Groupement des Acousticiens de Langue Française

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28-30 Mai 1975

AVEC LA PARTICIPATION

de l'A.F.C.E.T de l'I.R.I.A et du C.N.R.S

VOLUME I

Textes des Exposés

Organisées par
le LABORATOIRE C.E.R.F.I.A
(U.E.R Informatique)
à l'Université PAUL SABATIER
de TOULOUSE

Nous tenons à exprimer nos plus vifs remerciements à Monsieur le Président de l'Université Paul Sabatier, le Professeur L. LARENG, pour l'aide qu'il nous a fournie.

Pour le Groupe de la Communication Parlée.

Le Président, M. WAJKKOP

Comité Organisateur des Journées

MM. J. CAELEN
S. CASTAN
P.Y. CAZENAVE
D. DOURS
R. FACCA
G. PERENNOU

Groupement des Acousticiens de Langue Française

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28-30 Mai 1975

AVEC LA PARTICIPATION

de l'A.F.C.E.T de l'I.R.I.A et du C.N.R.S

VOLUME I

Textes des Exposés

Organisées par

le LABORATOIRE C.E.R.F.I.A

(U.E.R Informatique)

à l'Université PAUL SABATIER

de TOULOUSE

Les Sixièmes Journées d'Etude sur la Parole, ont été organisées sous le patronage du Groupement des Acousticiens de Langue Française, avec la participation de l'A.F.C.E.T., de l'I.R.I.A. et du C.N.R.S. par le Laboratoire C.E.R.F.I.A. (U.E.R. Informatique) à l'Université Paul Sabatier à Toulouse, les 28, 29 et 30 mai 1975.

Les thèmes de travail fixés pour cette rencontre étaient les suivants :

- Thème I

  A - Description prosodique

  B - Rôle de la prosodie en reconnaissance

  C - Rôle des contraintes sémantiques phonologiques, contextuelles.

- Thème II

  - Analyse et perception.

- Thème III

  A - Aide aux handicapés

  B - Sortie parlée

  C - Commande de processus.

Les exposés sont classés dans quatre catégories : les conférences plénières, prononcées par des personnalités invitées, les communications ordinaires, portant sur les thèmes IA, IC, II ; les exposés de synthèse portant sur les thèmes IB et III, et les communications de dernière minute. L'ensemble est suivi de trois tables rondes portant respectivement sur :

- segmentation en vue de la reconnaissance en continu
- simulation du fonctionnement de l'oreille
- analyse et détection du voisement

Les comptes-rendus se composent de deux volumes. Le premier comprend la majeure partie des communications ordinaires, et les communations concernant les exposés de synthèse. Le second comprend les textes des conférences plénières, les exposés de synthèse, les textes des communications de dernière minute, les discussions et les comptes-rendus des tables rondes.

THEME 1A

DESCRIPTION PROSODIQUE

UNE METHODE SYNCHRONE D'EXTRACTION EN TEMPS REEL DU FONDAMENTAL

J. LE ROUX

1 - INTRODUCTION

La méthode d'extraction est fondée sur l'analyse de la stabilité des systèmes linéaires du deuxième ordre.

Trois inégalités indiquent si trois échantillons successifs du signal peuvent être la sortie d'un système linéaire du deuxième ordre à entrée nulle. Sinon, il y a détection d'une entrée (bruit ou impulsion du fondamental).

Parallèlement, un modèle adaptatif du deuxième ordre détermine à chaque instant quel est le meilleur système du deuxième ordre ayant pour sortie le signal. L'erreur de prédiction entre la sortie du modèle et le signal réel au moment de la détection d'une entrée donne l'amplitude de cette entrée.

La comparaison de ces amplitudes permet d'éliminer les fausses détections éventuelles. La régularité des intervalles de temps séparant les détections permet de décider si elles sont dûes à un bruit ou à l'impulsion du fondamental.

La méthode s'oppose à celles utilisant la simple erreur de prédiction /1/ par la mise en oeuvre d'un critère portant sur le système. Un tel critère se révèle plus performant et ne peut être accessible qu'à une méthode récursive d'adaptation /2/. L'acuité du critère permet d'utiliser une prédiction linéaire du second ordre qui se prête à une technique d'adaptation particulièrement simple et rapide (temps réel). De plus, le critère se réduit dans ce cas à exprimer la stabilité du système, propriété qui peut être testée sur les valeurs sucessives des échantillons du signal.

2 - STABILITE DES SYSTEMES LINEAIRES DU DEUXIEME ORDRE

Pour un système linéaire du deuxième ordre, les échantillons successifs du signal s_n , s_{n-1} et s_{n-2} vérifient la relation

$$s_n = as_{n-1} + bs_{n-2} + e_n \qquad (1)$$

e_n étant l'entrée, la fonction de transfert associée est :

$$F(z) = \frac{1}{1 - az^{-1} - bz^{-2}} