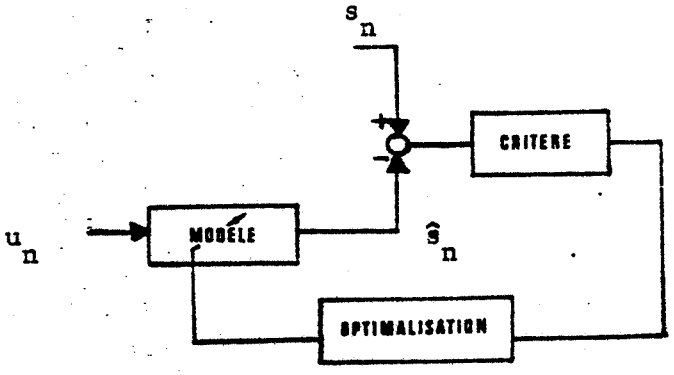
/25/ C. GUEGUEN, M. MATHIEU : Contribution des zéros à l'analyse de la parole, 7ème Journées d'Etude sur la Parole (GALF) (Comm. libres), Nancy, 1976

/26/ I. EL MALLAWANY : Approches à la détection et à l'analyse de l'intervalle de fermeture de la glotte,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

/27/ J. MENEZ, D. ESTEBAN, F. DUBUS : Analyse du signal vocal en vue de son codage à faible débit d'information,
7ème Journées d'Etude sur la Parole (GALF), Nancy, 1976

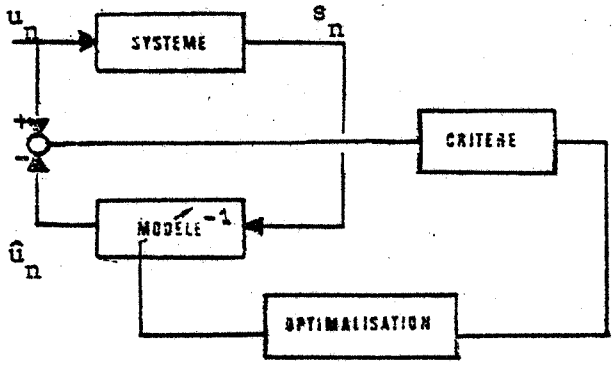


a) Identification par erreur de sortie

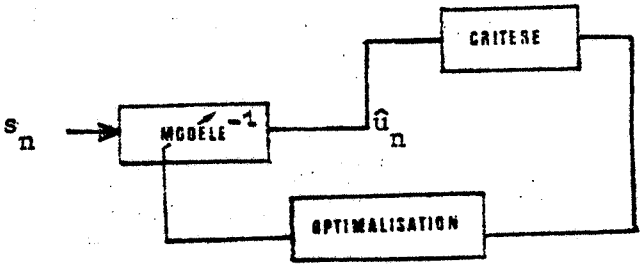


b) Application au signal

Figure 1



a) Identification par erreur d'entrée

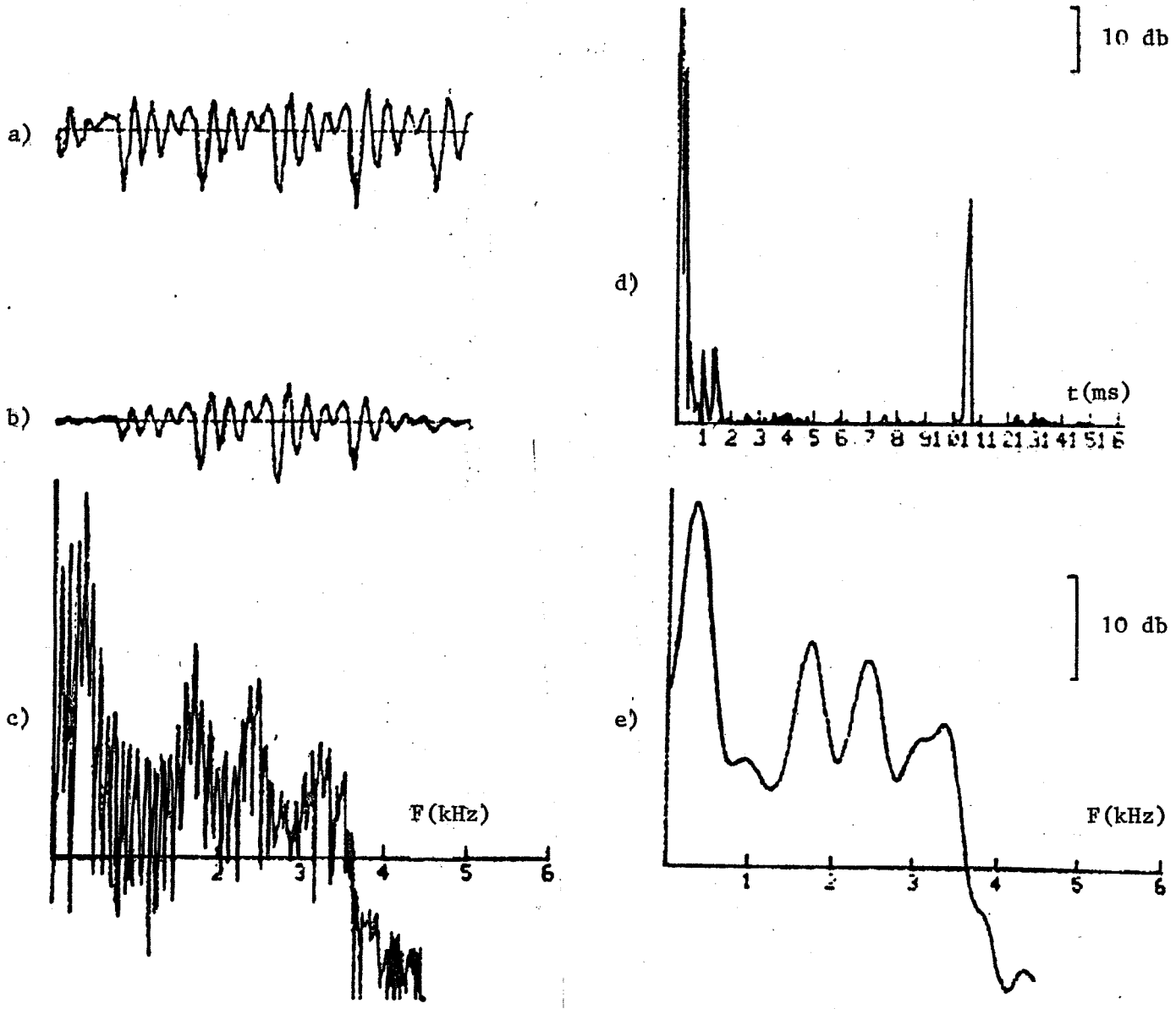


b) Filtrage Inverse

Figure 3

	Mots mémoire	Nombre de multiplications
. Elimination de Gauss	p^2	$p^3/3$
. Décomposition de Cholesky	$p^2/2$	$p^3/6$
. Algorithme de Levinson	2 p	2 p^2
. Algorithme de Durbin	2 p + 6	1.25 p^2
. Algorithme de Le Roux	2 p + 3	p^2 - p + 1

Figure 4 Algorithmes de résolution des équations normales



a) Signal original

b) Après fenêtre de Hamming

c) Spectre brut par FFT

d) Cepstre

e) Spectre lissé

Figure 2 Application de l'analyse spectrale à la parole (d'après CARAYANNIS - Thèse de Docteur-Ingénieur)

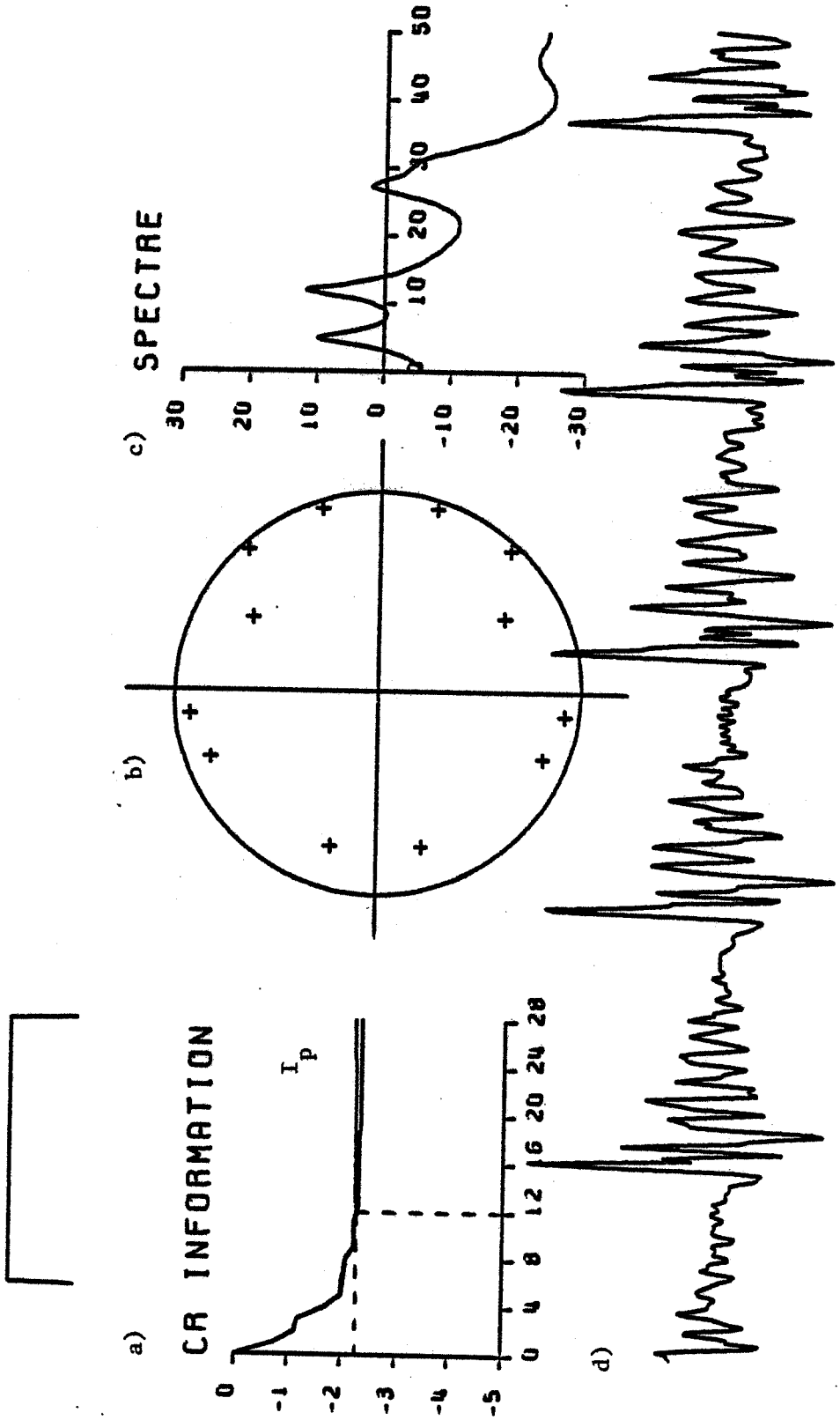


Figure 5 a) Critère d'ordre b) Poles du modèle c) Spectre déduit d) Signal original /a/

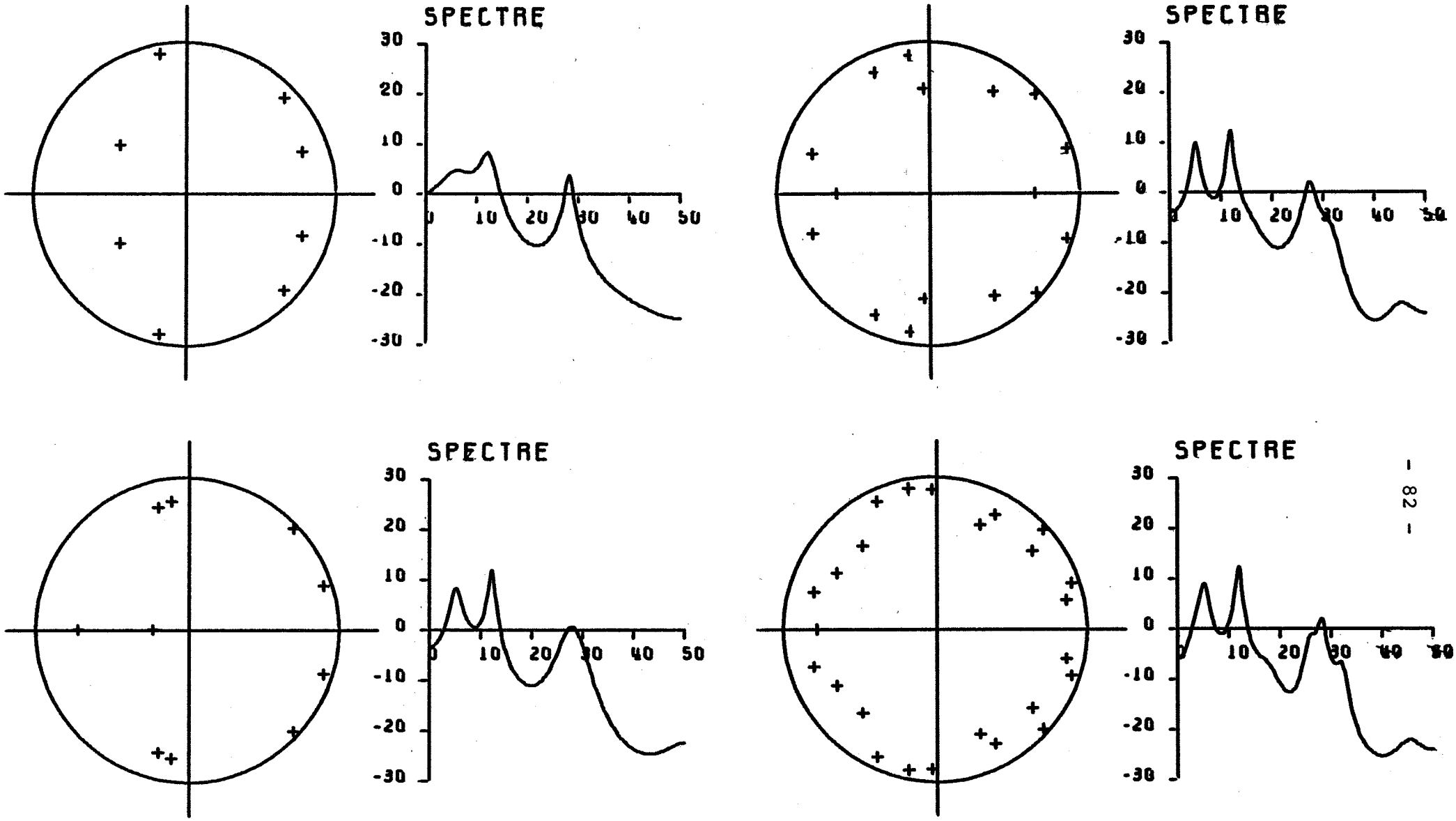


Figure 6 Modèles obtenus (poles et spectre) pour différents ordres p=8 p=10 p=16 p=25

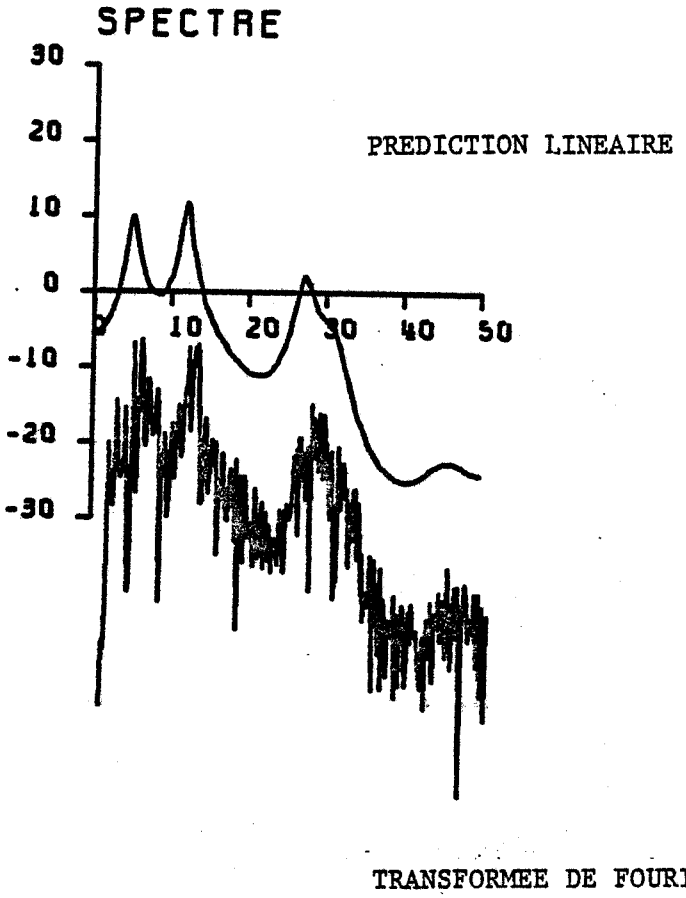


Figure 7 Comparaison des spectres obtenus par prédiction linéaire et FFT

ANALYSE DU SIGNAL VOCAL

DISCUSSION GENERALE

M. GUEGUEN à M. BERGER-VACHON :

Le système que vous proposez, déjà utilisé dans le domaine des images, fait appel à un débit informationnel très riche, qui dépasse largement les valeurs généralement utilisées pour le codage de la parole. Comment pensez-vous traiter ce surcroît d'information ?

Réponse de M. BERGER-VACHON :

Effectivement, le codage nécessite 64 x 6 bits pour 10 millisecondes de signal quasi-stationnaire, et il est certain que des transformations par matrice de Hadamard 16 x 16 (ne demandant que 16 x 4 x 100 = 6400 bits par seconde) sont plus proches du phénomène physique.

Actuellement, nous essayons un peu systématiquement l'étude, par la transformation de Hadamard, du signal vocal ; c'est en fonction des résultats que nous effectuerons les réductions nécessaires.

M. CARRE à M. GALAND :

Je regrette que les orateurs ne comparent pas leurs méthodes par rapport à celles déjà existantes.

Réponse de M. GALAND :

Effectivement, une multitude de détecteurs de mélodie ont déjà été proposés, mais une étude comparative d'un grand nombre de ces détecteurs représente un travail important. Un tel travail a été effectué en 1975 par un étudiant du MIT,

en liaison avec la Bell System (L. Rabiner). La question : "quelle est la meilleure méthode" ? n'a pas été élucidée dans ce travail. Tout dépend du type d'erreurs considérées et de l'application envisagée.

M. DREYFUS-GRAF à M. CARTIER (remarque):

Les résultats communiqués actuellement proviennent d'un premier prototype, le phono-décodeur I, nommé CHARLES.

J'avais proposé un prototype plus élaboré, nommé phono-décodeur II. Celui-ci est prévu pour distinguer, entre autres, les plosives p, t, k, des voyelles explosées, en particulier grâce à une analyse spectrale de l'explosion elle-même. Les descriptions des phono-décodeurs I et II, ainsi que leurs différences, sont résumées dans ma communication "Recognition of Coded Speech (Phonocodes)", publiée dans les "Proceedings of the IEEE", ICASSP, Philadelphia, April 1976.

COMMUNICATIONS DE DERNIERE HEURE

ANALYSE DU CONDUIT VOCAL

PAR INVERSION D'UN MODELE MATHEMATIQUE

J. GENIN

C.N.E.T. LANNION

Notre but est de retrouver les paramètres du conduit vocal à partir d'un enregistrement du signal de parole.

CHOIX D'UN MODELE.

Nous avons choisi un modèle de simulation du conduit vocal semblable à celui de notre réalisation câblée. La fidélité au phénomène naturel de la phonation nous paraît important dans cette étude. Aussi les paramètres de commande sont-ils de nature physiologique:

pression d'air dans les poumons,
écartement des cordes vocales,
tension des cordes vocales,
loi de section du conduit vocal, que nous espèrons au plus tôt remplacer par un jeu de paramètres articulatoires.

INVERSION DU MODELE.

On cherche ici à optimiser le choix des paramètres accessibles pour approcher au mieux le signal de parole enregistré. La méthode utilisée est une méthode de calcul de gradient.

Les toutes premières expérimentations menées à ce jour ont permis de retrouver les paramètres de constitution du modèle (masse des cordes vocales par ex.) sur un signal de parole de synthèse. Aucune expérimentation n'a encore été faite sur un signal de parole naturelle.

BIBLIOGRAPHIE

Simulation du conduit vocal en technologies analogiques.
J.L. COURBON. 6es J.E.P. TOULOUSE 1975

Choix d'un modèle mathématique pour la simulation du conduit vocal. Etude de la production des voyelles.
A. GUERARD Note interne CNET/CES/98

Identification de la fonction d'aire du conduit vocal dans un modèle simplifié sans perte.
A. GUERARD Note interne CNET/CES/100

Etude de la source sonore dans un système de synthèse de la parole: modèle des cordes vocales à deux masses.
M. GROJNOWSKI Note interne CNET/CES/101

Identification de la fonction d'aire et des coefficients d'atténuation du conduit vocal dans un système avec pertes.
A. GUERARD Note interne CNET/CES/103

Identification des caractéristiques physiques des cordes vocales et du conduit vocal dans un modèle, avec pertes, de production de sons.
A. GUERARD Note interne CNET/CES/107

FREQUENCE, INTENSITE ET DUREE : ETUDE COMPARATIVE DES FONCTIONS DANS LES PHRASES ENONCIATIVES SIMPLES ET ETENDUES

Mme CAELEN G. - M MAURAND G.
Université de TOULOUSE-LE-MIRAIL

INTRODUCTION :

Dans un premier stade de notre recherche, nous avons voulu étudier le comportement d'ensemble des trois paramètres prosodiques. Nous avons procédé à l'enregistrement de 38 phrases assez courtes (15 syllabes au plus) lues par 9 locuteurs (7 hommes + 2 femmes). Ces phrases ont été traitées par les analyseurs de parole (mélodimètre, intensimètre et oscillographe) qui ont reproduit à la vitesse de 100 mm/s les trois tracés correspondant aux trois paramètres : la fréquence fondamentale, l'intensité et la durée. Le mélodimètre a filtré la voix dans la bande 90-250 hz pour les hommes et 170-370 hz pour les femmes.

Nous ne nous sommes intéressés qu'aux voyelles seules en négligeant délibérémentles influences de l'entourage consonantique. Nous avons procédé à un codage parallèle de ces paramètres en divisant à chaque fois l'écart maximum entre minima et maxima de l'ensemble des réalisations vocaliques de chaque locuteur, en 4 niveaux. Il n'est pas question pour nous de reconnaître à ces niveaux une fonction, mais d'apprécier en première approximation la qualité et le sens des variations, d'étudier si le comportement mélodique des mots et groupes de mots est superposable ou non à ceux de l'intensité et de la durée, et de préciser enfin les fonctions respectives des 3 paramètres.

1. NOTIONS GRAMMATICALES : Thème (T) et Prédicat (P)

Sur le plan de la syntaxe, le thème représente ce dont on parle, le prédicat est le commentaire, l'information que l'on apporte au T. L'opposition T / P ne recouvre pas systématiquement l'opposition sujet / verbe.

2. ANALYSE DES TROIS PARAMETRES PROSODIQUES

1. Analyse de la fréquence

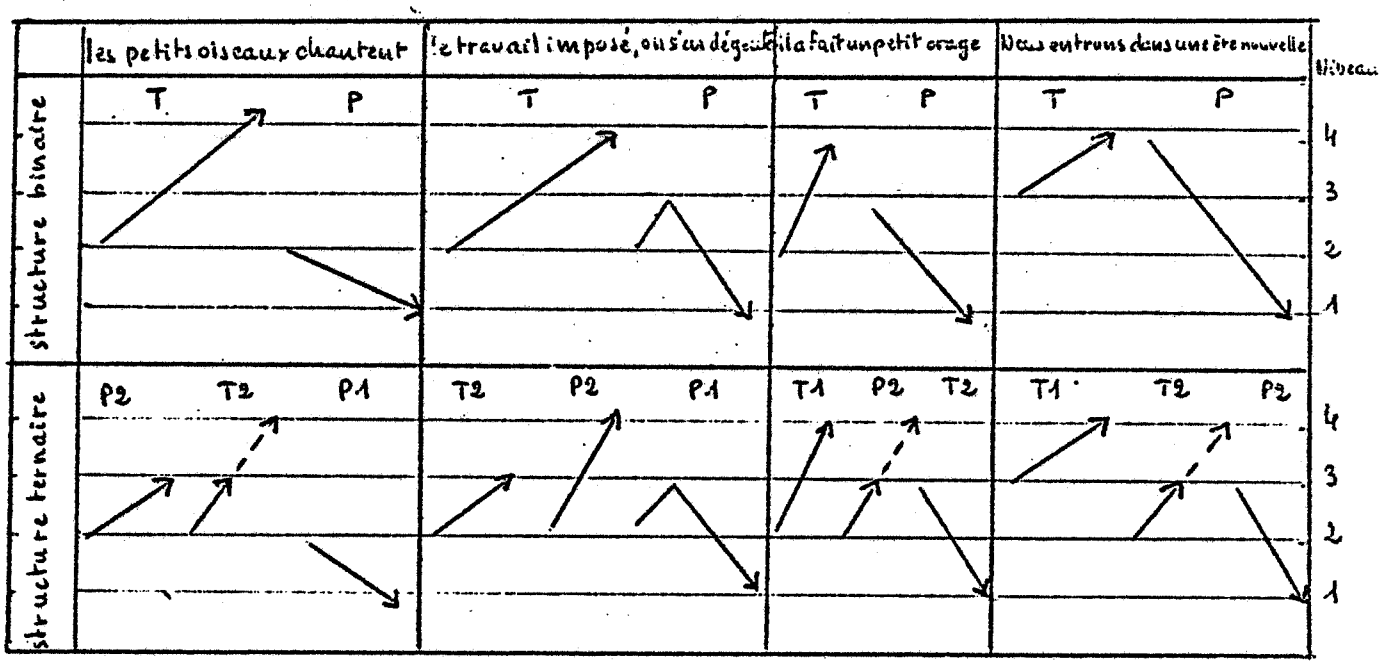
Les phrases que nous avons étudiées sont toutes mélodiquement structurées en deux parties : une partie ascendante (intonème positif ↑) qui caractérise le T, une partie descendante (intonème négatif ↓) qui caractérise le P.

A l'intérieur de cette structure on remarque que les monosyllabes

à fonction essentiellement grammaticale (article, pronom, préposition conjonction, verbe avoir, verbe être) sont prononcés à un niveau bas (15).

Les phrases à T et/ou P complexes sont toutes analysables de cette manière, mais parfois les locuteurs ont choisi de distinguer T et P de rang inférieur, à l'intérieur de cette structure phrastique binaire.

Nous rencontrons dans notre corpus 4 types de phrases. Ces phrases se décomposent de manière ternaire.



On constate que lorsque le T est complexe, c'est à dire composé d'un T et d'un P de rang 2, les inflexions sont positives. Lorsque le P se décompose à son tour en un T et un P de rang inférieur, les inflexions s'opposent, la pente négative suivant la pente positive, quelle que soit la nature du premier élément de rang 2.

Il semble donc que l'on puisse poser l'hypothèse que la structure mélodique de rang supérieur prévaut sur celle de rang inférieur et qu'elle impose aux seconds éléments des groupes complexes (qui coïncident avec les finales de T et P), la pente adéquate.

En d'autres termes, tout mot lexical est susceptible d'être mis en relief par la mélodie de la phrase. Lorsque le T est complexe, le premier élément (T2 ou P2) non terminal, dans la grande majorité des cas en français, est affecté d'un intonème positif de continuation mineure ; le deuxième élément terminal (T2 ou P2) est affecté d'un intonème positif de continuation majeure. Lorsque le P se décompose à nouveau en deux unités sous-jacentes, le premier (T2 ou P2) non terminal, est caractérisé par une inflexion mélodique ascendante, le deuxième (T2 ou P2) terminal, par l'inflexion descendante obligatoire de finalité.

La fonction de la mélodie serait alors triple : elle actualiserait la structure grammaticale générale de la phrase (fonction syntaxique), distinguerait des unités de rang inférieur au sein des groupes de sens (fonction démarcative), tout en indiquant la relation de subordination (fonction syntaxique de "subordination") des éléments de rang inférieur aux éléments de rang supérieur (voir les intonèmes de continuation mineure).

2. Analyse de l'intensité

Dans le tableau 1, il apparait très clairement que sont intensives les voyelles / A + O + E / qui se trouvent à la première syllabe de mot ou de groupe, et les voyelles accentuées. Aucun exemple de voyelle terminale n'a été trouvée lorsque la syllable n'était pas accentuée. C'est significatif. L'examen des voyelles prononcées faiblement apporte une contre épreuve qui confirme amplement nos premières observations. Il semble donc raisonnable de dire que l'intensité, dans la mesure où les conditions phonétiques sont favorables, (voyelles à grande et moyenne aperture) joue un rôle non négligeable dans la segmentation de la phrase en mots

lexicaux puisqu'elle en indique le début, les mots grammaticaux étant d'ailleurs prononcés systématiquement à voix plus faible (N1 ou 2).

Dans la mesure enfin où la syllabe accentuée est en général, en français, la dernière du groupe de sens, l'intensité permet de segmenter la chaine parlée en unités supérieures au mot en élevant régulièrement de quelques décibels les syllabes finales de groupe de sens (fonction constrastive).

3. Analyse de la durée

Le tableau 2 indique que les syllabes finales (de T ou P) accentuées sont longues (78 %). Une contre épreuve a confirmé notre point de vue. On peut donc affirmer que les syllabes finales de groupe, généralement accentuées par ailleurs, sont longues ou plus longues que la syllabe immédiatement précédente. La fonction de la durée est donc complémentaire de celle de l'intensité qui frappe les syllabes initiales, fonction également contrastive.

4. Corrélations entre mélodie et durée

La durée joue également un rôle dans la syntaxe de la phrase dans la mesure où elle permet d'opposer certaines structures. En effet lorsqu'on a une suite de mots appartenant à un groupe complexe du type : " dans une ère nouvelle " ou " sur un napperon " constituant tous deux le prédicat de la phrase, le recours à la mélodie seule est insuffisant pour délimiter les unités syntaxiques.

Dans l'un et l'autre cas, (représentatifs de phénomènes analogues pour d'autres phrases) la préposition et l'article sont, soit maintenus

soit prononcés à un niveau bas, alors que le reste du syntagme (substantif seul ou substantif + adjectif) est caractérisé par une impulsion négative très nette.

Si l'on est autorisé à découper le syntagme prépositionnel " dans une ère nouvelle " en deux éléments mélodiques : " dans un ère " T2 ↗ "nouvelle" P2 ↘, c'est que la voix allonge distinctivement la syllabe unique du substantif, c'est à dire que la durée établit une frontière syntaxique et sémantique entre le substantif et l'adjectif (nous ne nions pas dans le rapport de la durée de 1 à 3, le rôle de la consonne allongeante qu'est le /r/, mais ce rôle n'a pas encore été déterminé avec précision) alors que dans l'autre cas, la fin de la phrase est très progressivement allongée.

Notre recherche future s'orientera en particulier sur une étude plus fine de la fonction syntaxique de la durée.

CONCLUSION :

Cette étude constitue une première étape dans la recherche. Nous nous attacherons ultérieurement à donner une représentation exacte des phénomènes, à les chiffrer exactement, à approfondir enfin les corrélations entre fréquence, intensité et durée.

Dans la mesure où l'intensité caractérise les initiales de mot et les finales accentuées, où la durée caractérise les finales accentuées où la mélodie range dans les hautes et basses fréquences ces syllabes à accent, on peut dire que la durée est le seul paramètre spécifique de l'accent.

D'autre part, notre étude nous amène et nous autorise à poser plusieurs hypothèses que des recherches futures devraient ou non

confirmer : la mélodie actualise d'une manière globale la structure syntaxique de la phrase en indiquant les groupes de sens fondamentaux (T1 ou P1). A l'intérieur de ces groupes, le jeu complémentaire de la fréquence, de l'intensité et de la durée distingue les unités de rang inférieur, plus précisément les mots lexicaux sont identifiés grâce à l'existence des valeurs positives (absolues ou relatives) de leurs paramètres prosodiques, tandis que les mots grammaticaux sont caractérisés par un même comportement paramétrique : descente avec creux significatif des trois courbes de fréquence, intensité et durée.

TABLEAU 1 DUREE

Phrase	accentuée	accentuée	accentuée	non accentuée	Groupe	mot
1	+	+				
2	+	+				
3		+				
4			+			
5		+				
6						
7				+		
8		+			+	
9	+	+				
10	+	+				
11	+	+				
12	+	+				
13	+					
14	+					
15	+					
16	+	+				
17	+	+				
18	+				+	
19	+	+				
20		+	+			+
21		+				+
22	+	+				+
23			+	+		
24		+	+			
25		+				
26	+	+				
27	+	+				
28	+	+				
29	+	+				
30		+	+			
31		+				
32	+	+				
33		+	+			+
34		+				
35						
36	+					
37	+	+				
38	+	+				

Les n° des deux tableaux correspondent aux n° de phrases

TABLEAU 2 INTENSITE

N°	Phrase, groupe et mot	groupe seulement	mot seulement	accentuée	accentuée	phrase accentuée	de mot accentuée
1	PG		+ +	+			
2	P G		+	+		+	
3	P G		+	+		+	
4	P M		+		+		
5			+	+	+		
6				+			
7	P M		+	+			
8			+	+			+
9		+		+		+	
10		+ +	+			+	
11		+ +	+		+		
12	P M	+	+	+	+	+	
13			+	+			+
14			+	+			
15	P G	+ +	+	+		+	
16			+	+			
17			+	+			
18			+		+		
19			+				
20			+	+ +			
21		+	+ +	+			
22		+	+ + +	+			
23	P M				+ +		
24			+ + +				
25	P G		+	+		+	
26			+ +	+			
27		+	+	+			
28				+			
29			+ +	+			
30	P M		+			+	
31				+			
32		+	+	+			
33			+	+			
34			+	+			
35	P G		+	+		+	
36				+		+	
37			+	+			
38				+			*

CORPUS DE L'ETUDE

1 Le premier repeint
2 La première peint
3 Les petits oiseaux chantent
4 Marguerite a un bonnet
5 Il avait un signalement
6 Il attrape tout
7 L'armée forte protégera le pays
8 Il portait un signe allemand
9 Il afait un petit orage
10 Il a dessiné trois petites roues
11 Il a remarqué trois petits trous
12 L'été renouvelle la population du village
13 Nous le ferons petit à petit
14 Depuis un moment, on sent des gouttes
15 Ce que j'ai dessiné, c'est un oeuf
16 En ce moment, il parle d'une autre
17 En ce moment, il parle du nôtre
18 Le travail imposé, on s'en dégoûte
19 Je pêche dans un petit torrent
20 Nous entrons dans une ère nouvelle
21 Nous avons mangé sur un napperon
22 Nous avons mangé sur une nappe ronde
23 Marguerite a un beau nez
24 Il a de bonnes aptitudes
25 Je trouve l'eau très noire
26 Ils sont treize amis

27 Le local est ouvert
28 Le paysage est tout vert
29 Ils sont très amis
30 L'arme est forte et redoutable
31 Il est apte à tout
32 Ce costume, c'est un neuf
33 Nous sommes n'importe où
34 Nous sommes partout
35 Sa toilette est faite
36 Sa toile était faite
37 Ce travail est fait à rebours
38 Cette tâche est faite à rebours

BIBLIOGRAPHIE

(1) DELATTRE Les dix intonations de base du Français

(2) DI CRISTO Recherches sur la structuration prosodique de la phrase française (essai d'analyse phonosyntaxique) 6 ° Journées d'Etude sur la Parole Toulouse mai 75

(3) DUBOIS-LAGANE La nouvelle grammaire du Français La rousse

(4) FAURE G. (1972) Contribution à l'étude de la fonction prédicative de l'intonation Travaux de l'Institut de Phonétique d'Aix I,3-13

(5) GARDE P. L'accent P.U.F.

(6) LEON P. & MARTIN P. (1969) Prolégomènes à l'étude des structures intonatives. Didier Montréal, Paris, Bruxelles

(7) LEON P.R. (1971) Essais de phonostylistique Studia phonetica 4 Didier, Paris

(8) LEON R. Problèmes de l'étude intonative. Bulletion d'audio-phonologie. La mélodie de la parole.

(9) MARTIN P. Intonation et reconnaissance automatique de la structure syntaxique 6° Journées d'Etude sur la Parole Toulouse 75

(10) MAURAND G. (1971) Phonétique et phonologie du parler occitan d'Ambialet . Thèse de Doctorat d'Etat

(11) MAURANDG. (1974) Contribution à l'étude du rôle syntaxique de l'intonation . Grammatica

(12) ROSSI M. & CHAFCOULOFF H. (1972) Les niveaux intonatifs Travaux de l'Institut de Phonétique d'Aix 1,16 7-76

(13) ROSSI M. L'intonation prédicative dans les phrases transformées par permutation (1973) Linguistics 103, 64-94

(14) VAISSIERE J. (1974) Fréquence fondamentale des phrases déclaratives en Français. 5° J. E. P. Orsay mai 74

(15) VAISSIERE J. Caractérisation des variations de la fréquence du du fondamental dans les phrases françaises 6° J. E. P. Toulouse mai 75 39-50

CONTRIBUTION DES ZEROS A LA MODELISATION DU SIGNAL DE PAROLE

C. GUEGUEN et M. MATHIEU

1 - INTRODUCTION

L'analyse du signal peut être avantageusement abordée par des techniques de modélisation. Dans cette approche le signal à analyser est confronté avec la sortie d'un modèle dont les paramètres sont ajustés au sens d'un écart minimal entre le signal de référence et son estimation. Les techniques usuelles considèrent un modèle générateur du type autorégressif (AR) et diverses procédures rapides d'identification ont été proposées /1/. L'extension aux modèles linéaires comportant pôles et zéros est un problème plus délicat car typiquement non-linéaire. Son intérêt est cependant évident car la présence d'un zéro dans le processus générateur du signal nécessite un nombre parfois important de pôles pour son approximation par un modèle autorégressif.

Cet article propose l'élaboration de modèles complets (ARMA) comportant pôles et zéros par un processus itératif. Les opérations requises en ce cas ne sont autres que linéaires car fondées sur l'adaptation de modèles autorégressifs (AR) dont certaines parties sont convenablement inversées (MA). Cette inversion du modèle trouve sa justification dans celle de la matrice d'autocorrélation du signal, problème considéré ici comme fondamental.

2 - MODIFICATION DE L'AUTOCORRELATION PAR UN SYSTEME LINEAIRE

En l'absence de données précises sur l'excitation d'entrée qui, à travers le modèle, produit le signal, on est réduit sur celle-ci à des conjectures. La plus simple et la moins contraignante, en général, consiste à admettre que cette entrée est un bruit (éventuellement coloré). Il convient donc alors d'examiner, dans le cas gaussien le plus souvent retenu, les modifications imposées à l'autocorrélation de l'entrée par le passage dans un modèle linéaire. Cette étude peut être abordée sous l'angle des fonctions de transfert en z ou de leurs traductions matricielles.

2.1 - Modèle autorégressif (AR)

On considère un processus défini par l'équation récurrente (1) liant l'entrée u_n à la sortie s_n :

$$a_o s_n + a_1 s_{n-1} + \ldots + a_p s_{n-p} = u_n \qquad (1)$$

L'application de la transformée en z (conditions initiales nulles) donne :

$$A(z).S(z) = U(z) \quad \text{avec} \quad A(z) = \sum_{i=o}^{p} a_i \, z^{-i}$$

La fonction de transfert correspondante ne possède que des pôles qui seront supposés stables. Considérant les transformées de l'autocorrélation de l'entrée et de la sortie :

$$R(z) = S(z)\, S(z^{-1}) \quad \text{et} \quad W(z) = U(z)u(z^{-1}) \qquad (2)$$

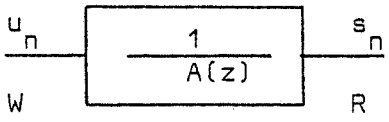
on a directement :

$$W(z) = A(z)\, R(z)\, A(z^{-1}) \qquad (3)$$

Cette équation se traduit de manière matricielle en égalant les coefficients des termes de même degré en z. Introduisant les matrices d'autocorrélation R et W (de Toeplitz) d'éléments r_i et w_i et la matrice A définie par :

$$A^T = \begin{bmatrix} a_o & a_1 & \ldots & a_p & 0 & \ldots & 0 \\ 0 & a_o & \ddots & & a_p & \ddots & 0 \\ 0 & \ddots & \ddots & a_o & & \ddots & a_p \end{bmatrix} \qquad (4)$$

La relation matricielle correspondante s'écrit en supposant toujours pour les matrices les dimensions convenables :



$$A^T R A = W \qquad (5)$$

Cette relation définit implicitement R (matrice d'autocorrélation de la sortie) à partir du modèle $\{a_i\}$ et de W (autocorrélation de l'entrée).

La formule (5) offre une généralisation de la prédiction linéaire usuelle (LPC). En effet, celle-ci ne correspond qu'à la minimisation du terme diagonal de W, c'est-à-dire :

$$\underline{a}^T R \, \underline{a} \qquad \text{avec} \qquad \underline{a}^T = [a_o \ \ldots \ a_p] \quad \text{et} \quad a_o = 1 \qquad (6)$$

La matrice apparaît comme l'autocorrélation de l'erreur de prédiction. Ce résiduel n'est un bruit blanc que dans la mesure où l'ordre du modèle et celui du système coïncident. La formule (5) est, après définition d'une norme matricielle convenable, la source de divers compromis entre l'erreur quadratique moyenne et la blancheur du résidu.

2.2 - Modèle à moyenne mobile (MA)

Un modèle à moyenne mobile est qualifié par une relation de récurrence du type :

$$s_n = b_o\, u_n + b_1\, u_{n-1} + \ldots + b_q\, u_{n-q} \qquad (7)$$

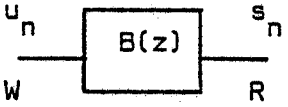
La fonction de transfert en z correspondante ne comprend que des zéros et s'écrit donc :

$$S(z) = B(z)\, U(z)$$

où l'on supposera que les zéros de B(z) sont de modules inférieurs à 1 (phase minimale). Il en résulte que :

$$R(z) = B(z)\, (U(z).U(z^{-1}))\, B(z^{-1}) \qquad (8)$$

Cette relation polynomiale se traduit de façon matricielle en fonction de R et W en introduisant une matrice B^T conformément à (4) :



$$R = B^T\, W\, B \qquad (9)$$

Dans le cas d'une entrée blanche (W=I), la recherche d'un modèle MA pour un R donné s'interprète comme une factorisation matricielle particulière. On note que la corrélation r_i de la sortie est dans ce cas finie puisque $r_i = 0 \quad i > q$.

2.3 - Modèle mixte (ARMA)

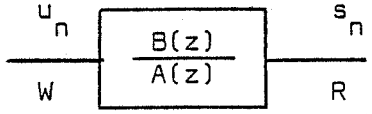
Supposant dans le cas général que la fonction de transfert du modèle comporte à la fois des pôles et des zéros conformes aux deux expressions précédentes, on a :

$$S(z) = \frac{B(z)}{A(z)}\, U(z)$$

Ce qui se traduit par :

$$A'z)\, R(z)\, A(z^{-1}) = B(z)\, U(z)\, U(z^{-1})\, B(z^{-1}) \qquad (10)$$

dont l'homologue matriciel est la relation :



$$A^T R A = B^T W B \qquad (11)$$

Pour une entrée blanche (W=I), la recherche du modèle s'interprètera alors par la détermination de $\{a_i\}$ tels que leur corrélation résiduelle soit explicable par des zéros.

3 - INVERSION DES MATRICES DE TOEPLITZ

3.1 - Importance du problème

L'inversion des matrices de corrélation pourvues de la forme de Toeplitz joue, à notre sens, un rôle central en analyse des séries temporelles. Il est bien connu, en effet, que la détermination des coefficients de prédiction linéaire passe implicitement par le calcul de R^{-1}. Mais, de plus, cette matrice représente la covariance des paramètres estimés. Les exemples sont nombreux où la connaissance de cette matrice est requise. Pour n'en citer que les principaux :

- la matrice R^{-1} du maximum de vraisemblance /2/ n'est, en général, qu'approximée (avec une médiocre précision pour les faibles dimensions) ;
- cette même matrice est celle du gain asymptotique du filtre de Kalman dont l'évaluation fournit un gradient sous optimal peu coûteux et une initialisation judicieuse du filtre.

Mais nous montrerons aussi que cette inversion est, moyennant certaines précautions, la clef du renversement de modèles AR en modèles MA équivalents.

De nombreux algorithmes d'inversion ont été fournis jusqu'à une période récente (Zohar, Trench, ... Akaike). Ils sont basés sur une procédure itérative équivalente à celle utilisée dans la recherche du prédicteur linéaire optimal (Levinson, Robinson, Durbin, Whittle). Une contribution particulièrement intéressante est celle de Rissanen qui sera utilisée au paragraphe 4.1 et qui se fonde sur la décomposition de R en facteurs triangulaires. Markel propose une expression analytique de l'inverse qui a pour inconvénient de faire figurer les prédictions linéaires d'ordres successifs (alors que la dernière est suffisante en théorie).

Mais l'expression la plus séduisante est donnée par Durbin /3/ dans une forme proche de celle ici introduite. Cependant, la formulation est inutilement compliquée puisqu'elle revient à calculer l'inverse de la matrice de corrélation de dimension p à partir d'un prédicteur d'ordre p+1, celui-ci dépendant du coefficient r_{p+1} non contenu dans la matrice.

3.2 - Expression de l'inverse

Soit à déterminer l'inverse d'une matrice R à partir de la connaissance du prédicteur correspondant : $a_o\ a_1\ \ldots\ a_p$ (a_o=1, en général).

Notant par J la matrice dont les éléments de l'anti-diagonale sont des 1, on a

la propriété (permutation des lignes et des colonnes) :

$$R = J\,R\,J \qquad \text{d'où} \qquad R^{-1} = J\,R^{-1}\,J \tag{12}$$

L'inverse est donc, comme R, symétrique et per-symétrique.

Supposant que le prédicteur est optimal et que l'ordre du modèle est le bon (inférence des r_i d'ordre supérieur à p), on a :

$$A^T R\,A = I \qquad \text{et} \qquad A\,A^T R\,A\,A^T = A\,A^T \tag{13}$$

Il en résulte que la pseudo-inverse de AA^T est R. On recherchera donc une expression de R^{-1} dans les termes de cette matrice.

La dimension de A^T étant choisie pour posséder exactement (p+1) lignes, cette matrice peut être partitionnée en deux blocs

$$A^T = [A_o^T,\ A_1^T] \tag{14}$$

On prouve alors par simple vérification les propositions équivalentes suivantes :

$$\begin{aligned} &\text{(i)} && R^{-1} = A_o\,A_o^T - A_1^T\,A_1 \\ &\text{(ii)} && R^{-1} = A_o^T\,A_o - A_1\,A_1^T \\ &\text{(iii)} && \text{Diag}\,(R^{-1},\ -R^{-1}) = A\,A^T - J\,A\,A^T\,J \end{aligned} \tag{15}$$

Les formules donnent donc, en particulier, une solution explicite au calcul de R impliqué en (5) pour W = I.

4 - RENVERSEMENT D'UN MODELE AR en MODELE MA

La détermination d'un modèle à moyenne mobile (7) à partir d'un modèle autorégressif (1) et sa réciproque sont à l'origine un simple problème d'inversion :

$$B(z) = 1/A(z) \qquad \text{ou} \qquad A(z) = 1/B(z)$$

Mais il est évident que l'on ne peut dans la pratique se contenter de cette solution. Bien que A(z) soit stable, B(z) comportera, en général, un grand nombre de termes non négligeables. Dans ces conditions, c'est le processus de réduction de l'ordre qui joue un rôle essentiel.

4.1 - Inversion des facteurs de Cholesky

La matrice de corrélation R peut être décomposée en facteurs triangulaires par des algorithmes rapides donnés par Rissanen et qui généralisent ceux de la prédiction linéaire /4/.

Cette décomposition s'écrit sous la forme :

$$R \begin{bmatrix} a_{op} & o & & o \\ a_{1p} & a_{op-1} & & \\ \vdots & \vdots & \ddots & o \\ a_{pp} & a_{p-1\,p-1} & \cdots & a_{oo} \end{bmatrix} = \begin{bmatrix} b_{op} & b_{1p} & \cdots & b_{pp} \\ o & b_{op-1} & \cdots & b_{p-1\,p-1} \\ & & \ddots & \vdots \\ o & & o & b_{oo} \end{bmatrix}$$

C^{-1} $\qquad$ C^{T}

Elle fait apparaître comme première colonne de C^{-1}, le prédicteur AR d'ordre p et comme première ligne de C^{T}, le modèle MA correspondant. Le modèle MA se construit donc à partir des résidus successifs de la prédiction linéaire et ceci est particulièrement net dans l'algorithme proposé par Le Roux /5/ Il peut aussi s'obtenir par inversion de la matrice triangulaire C^{-1} qui se prête à une formulation récurrente aisée.

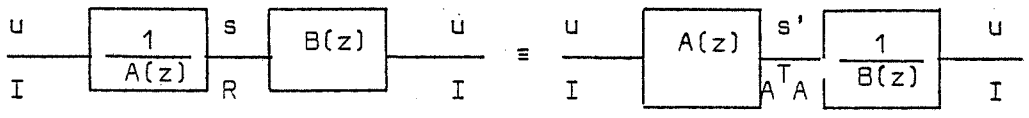
Cependant, une telle technique ne peut être que difficilement utilisée car les ordres des deux modèles sont liés. Si l'approximation d'un modèle autorégressif par un ordre inférieur est aisément maîtrisée, il n'en va pas de même pour un modèle MA.

4.2 - Solution par filtrage inverse

Soit un signal s de covariance R dont on suppose qu'il est la réponse d'un système autorégressif à un bruit blanc, on désire construire le modèle MA approximatif d'ordre fixé tel que :

$$A^{T} R A = I \qquad\qquad R = B^{T} B$$

Pour cela, on considère les deux schémas équivalents obtenus par simple inversion des filtres :



Appliquant les formules (5) et (9), on a :

$$A^{T} R A = I \qquad \text{et} \qquad B^{T} (A^{T}A) B = I$$

Le modèle MA défini par B(z) est donc obtenu par deux prédictions linéaires successives. Le coût d'une telle opération reste cependant limité par l'efficacité des algorithmes actuels /5/. Si la matrice R est donnée a priori le contrôle de la précision s'effectue par l'ordre p du modèle AR tandis que l'ordre q du modèle MA peut encore être choisi.

La procédure présentée rejoint une méthode introduite par Durbin dans un contexte statistique mais avec une approche plus directe. Celle-ci peut être interprétée comme une double "inversion" qui aboutit à l'identité entre R et $(B^{t}B)$.

5 - IDENTIFICATION DE MODELES MIXTES ARMA

5.1 - Difficultés du problème

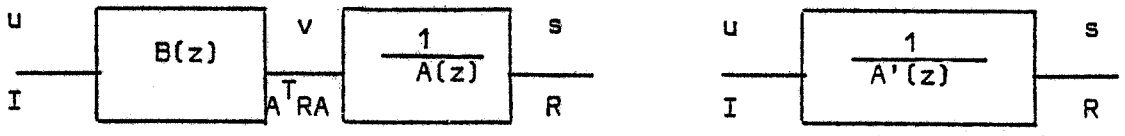
La détermination des paramètres d'un modèle mixte est un problème non linéaire du fait que des quantités inconnues E(z) (bruit) et B(z) (partie MA) interviennent sous forme de produit. Il serait souhaitable, dans ces conditions de pouvoir au moins, séparer les problèmes d'estimation des pôles et des zéros. L'idée, en général retenue, est de profiter de la nature finie de la corrélation par les zéros pour établir une équation de Yule-Walker convenablement décalée vers les indices supérieurs pour éviter la corruption des r_i par les zéros.

Dans cette optique, nous avons introduit une technique ramenant l'équation en question à la forme de Toeplitz en superposant l'équation transposée. Les algorithmes habituels sont ainsi applicables mais l'imprécision sur les r_i d'ordre supérieur entraîne de fréquentes instabilités numériques.

Le processus d'identification retenu est donc une solution itérative en deux temps alternés.

5.2 - Estimation des zéros connaissant les pôles

Le modèle mixte étant décomposé sous la forme d'une cascade AR - MA, la covariance de la variable intermédiaire v est facilement évaluée par la formule (5) pour un A(z) connu.



Le résidu v du filtrage inverse correspondant doit alors être modélisé en un système MA. On utilise, à cet effet, la technique de double inversion proposée en 4.2.

5.3 - Estimation des pôles connaissant les zéros

On suppose disposer d'un modèle AR d'ordre élevé 1 ajusté au signal. La fonction de transfert correspondante 1/A'(z) est donc une bonne approximation de celle du modèle ARMA : B(z)/A(z).

On en déduit que 1/B(z) A'(z) est une bonne approximation de 1/A(z). Ce modèle autorégressif d'ordre (q+1) est donc équivalent à un modèle d'ordre p à déterminer. Nous proposons pour cela /6/ d'utiliser l'algorithme de Durbin "à l'envers" de façon à réduire le modèle d'ordre supérieur à l'ordre p en respectant les corrélations des signaux. Cette démarche se justifie directement par des considérations spectrales /6/.

5.4 - Application de la procédure

Des tests effectués sur des signaux synthétiques à partir d'une estimation initiale de faible qualité ont montré une convergence rapide vers le système ARMA théorique avec un léger biais dans l'estimation des zéros. L'application au signal vocal développée dans les planches 1 à 3 soulève un problème délicat de détermination de l'ordre optimal de chacun des modèles AR et MA. Cependant, il apparaît dans la pratique que les cas sont nombreux où la méthode permet d'obtenir une modélisation spectrale qui, à qualité égale, requiert un nombre plus faible de paramètres que le modèle purement AR correspondant.

REFERENCES

/1/ J. MAKHOUL : Linear prediction : a tutorial review,
Proc. of IEEE, Vol 63, n° 4, April 1975

/2/ F. ITAKURA, S. SAITO : Analysis synthesis telephony based on the maximum likelihood principle,
6 Int. Congr. Acoustics, Paper C 55, August 1968

/3/ J. DURBIN : Efficient estimation of parameters in moving average models,
Biometrika, Vol 46, pp 306-316, 1960

/4/ J. RISSANEN : Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factoring positive matrix polynomials,
Mathematics of computation, Vol 27, n° 121, January 1973

/5/ J. LE ROUX : Optimisation du calcul des coefficients de corrélation partielle,
7ème JOURNEES d'ETUDE sur la PAROLE, Nancy, 1976

/6/ M. MATHIEU : Analyse de l'EEG par prédiction linéaire,
Thèse de Docteur-Ingénieur, Université Paris VI, à paraître

TABLES RONDES

1°) AIDE AUX HANDICAPES

COMPTE RENDU DE LA TABLE RONDE SUR L'AIDE AUX HANDICAPES.

Présidents : MM. LORAND et GUEGUEN
Rapporteur : M. GUEGUEN

M. LORAND

ouvre la table ronde en remerciant les organisateurs et en précisant que les Présidents ne se reconnaissent, malgré quelques études sur le sujet, qu'une compétence limitée pour aborder de manière générale le problème de l'aide aux handicapés. Aucune règle n'est dont préétablie pour la conduite de cette table ronde, dont le succès dépendra essentiellement de la participation active de l'assistance.

Il insiste sur la pertinence des études sur la parole : analyse, synthèse, reconnaissance dans le domaine, en mettant en relief la distinction entre handicapés moteurs et sensoriels. Ce sont ces derniers, que leur handicap, soit d'origine auditive ou visuelle, qui feront l'objet essentiel des discussions.

Il entame un rapide historique de la question en se basant sur un article (non publié) de M. GUEGUEN, qui décrit rapidement l'état de l'art à l'Etranger jusqu'en fin 1974. L'auteur précise que le bilan en question était partiel et très coloré par les orientations de l'article. On décide de dresser un panorama rapide des études en France, menées par les participants et de donner la parole aux utilisateurs pour qu'ils expriment leurs besoins.

1. BILAN DES ETUDES EN FRANCE.

M. LAMOTTE

rappelle les bases du système DIFA (division de fréquences audibles) qui opère une transposition et un tassement du spectre vers les basses fréquences. On effectue, en général, une division par 2, ce qui conserve globalement la forme du spectre (positions relatives des formants). Le but visé est une aide rééducatrice utilisant un retour auditif dans la bande perçue par le malentendant.

Des expériences préliminaires ont été menées à Nancy (Malgrange) sur 4 enfants sourds et 4 enfants témoins. Les déficiences moyennes concernées étaient de 70 à 80db vers 125Hz et plus de 100db à 1000Hz. L'appareil est actuellement en expérimentation plus poussée à l'Hôpital de Nancy (Professeur WAYOFF).

Mme HATON

complète la bibliographie après 1974 en signalant un vaste effort concerté de divers Laboratoires en Suède et aux U.S.A. Elle fait état d'un document bibliographique qui, selon le voeu général, sera intégré au compte rendu. Elle insiste sur l'intérêt des aides tactiles, sens dont certaines caractéristiques se rapprochent de celles du système auditif.

Elle aborde l'exposé du projet SIRENE au Laboratoire d'Informatique de l'Université de Nancy, en classifiant les défauts observés : défauts d'intelligibilité très critiques (articulation, liaison entre sons, respiration, rythme, durée) ; défauts de qualité moins importants (distinction voisé/non voisé, fondamental, ...).

Le projet se donne pour but la rééducation du langage et se distingue par l'usage systématique du calculateur en ligne, usage peu répandu en France dans cette application. Le rôle du calculateur est d'analyser et stocker les formes (établissement d'un fichier de l'enfant), de les confronter avec un système de reconnaissance automatique de la parole, de permettre un auto-apprentissage qui ne minimise pas le rôle important du rééducateur, d'apporter une objectivité totale dans la mesure des performances de l'élève.

Après une étude théorique, le projet est en cours d'implantation. Il a déjà suscité des contacts avec l'Institut des Jeunes Sourds de Paris (Professeur DISSOUBRAY).

M. LAMOTTE

fait état de projets du même type et insiste sur le rôle fondamental de l'orthophoniste. Face aux difficultés d'une représentation visuelle de la parole, il attire l'attention sur l'intérêt des aides tactiles comme complément de la lecture labiale pratiquée couramment par le sourd.

Selon lui, la rééducation doit se baser sur des exercices simples, relatifs à des paramètres du type intensité, mélodie, ..., qui peuvent être représentés visuellement sans effort excessif. Mais bien plus, la représentation doit offrir le moyen d'effectuer un transfert significatif vers la mise en oeuvre des organes phonatoires.

M. PINEL

mentionne une série d'expérimentations utilisant le synthétiseur CIPHON, visant à produire une parole lente sans modifier pour autant les formants et la mélodie. Il précise que la technique s'adresse aux enfants qui manifestent une compréhension lente et que le rythme de l'élocution peut être progressivement accéléré. Un autre intérêt est le caractère simplifié et schématique du signal engendré.

L'efficacité de la parole lente en rééducation est contestée par divers participants et M. LIENARD signale divers autres dispositifs commercialisés produisant le même effet.

M. LIENARD

résume les travaux effectués dans son Laboratoire par M. BART (élève professeur à l'Institut des Jeunes Sourds). L'attention se centre sur l'usage des techniques de reconnaissance de la parole en rééducation. On propose un dispositif de visualisation comparée des séquences émises par le professeur et l'élève sur un écran comportant une cible à atteindre. La réalisation générale d'un tel dispositif pourrait apparaître une gageure étant donnés les caractères très différents des voix des deux protagonistes. Cependant, si l'on se restreint à certains aspects partiels, comme le rythme par exemple, le but peut être atteint. On dispose, en effet, en reconnaissance de la parole de procédures de segmentation qui peuvent être appliquées aux deux élocutions pour déceler les différences et coïncidences, et aboutir à une note de ressemblance. Le système a été réalisé en simulation.

Un autre aspect des travaux menés au L.I.M.S.I. fait appel au caractère normalisé de la voix synthétisée par diphonèmes. On espère ainsi procurer un signal de parole très reproductible, bien adapté à la communication avec les malentendants.

MM. TOUSSIGNANT, TESTON, CARTON :

Sur une question relative à l'état de l'art dans l'analyse des voix de sujets laryngecthomises, certains travaux sont signalés à Aix-en-Provence (4 mémoires) et à Nancy (3 mémoires). Le traitement à ce sujet n'apparaît cependant pas privilégié.

M. LORAND

rappelle des études menées au C.N.E.T. et qui ont déjà fait l'objet de publications (6ièmes Journées du GALF, Toulouse). Sur une idée de P. PIMONOV, le système VENUS opère une transformation de la parole grâce à un vocoder à canaux, décalé vers les fréquences graves. Sept porteuses modulées en amplitude dans la bande 70-300Hz constituent le support d'information. Suite à son expérimentation à Fougères, il est apparu que le dispositif était trop sophistiqué pour se prêter à un apprentissage rapide.

En conséquence, les études se sont orientées vers le système PARME, qui superpose le signal original à des bandes transposées. Les fréquences supérieures à 1000Hz sont ainsi codées sous forme de deux bruits injectés dans les bandes 0 à 500Hz et 500 à 1000Hz. Bien qu'aucun effort de miniaturisation n'ait été réalisé, le système a été expérimenté sous forme d'appareil de classe pendant trois ans, à Fougères. Les résultats sont encourageants mais ne doivent pas cacher la nécessité d'aboutir à un dispositif transportable, utilisable facilement en dehors de la classe.

M. GUEGUEN

décrit le système d'aide visuelle en cours d'élaboration à l'E.N.S.T., et qui se destine principalement à la rééducation, mais aussi éventuellement, à la lecture d'un vocabulaire réduit. Le schéma d'ensemble comporte un bloc d'analyse extrayant du signal divers paramètres, un bloc de réduction effectuant une transformation de coordonnées concernant les paramètres dans un plan, un bloc de visualisation (écran d'oscilloscope ou moniteur TV). Une chaine parallèle de marquage règle l'intensité (intensité de parole, présence de fondamental). Le système ne nécessite la connexion avec un calculateur que lors de l'apprentissage (réalisé une fois pour toutes pour un corpus donné).

Le système, dont l'idée remonte à 1971 (Assises de la prothèse auditive, Paris), est en cours d'expérimentation. Des discriminations types proposées par M. DISSOUBRAY (O, ON, AN par exemple) sont réalisées en temps réel sur l'écran, avec une bonne insensibilité au locuteur. Mais, comme tout appareil de ce genre, la faiblesse essentielle réside dans la pauvreté de la signification physiologique (position des organes phonatoires, etc...) de l'information de retour.

MM. DISSOUBRAY, CAELEN, LIENARD, LORAND :

L'attention est attirée sur les travaux du Professeur CHOUARD (et aussi Mc LEOD et Coll.) qui implantent des électrodes dans la membrane basilaire de sourds profonds pour exciter par des impulsions déduites du signal vocal les terminaisons nerveuses disponibles. M. CAELEN décrit rapide-

ment le dispositif expérimental qui consiste, après ouverture de l'os du rocher, à implanter au niveau des cellules ciliées, 8 électrodes convenablement réparties. Les problèmes d'isolation et de rejet des téflons protecteurs sont d'importance considérable.

Un consensus général, sollicité par MM. LIENARD et LORAND, se fait sur le danger qu'il y a à introduire dans le public l'idée prématurée qu'une technique de ce type pourrait, sans la confirmation d'une longue expérience, révolutionner l'état de l'art.

2. POINT DE VUE DES UTILISATEURS.

M. DISSOUBRAY

est invité, en tant que praticien de l'apprentissage du langage chez les enfants sourds et en tant que personnalité bien au courant des développements techniques récents, à faire valoir le point de vue des utilisateurs et leurs besoins.

Il attire l'attention sur l'importance fondamentale de l'auto-contrôle dans la production de la parole. En cas de défaillance de ce contrôle, un appareillage précoce (3 à 6 mois) se révèle très favorable pour profiter de la période de maturation corticale des aires auditives et la création de circuits synoptiques (jeux lumineux par exemple).

Malgré l'apparition des appareils dont il a été fait mention, les aides techniques efficaces pour cet auto-contrôle sont encore trop rares. Certaines écoles disposent d'appareils en nombre réduit permettant la mesure de paramètres partiels (intensité, mélodie, ...). Mais, même les appareils plus évolués, du type spectroscope (rangée de tubes s'éclairant en fonction de l'énergie dans les bandes de fréquence par exemple), donnent :

- des images compliquées,
- évoluant trop rapidement.

Ces aides ne peuvent être utilisées que dans des cas restreints (syllabe unique).

Les souhaits du praticien dépendent du but recherché : soit apprentissage de l'élocution, soit compréhension de la parole.

a) Les aides d'apprentissage doivent laisser une large part à l'auto-correction physiologique. Le modèle à retenir peut être celui de la reconnaissance de la parole, même si celle-ci ne vise qu'à obtenir de bons taux pour un nombre réduit de mots ou de phonèmes. A ce titre, l'expérience rapportée par M. CARTIER sur des ensembles CVC paraît intéressante. Un ordre de grandeur de 10 à 20 mots reconnus pourrait déjà être satisfaisant. Le système doit pouvoir être utilisé par l'élève, de sa propre initiative, sans présence obligatoire du professeur. L'apprentissage s'exercera, si possible, au niveau du mot faisant intervenir l'auto-contrôle de quelque chose d'utile (même limité) et motivant. Il est, d'autre part, important en complément de faire entendre l'élocution de référence pour bénéficier des restes auditifs du sujet.

L'appareil d'apprentissage peut donc prendre la forme d'un dispositif de reconnaissance automatique utilisant une image au niveau du mot et faisant figurer sous forme écrite le message reconnu (même faux ?).

b) En matière de compréhension de la parole, la lecture labiale est essentielle. Certains malentendants réalisent des performances étonnantes, pourvu que le locuteur possède une bonne articulation (ne pas "parler petit"). Mais certaines discriminations demeurent délicates. Par exemple :

groupes de consonnes :

(PE , BE , ME) (TE , DE , NE) (KE , GUE , RE) ;

groupes de voyelles :

(O , ON , IN) (A , AN) (I , E , AI).

La détection d'un faible nombre d'indices permettrait de lever de nombreuses ambiguités, comme le caractère voisé (V), non voisé (NV), plosif (P), non plosif (NP). On a, par exemple :

	PE	VE	ME
caractères	NV	V	V
	P	P	NP

Il est à noter que les voyelles ont moins d'importance dans l'intelligibilité et que certaines se lisent très bien sur les lèvres.

Dans cet esprit, l'appareil d'UPTON (présenté aux Journées d'Etudes précédentes) est une idée ingénieuse, possédant la complémentarité recherchée avec la lecture labiale.

3. DISCUSSIONS ET CONCLUSIONS.

La discussion s'élargit autour des thèmes précédents et de la disparité entre les dispositifs actuels, leurs prix et les besoins des utilisateurs.

Mme HATON, MM. TESTON, LORAND

engagent une discussion sur les détecteurs de "pitch" du marché et les mérites respectifs des appareils individualisés, tels qu'ils existent en Suède pour la mise en évidence de traits pertinents séparés. Le détecteur de fondamental idéal n'existe encore pas, surtout dans le contexte d'une salle de classe.

Mlle VAYSSIERE

fait part d'une expérience pratique autour d'un programme de BBN, où le caractère ludique et l'esprit de compétition sont exploités pour motiver l'auto-éducation et le contrôle de la voix. Elle signale d'autre part une expérience de perception de parole découpée et répétée syllabe par syllabe. L'impression auditive est ainsi plus clairement perçue (accentuation, longueur, hauteur) que dans la parole continue, même chez les sujets normaux.

M. DREYFUS-GRAF

aborde le problème des handicapés moteurs et de la commande à la voix des prothèses (ici, fauteuil roulant). Il propose, dans ce but, un code plosive-fricative ou le recours au phonocode SOTINA. Devant l'objection soulevée par le caractère artificiel du code, il met en avant certaines similitudes entre les mots d'un vocabulaire naturel restreint et leur interprétation en SOTINA. M. GUEGUEN signale qu'une solution vocale est elle aussi envisagée dans le projet pilote SPARTACUS, patronné par l'I.R.I.A.

M. CARTON

s'intéresse à l'apprentissage des langues étrangères, intonation en particulier. Il fait allusion au travail de MARTIN et LEON à Toronto, qui visualisent sur écran T.V. les courbes comparées maître-élève, mais la technique se révèle peu efficace.

Indépendemment de l'appareil, le principal, c'est-à-dire la recherche pédagogique, demeure à faire. Celle-ci est nécessaire, fort longue et trop souvent négligée. Faute de quoi, l'appareil n'est utile qu'au tout premier stade comme instrument de sensibilisation.

M. TESTON

cite, dans le même cadre, les travaux très contreversés, de SOUVAG et GOUBERINA, et ceux plus systématiques de RENARD et LANDERCY (Toulouse, Litz). Il rejoint complètement M. CARTON sur l'importance de la pédagogie et insiste sur les études systématiques qui sont encore à entreprendre de la part des phonéticiens.

MM. LORAND et DISSOUBRAY

résument les interventions et tirent les premières conclusions, d'où il apparaît qu'un long et patient travail est encore nécessaire pour combler les voeux des praticiens, réussir à donner aux handicapés sourds toute l'aide technique qu'ils méritent, et leur faire bénéficier ainsi des progrès récents des recherches en traitement de la parole.

REEDUCATION DES DEFICIENTS AUDITIFS : REVUE DES TRAVAUX DANS LE MONDE. TABLE RONDE SUR "L'AIDE AUX HANDICAPES".

7èmes J.E.P. GALF Nancy, Mai 1976.

Marie-Christine HATON

Laboratoire d'Informatique
Université de NANCY I.

I- INTRODUCTION

Le développement des aides à la communication avec autrui est un des facteurs indispensables pour la réinsertion sociale des mal-entendants. Les sourds ont aujourd'hui accès à un certain nombre d'appareils qui leur offrent une aide réelle dans les circonstances où l'ouïe joue un rôle important (voir (1) et (2)) : avertissement, enseignement, acquisition de l'information, communication à distance, communication entre personnes... C'est à ce dernier aspect seulement que nous nous intéressons ici.

On fait généralement une distinction entre les aides à la compréhension et les aides à la production de la parole. Les premières doivent plus particulièrement fournir au sujet une information complètant celle apportée par la lecture labiale. Les secondes doivent amener le sujet à s'exprimer de façon intelligible pour son interlocuteur. Les deux problèmes sont en fait liés et, s'il existe des appareils spécifiques, il arrive souvent que les deux types d'aides se confondent.

Ces divers points ont fait l'objet de nombreuses études aux USA, en Suède, au Japon et, depuis peu, en France, particulièrement en liaison avec les recherches sur l'analyse et la reconnaissance de la parole.

Nous tentons ici de faire une revue des résultats obtenus et des travaux actuellement menés dans le monde sur ce sujet.

II- LES AIDES AUDITIVES

On a l'habitude de les classer parmi les aides à la réception de la parole. Il est évident cependant qu'elles ont leur importance dans l'apprentissage de la parole. Nous n'envisageons pas ici les problèmes liés à la médecine et à la chirurgie. Citons simplement deux des travaux les plus récents dans ce domaine sur :

- les implants cochléaires dont l'efficacité pourrait être accrue grâce à l'insertion de douzaine d'électrodes
- la stimulation électrique directe des zones du cerveau correspondant à l'audition qui est expérimentée par un groupe de chercheurs américains.

On peut distinguer deux catégories dans les aides autitives:

1- Les amplificateurs qui accentuent tout ou partie du domaine de fréquences de la parole, généralement portatifs et que nous ne détaillons pas ici. (Certains dispositifs sont spécialement prévus pour la rééducation de la parole comme le "Speech Trainer" développé par SCASE en Norvège (54)).

2- Les "transposeurs de fréquence", dont l'idée a été émise dès les années 50 par L. Pimonow en particulier, qui déplacent vers les fréquences plus basses les fréquences du spectre de la parole. En 1966, Johannson (3) proposait la transposition d'une partie seulement de la bande de fréquence de la parole afin de mettre grossièrement en évidence les consonnes non voisées par superposition sur le spectre non filtré. Cette transposition partielle a aussi été exploitée pour la mise en évidence des fricatives (4). On trouve également cet apport de sons de compensation dans le travail de Lafon.

L'appareil DIFA (diviseur des fréquences audibles) de Lamotte et Vigneron (6) permet de comprimer en bloc vers les basses fréquences la zone des fréquences normalement audibles avec un coefficient de division réglable. Il est expérimenté en classe de rééducation comme le système PARME de Lorand et al. (7) qui opère une compression sélective permettant d'obtenir un niveau sonore plus élevé dans les consonnes plosives.

Dans les cas où les pertes auditives sont supérieures à 80 db dans le domaine de la parole, il est indispensable d'apporter au sujet une contre-réaction autre qu'auditive pour la compréhension et la production de la parole. Cette idée avait été émise dès 1928 par Hudgins (9) (10). Suivant le cas, les signaux acoustiques sont accompagnés de signaux visuels ou tactiles.

III- LES AIDES VISUELLES

Elles permettent la visualisation d'un paramètre ou d'un ensemble de paramètres en vue de la rééducation.

1- Paramètres isolés

- Emission d'un son quelconque : Par exemple le "Teddy Bear" (West Virginia University's Department of Electrical Eng neering) dont les yeux s'allument si l'enfant produit un son.

- Paramètres prosodiques

Martony (11) propose une visualisation instantanée de la de la fréquence fondamentale de la voix par déplacement d'une aiguille sur un cadran ; une lampe s'allume lorsqu'un seuil préfixé est dépassé. Hansen (12) la réalise grâce à un système de stroboscopie optique.

Les variations du pitch dans le temps (mélodie) ont été très souvent visualisées sur écran cathodique. Ces aides sont utilisées pour les problèmes de commande du pitch et aussi pour la rééducation du rythme. Citons dans ce domaine le "F_0 indicator" réalisé à Stockholm (13), l'indicateur dérivé du "Pitch Period Indicator" développé par Martony, Phillips et al. (14), les travaux de Knudsen (15), de Levitt et al. (16) ou de Wood qui, lui, visualise un cepstrogramme (17).

Pour l'intensité des sons :

Valeur instantanée : dès 1947, Pronovost (18) proposait un "Vu mètre" où l'intensité commandait le déplacement d'une aiguille devant un cadran.

Variations dans le temps : ce type de visualisation est utilisé comme indicateur de rythme, d'accentuation, etc... Citons le "Pitch-Intensity Indicator" de Watanabe et Okamura (19) utilisé pour l'entraînement à la stabilité des cordes vocales ou le "Florida" de Holbrook (55).

Rythme :

En 1959, Knudsen (15) proposait un indicateur de rythme associé à un "s - indicateur".

Pickett (22) propose la rééducation du rythme dans l'association consonne-voyelle. La segmentation est fondée sur le fait que la puissance en basses fréquences est faible pour les consonnes par rapport aux voyelles. Barth (23) fait une segmentation phonétique par la recherche des maxima d'une certaine "fonction d'instabilité".

Citons enfin Risberg (42) qui utilise un système voisin de (22) pour l'apprentissage du rythme et de certains sons comme /r roulé/ et /s/.

2- Sons élémentaires

- Spectres à court-terme (distribution instantanée de l'amplitude sur l'échelle des fréquences).

En 1944, le "Visible Speech Translator" visualisait le spectre à court-terme grâce à l'association d'un analyseur à bancs de filtres et d'écrans phosphorescents . R.E. Stark (46) en a expérimenté une version modifiée pour la rééducation des plosives voisées et non voisées qui sont fréquemment l'objet de confusions.

Un des systèmes les plus connus actuellement est le "Lucia Spectrum Indicator" developpé par Risberg au KTH de Stockholm (24) ; l'écran consiste en une matrice de 10 x 20 lampes à incandescence, chacune des 20 colonnes correspondant à une zone de fréquence. On rencontre maintenant des spectroscopes chez un certain nombre de constructeurs.

- Fricatives

Dans les cas de déficience aigue ou totale dans les fréquences élevées, le /s/ est difficile à prononcer -en particulier à distinguer de /ʃ/. Le plus souvent, c'est le taux de passages par zéro du signal temporel qui est calculé pour commander le déplacement d'une aiguille sur un cadran. On peut citer les travaux de Borrild (25), Risberg (24), Guttmann et al. (26).

- Nasalité

Dans le "N-indicator" du KTH (24), la vibration à la surface de l'aile du nez est captée grâce à un microphone de contact et le taux de nasalité est calculé par rapport de l'intensité de ce signal à celle du signal de parole. Une méthode voisine est utilisée par Fletcher (27).

- Fréquences de formants

Dans certaines limites, il est possible d'établir une relation entre la position de la langue et les fréquences des deux premiers formants, F_1 et F_2 et montrer ainsi la position de la langue dans le conduit vocal. Cette visualisation est apparue dans la littérature en 1948, dans les travaux de Potter (28) mais pas en temps réel.

Thomas (29) obtient F_1 et F_2 après filtrage passe-bande dans les zones 250-1000 Hz et 700 -3200 Hz. L'écran est constitué d'une matrice de 12 x 12 lampes au néon.

On retrouve des idées voisines chez Watanabe et Kisu (31) ou Pickett (32).

Kalikow (33) utilise ce type d'aide pour l'apprentissage d'une langue étrangère (l'anglais par des Espagnols).

Boston (34) visualise une ellipse dont les axes sont proportionnels à F_1 et F_2 sur un cadran rond ; l'ellipse figure la bouche d'un personnage.

- Autres visualisations

Pickett et Constam (32) mettent en abscisse et en ordonnée respectivement les fréquences moyennes au-dessous et au-dessus de 1000 Hz. Les points représentatifs des voyelles /i/ /e/ /æ/ /o/ et /u/ forment un demi-cercle. Cette aide fait partie d'un ensemble développé au Gallaudet College.

Crichton et Fallside (36) font appel à l'ordinateur pour visualiser la fonction d'aire du conduit vocal grâce aux techniques de prédiction linéaire.

Gueguen (37) fait suivre une analyse par Vocoder d'une réduction de données par analyse factorielle. Un mot est représenté par une trajectoire dans le plan optimal obtenu.

L'évolution de la distribution d'énergie acoustique suivant les fréquences dans le temps est montrée dans les spectrogrammes. Certains caractères restent relativement invariants, comme par exemple, les différences très nettes entre les fricatives non voisées et les voyelles ou l'occlusion des consonnes plosives. Cependant, la visualisation est difficilement interprétable pour un observateur non spécialisé.

Cohen (41) visualise un son soutenu par un point dans un plan où les abscisses et les ordonnées correspondent respectivement aux taux de composantes sinus et cosinus détectés dans l'enveloppe du spectre.

D'autres systèmes fournissent des figures de Lissajous.

Citons le "Voice Visualizer" développé par Lerner et Fano (38) au MIT = la figure dépend du signal et se trouve être relativement indépendante de la fréquence du fondamental. Schulte (39) qui a expérimenté à Heidelberg ce qu'il nomme des "Phoneme Visualizing Systems" note que ces figures fournissent une information utile sur la durée, le voisement et la friction.

3- Visualisation de plusieurs paramètres

Un des systèmes les plus cités a été réalisé par Upton (40), ingénieur de la "Bell Helicopter" devenu sourd : les "Upton eyeglasses". Un petit analyseur extrait l'information pertinente et la visualise par de petites lumières sur les verres de lunettes. Cette prothèse, qui est plutôt une aide à la compréhension, a été décrite pour la première fois en 1967. Elle est maintenant l'objet de recherches au Gallaudet College où le Dr O. Cornett étudie un autre projet destiné à fournir une indication automatique de l'information non visible sur les lèvres.

Dans le même ordre d'idées, Traunmüller (30) propose une aide visuelle comprenant 10 lampes diodes disposées en arc de cercle autour de la bouche du locuteur qui donnent une indication sur le centre de gravité du spectre et la nasalité. Une aide vibrotactile lui est associée.

En 1968, Pickett et Constam(35) décrivent le "Gallaudet Visual Speech Trainer" qui permet de visualiser à la fois le pitch, l'instant d'attaque des associations consonne-voyelle, le spectre des voyelles, l'intensité seule et en fonction du pitch.

Un appareil qui visualise à peu près les mêmes paramètres est présenté par Borrild dans (25).

Goldberg (44) propose une visualisation qui permet la distinction entre voyelles et consonnes et renseigne sur la durée du phonème et le lieu d'articulation. Chaque phonème est représenté par une barre verticale vers le haut pour les consonnes et vers le bas pour les voyelles. La position en abscisse figure le lieu d'articulation.

Le "VSTA" développé par Stewart et al. (45) permet de montrer tout paramètre pouvant être représenté par une tension variant dans le temps comme l'intensité, la nasalisation, la fonction de voisement, la fréquence du fondamental pour les sons voisés et une fréquence moyenne pour les sons non voisés.

IV- LES AIDES TACTILES

Selon Békésy (42), il existe une certaine similarité entre l'oreille et la peau du point de vue de la perception. Ceci est noté également par Pickett (43), qui précise que cette similitude se rencontre dans l'aptitude à percevoir le rythme et l'évolution des paramètres dans le temps. De plus, le sens de la vue reste disponible, pour la lecture labiale par exemple. Il fait cependant remarquer la capacité limitée de la peau à faire une discrimination fréquentielle pour les fréquences élevées. D'autres auteurs comme GOFF (57) signalent une détérioration rapide au-dessus de 200 Hz alor que l'information linguistique pertinente se situe entre 200 et 3500 Hz environ. D'autres mesures de la résolution temporelle de la peau montre qu'elle est bien inférieure à celle de l'oreille. Békésy trouve que le temps requis par un stimulus tactile pour produire la sensation maximale est bien supérieur à 1s alors qu'il ne serait que de 0,02s pour l'oreille. D'autre part, l'effet met une seconde pour disparaître totalement. En fait la peau est plus sensible à l'effet d'intensité qu'à l'effet de durée.

Un certain nombre de chercheurs ont tenté d'utiliser les

"Vocoders" en aides tactiles. L'idée de base est de moduler l'amplitude d'un jeu de vibrateurs par la sortie des canaux de l'analyseur spectral. Les fréquences des vibrateurs sont choisies dans une gamme de fréquences mieux perceptibles par la peau que celles des filtres. Très souvent les stimuli sont appliqués sur les extrêmités des doigts de la main. Citons, après 1950, les travaux de Lövgren et Nykvist (58), Edmondson (48), Pickett (43), Kringlebotn (47) qui transpose les fréquences du signal de parole dans la zone 0-800 Hz. Comme la plupart des auteurs, ce dernier trouve l'expérience encourageante dans l'identification de quelques mots mais conclut à l'impossibilité quasi-certaine de comprendre le discours continu à partir de ces aides tactiles. On retrouve la même conclusion de la part de Washington et al. du MIT. Dans leur Vocoder tactile appelé "Vocotac" construit en 1956, les sorties de six filtres commandent une matrice à six points de type Braille. Une aide plus complexe (l'analyse fréquentielle est faite à l'aide de petits tuyaux résonants appliqués sur les doigts de la main) a été décrite par Guelke et Huyssen (59). Après un entraînement de courte durée, les sujets étaient capables de distinguer les durées des sons, les glissements de fréquences des diphtongues et la classe des consonnes par rapport à celle des voyelles.

Depuis quelques années, les chercheurs s'orientent vers la création de nouvelles aides où l'information apportée n'est plus de nature spectrale. Parmi ces aides,on peut noter le "Single Vibrator Rhythm Indicator" de Boothroyd (49) ou le "Multivibrator Pitch Indicator" de Willemain et Lee (50). Le "Fonator-System" construit par Siemens et testé par Schulte et son équipe à Heidelberg (51) renseigne en particulier sur les paramètres qui ne peuvent être lus sur les lèvres comme la durée, le voisement, la hauteur, l'arrangement rythmique des sons entre eux. Le vibrateur est attaché au niveau du poignet.

V- REEDUCATION ASSISTEE PAR ORDINATEUR

Le système d'apprentissage d'une langue étrangère de Kalikow et Swets (33) est l'un des premiers systèmes centré autour d'un ordinateur.

En 1973, Nickerson et Stevens (35), (après s'être posé la question : "Can a computer help ?") réalisaient le premier système de rééducation assistée par ordinateur. Quatre algorithmes de jeux sont utilisables pour la rééducation de différents paramètre de la voix. Le système est expérimenté à la "Clarke School for the Deaf", aux USA où sont également developpés un certain nombre d'appareils de rééducation (20), (52), (53).

Le projet SIRENE, qui a été présenté en 1975 (8) propose un certain nombre de procédures de jeux rééducatifs et introduit l'utilisation d'un système automatique de reconnaissance de la parole pour l'appréciation objective des performances de l'élève et l'entraînement à la prononciation des mots d''un vocabulaire limité. L'utilisation d'un ordinateur permet en outre de développer des méthodes d'analyse spécifiques de la voix des sourds (5) (56).

VI- TENDANCES ACTUELLES ET CONCLUSION

Deux points permettent de penser que va s'amenuiser le fossé séparant les non-entendants des normalement entendants :

- l'apparition de l'ordinateur et l'accroissement de l'intérêt pour les problèmes d'analyse, de codage, de reconnaissance et de synthèse de la parole,
- la miniaturisation des circuits, l'abaissement des coûts et l'utilisation des micro-processeurs.

Les progrès réalisés dans le domaine de la "compréhension" de la parole par les ordinateurs peuvent conduire au développement d'aides à la communication de plus en plus sophistiquées. Peut-être même verra-t-on se réaliser un jour le vieux rêve de la machine à écrire phonétique.

Les travaux concernant la compression de l'information et le codage de la parole en vue de sa transmission doivent conduire à une meilleure appréciation objective des défauts de la parole en jouant sur les paramètres relatifs aux facteurs qui en conditionnent la qualité et l'intelligibilité.

Les progrès effectués en synthèse permettront d'apprécier

la contribution de ces facteurs : il est par exemple important de déterminer quel est l'effet des paramètres articulatoires sur les traits prosodiques. Les problèmes principaux dans ce domaine sont d'arriver à une voix "naturelle", de bonne intelligibilité et présentant des caractères suprasegmentaux corrects, intonation et rythme en particulier. Ces problèmes sont précisément ceux que rencontre l'enfant sourd lorsqu'il doit apprendre à commander son propre synthétiseur...

Enfin, les progrès de la miniaturisation permettent de penser que l'on pourra équiper, et de façon économique, l'enfant sourd d'un micro-calculateur de poche. Il reste à progresser dans l'utilisation efficace d'une telle capacité de calcul, dans une collaboration entre chercheurs et enseignants des déficients auditifs, pour la conception des aides à la compréhension et la production de la parole, et comme nous le pensons personnellement, pour le développement d'aides au diagnostic et de méthodes d'évaluation des procédures de rééducation.

BIBLIOGRAPHIE

(1) H. LEVITT and J.R. NELSON, "Experimental Communication aids for the deaf", IEEE Trans. Audio. Electroacoust., Vol. AU-18, pp 2-6, March 1970.

(2) A. BOOTHROYD, "Technology and deafness", Volta Review, Vol 77, N°1, 1975.

(3) B. JOHANNSON, "The use of the transposer for the management of the deaf child", Journal Int. Audiology, Vol 5, pp 362-72, 1966.

(4) N. GUTTMAN et al., "Articulatory training of the deaf using low-frequency surrogate fircatives", J.S.H.R., 13, pp 19-29, 1970.

(5) M.C. HATON et J.P. HATON, "Essai de caractérisation des voix d'enfants sourds par analyse polynomiale de la mélodie", 6èmes JEP, GALF, Toulouse, Mai 1976.

(6) M. LAMOTTE et C. VIGNERON, "Utilisation d'un diviseur de fréquences audibles en éducation orthophonique", 6èmes JEP, GALF, Toulouse, Mai 1975.

(7) P. LORAND et al., "Un système destiné à la rééducation des déficients auditifs profonds : le système P.A.R.M.E.", 6èmes JEP, GALF, Toulouse Mai 1975.

(8) J.P. HATON et M.C. HATON, "SIRENE : un projet de Système Intéractif pour la Rééducation vocale des Enfants Non-Entendants", Rev. Gén. Ens. Déf. Auditifs, pp 203-9, 4ème trimestre 1975.

(9) C.V. HUDGINS, "Visual aids in correction of speech", Volta Review, 37, pp 637-43, 1935.

(10) C.V. HUDGINS, "A study of respiration and speech", Volta Rev., Vol 38, pp 341-7, 1936.

(11) J. MARTONY, "On the correction of the voice pitch level for severely hard of hearing subjects", Amer. Ann. Deaf, 113, pp 195-202, 1968.

(12) V.M. HANSEN, "Speech education with deaf children", UNESCO Seminar held in Denmark, Aug.Sept. 1968.

(13) A. RISBERG, "Visual aids for speech correction", Amer. Ann. Deaf, 113, pp 178-94, 1968.

(14) J. MARTONY, "Visual aids for speech correction : summary of three years' experiences", In G. Fant (Ed.), Speech communication ability and profound deafness. Washington, pp 345-9, 1970.

(15) E.R. KNUDSEN, "Articulation amplifiers and other aids for speech training with deaf children", Nord-Til. Tale og Stemme, 19, pp 149-162, 1959.

(16) H. LEVITT, "Speech processing aids for the deaf : an overview", IEEE Trans. Audio. Electroacoust, Vol AU-21, pp 269-73, June 1973.

(17) M.L. WOOD, "Computer-generated spectrograms and cepstrograms" M.S. Thesis, MIT, June 1971.

(18) W. PRONOVOST, "Visual aids to speech improvement", J.S.H.D., 12, pp 387-91, 1947.

(19) A. WATANABE and H. OKAMURA, "Effect of Speech training by 2. dimensional pitch-intensity indicator on deaf children", J.A.S. of Japan, 31, N°3, pp 179-88, Mar. 1975.

(20) A. BOOTHROYD et al., "Use of a computer-based system of speech training aids for deaf persons", Volta Rev., vol 77, N°3, 1975.

(21) A. RISBERG, "A critical review of work on speech analyzing hearing aids", IEEE Trans. Audio Electroacoust., Vol AU-17, pp 290-7, Dec. 1969.

(22) J.M. PICKETT, "Recent research on speech-analyzing aids for the deaf", IEEE Trans. Audio Electroacoust., AU 16, pp 227-34, 1968.

(23) S. BARTH, "Application des procédés de reconnaissance automatique de la parole à l'aide aux déficients auditifs profonds", Thèse I.N.J.S., 1975.

(24) A. RISBERG, "Some comments on the development of new technical aids for the deaf", 6èmes J.E.P., Galf, Toulouse, Mai 1975.

(25) K. BORRILD, "Experience with the design and use of technical aids for the training of deaf and hard of hearing children", Amer. Ann. Deaf, 113, pp 168-77, 1968.

(26) N. GUTTMANN et al., "Articulatory training of the deaf using low-frequency surrogate fricatives", J.S.H.R., 13, pp 19-29, 1970.

(27) S.G. FLETCHER and D.A. DALY, "Nasalance in utterances of hearing-impaired speakers", J. of Comm. Dis., 9, pp 63-73, 1976.

(28) R.G. POTTER and G.E. PETERSON, "The perception of vowels and their movements" J.A.S.A., 20, pp. 528-35, 1948.

(29) I.B. THOMAS et al., "Articulation training through visual speech patterns", Volta Rev., 72, pp 310-8, 1970.

(30) H. TRAUNMULLER, "Lipreading aid tested in continuous speech", Speech Transmission Lab., Quaterly report, 4/1974.

(31) A. WATANABE and S. KISU, "Articulatory trainer for vowels by inverse filter", Faculty of Engineering, Kumamoto University, Feb 1975.

(32) J.M. PICKETT and A. CONSTAM, "A visual speech trainer with simplified indication of vowel spectrum", Amer. Ann. Deaf, 113, pp 253-8, 1968.

(33) D.N. KALIKOW and J.A. SWETS, "Experiments with computer-controlled displays in second-language learning, IEEE Trans. Audio Electroacoust., 20, N°1, pp 23-8, 1972.

(34) D.W. BOSTON, "Synthetic facial communication", Brit. J. of Audiology, 7, pp 95-101, 1973.

(35) R.A. NICKERSON and K.M. STEVENS, "Teaching speech to the deaf : can a computer help ?", I.E.E.E. Trans. Audio Electroacoust., 21, pp 445-55, Oct. 1973.

(36) R.G. CRICHTON and F. FALLSIDE, "The development of a deaf speech training aid using linear prediction analysis", S.C.S. Stockholm, Aug. 1974.

(37) C.J. GUEGUEN, "Un système de transcription visuelle de la parole", Publication int. à l'ENST, Sept. 1974.

(38) R.M. LERNER and R.M. FANO, "Vocoder", Prog. Rep., RLE, M.I.T.,1952.

(39) K. SCHULTE, "Experimental comparison between three oscilloscope phonemes visualizing systems", In G. Fant (Ed.), Washington, pp 355-9, 1970.

(40) H.W. UPTON, "Wearable eyeglass speechreading aid", Amer. Ann. Deaf, 113, pp 222-9, 1968.

(41) M.L. COHEN, "The ADL sustained phoneme analyzer", Am. Ann. of the Deaf, 113, pp 247-52, 1968.

(42) G. BEKESY, "Similarities between hearing and skin sensations", Psychological Review, 55, pp 1-22, 1959.

(43) J.M. PICKETT, "Tactual Communication of speech sounds to the deaf : comparison with lipreading", J.S.H.R., 28, pp 315-30, 1963.

(44) A.J. GOLDBERG, "Visual feature indicator for the severely hard of hearing" IEEE Tr. A.U., 20, n°1, pp. 16-23, 1972.

(45) L.C. STEWART et al., "The VSTA : An approach to the speech training instrumentation problem", Carnahan Conf. on Electronic Prosthetics, Lexington, Kentucky, Sept. 1973.

(46) R.E. STARK, "Teaching/ba/and/pa/to deaf children by means of visual speech displays", 79th Meeting A.S.A., Philadelphia, April 1970.

(47) M. KRINGLEBOTN, "Experiments with some visual and vibrotactile aids for the deaf", Amer. Ann. of the Deaf, 113, 2, pp 311-7, 1968.

(48) W.H. EDMONDSON, "Preliminary results with a new vibrotactile speech training aids for the deaf", S.C.S., Stockholm, Aug.1974.

(49) A. BOOTHROYD, "Sensory aids research project-Clarke School for the Deaf", in G. Fant ed. "Speech communication ability and profound deafness", A.G. Bell Assoc. for the deaf, 1972.

(50) T.R. WILLEMAIN et F.F. LEE, "Tactile pitch feedback for deaf speakers", Volta Rev., 73, pp 541-54, 1971.

(51) K. SCHULTE, "The Fonator-System- A vibroacoustic aid for speech and communication", Technical Paper. Päd Hochschule, Heidelberg, R.F.A.

(52) P. ARCHAMBAULT et al., "Instrumental aids to speech diagnosis, development, remediation and drill", 47th meeting of the Conventio of american instructors of the deaf, Greensboro, 1975.

(53) A. BOOTHROYD and M. DAMASHEK, "Development of small speech training aids", Clarke School for the Deaf, Northampton Mass., Feb. 1976.

(54) Speech trainer model 711M, Technical Paper, SCASE, Bergen, Norway.

(55) A. HOLBROOK, "Modification of speech behavior with preschool deaf children by means of spectrum control", AOEHI Bulletin, 1971.

(56) M.C. HATON et J.P. HATON, "Une méthode de représentation du signal vocal en base adaptative", 7èmes J.E.P. GALF, Nancy, Mai 1976.

(57) G.D. GOFF, "Differential discrimination of frequency of cutaneous mechanical vibration", J. of Experimental Psychology, 74, pp 294-9, 1967.

(58) A. LOVGREN and O.NYKVIST, "Speech transmission and speech training for the deaf child by visual and tactual means using special devices" (en suédois), Nordisk Tidskrift Dövundervvisn, pp 122-43, 1959.

(59) R.W. GUELKE and R.M.J. HUYSSEN, "Development of apparatus for the analysis of sound by the sense of touch", JASA, 31, pp 799-809, 1959.

2°) TRANSCRIPTION GRAPHEME-PHONEME

TABLE RONDE

TRANSCRIPTION GRAPHEME-PHONEME

Jacques GENIN

CNET-ETA

On a abordé les points suivants:

. les applications de la transcription graphème-phonème;
. les problèmes posés, les performances des réalisations actuelles.

Les applications.

Il est facile de trouver des applications à la synthèse de la parole. Celles qui nécessitent une transcription phonétique automatique d'un texte orthographique sont plus difficiles à cerner.

STEPHAN cite comme devant être opérationnelle en 1978/1980 la lecture de lignes d'annuaire, ceci dans le cadre de la modernisation des centres de renseignement téléphonique. Nombre de problèmes spécifiques sont posés ici pour les noms propres, les abréviations, les sigles, qui ne font l'objet d'aucune règle logique bien définie. L'automatisme total et infaillible est impossible. Il faudra sans tolérer des erreurs.

LONGCHAMP cite comme devant être opérationnelle en 1978/1980 la lecture des lignes d'annuaire, ceci dans le cadre de la modernisation des centres de renseignements téléphoniques. Nombre de problèmes spécifiques sont posés ici pour les noms propres, les abréviations, les sigles qui ne font l'objet d'a

LONGCHAMP cite la lecture pour les handicapés. C'est aussi le but annoncé par le Pr. ALLEN au M.I.T. (On n'oubliera pas qu'une telle application suppose résolu le problème de la lecture optique.) Ce sujet ne semble pas, en France, recueillir un intérêt très marqué.

CASTAN cite les applications dans un contexte informatique. L'aide de pupitrage pour ordinateur, si elle se contente d'un vocabulaire limité, ne ressort pas de ce chapitre. Par contre le format P (ou S ?) du FORTRAN permettant l'édition vocale de tout texte paraît plus intéressant.

RODET cite l'adaptation des systèmes informatiques aux non-voyants. Une firme américaine (A.S.I.) a mis en place un tel système, appelé à l'aide d'un clavier téléphonique, qui répond oralement. Cependant, il semble que dans cette affaire la partie synthèse de parole utilise des méthodes rudimentaires.

On cite également:
l'ensiegnement assisté par ordinateur;
les recherches linguistiques sur l'écrit, quand elles demandent à traiter un corpus important;
la reconnaissance de la parole;
la pédagogie de la lecture.

Dans le champ plus général des applications de la réponse vocale, on retiendra comme nécessitant une transcription phonétique automatique celles

portant sur des messages très évolutifs ou sur un vocabulaire très étendu.

On ne négligera pas les applications sortant du cadre de la réponse vocale: reconnaissance de la parole, recherches linguistiques.

Problèmes posés, performances accessibles.

Les équipes représentées qui ont effectivement travaillé sur le sujet sont:

L.I.M.S.I.	(TEIL)
I.U.T LANNION	(DIVAY, GUYOMARD)
E.N.S.E.R.G.	(GUERIN)
I.B.M.	(NEMETH)
E.N.S.T.	(MICLET, LE ROUX)
C.N.R.S. linguistique	(Mme CATACH)
C.E.R.F.I.A.	(CASTAN, PERENNOU)

Par ailleurs plusieurs participants à la table ronde ont une bonne connaissance des études entreprises à l'étranger sur le sujet.

-Dictionnaire phonétique ou règles.

Au premier niveau de la transcription il semble y avoir une concurrence entre l'application de règles et l'exploitation d'un dictionnaire phonétique. La langue française semble se contenter raisonnablement de la première solution. L'adjonction d'un dictionnaire contenant les mots les plus courants permettrait peut-être d'améliorer les performances.

Par ailleurs il paraît impossible, du moins dans certaines applications, de limiter de quelque façon que ce soit l'étendue du vocabulaire. Un système universel devrait traiter quelque 10.000 à 40.000 mots.

Mme CATACH signale que 184 mots couvrent 60% de n'importe quel texte. Il faut sans doute séparer le langage général et le vocabulaire spécialisé.

Il ressort de la discussion qu'un jeu de règles est indispensable (sauf application sur un vocabulaire réduit, pour lequel le problème de la transcription ne se pose pas.), un dictionnaire permettant de gagner du temps pour les mots les plus fréquents.

Il semble que le problème de l'élision du /ə/ obéisse également à des règles.

-Les liaisons.

Si le type de liaison à effectuer découle de l'utilisation de règles de réécriture, la décision de faire ou ne pas faire cette liaison implique, elle, l'utilisation d'un lexique contenant les formes fléchies des verbes, ou une analyse syntaxique permettant de reconnaître dans le texte les mots déterminants et déterminés. Ceci ne s'applique qu'aux seules liaisons obligatoires.

Les syntagmes figés tels que "mot à mot", "vis à vis", ne répondront à aucune règle générale. Ici le recours au lexique est nécessaire.

Mme CATACH signale que les liaisons adjectif-nom ne se font qu'avec les 50 adjectifs les plus fréquents. Ceci renforce l'utilité du lexique.

A l'E.N.S.T. on a choisi la règle simple de faire la liaison après les mots courts, les considérant comme généralement grammaticaux.

-Les homographes.

Nombre de règles mal connues semblent résoudre les problèmes de mots tels que "président", "précèdent"... Ces règles ne s'appliquent en fait qu'aux homographies de désinences; les vrais mots homographes nécessitent une analyse syntaxique ("couvent") ou sémantique ("fils"). On notera que les cas litigieux sont ceux où la syllabe précédent comporte une voyelle qui ne prend pas d'accent ("i", "a", "o", "u" parfois aussi la consonne "r").

-La prosodie.

Il faut obtenir en plus de la séquence de phonèmes une information permettant d'obtenir une parole intelligible et de bonne qualité. Le C.N.E.T. a montré une méthode de synthèse utilisant divers marqueurs de nature syntaxique tels que:

type de phrase (énonciative, interrogative...),
fin du syntagme nominal sujet,
fin du syntagme verbal,
fin des divers groupes de sens,
localisation du mot interrogatif (éventuellement),
marques usuelles de ponctuation (, . ;).

Ces marqueurs sont actuellement donnés "à la main". Une analyse syntaxique ou sémantique pourra-t-elle fournir ces informations de façon automatique?

Diverses approches théoriques ont été faites de ce problème (Ph. MARTIN, J. VAISSIERE), mais actuellement il n'existe pas de programme d'analyse syntaxique opérationnel qui résolve la question.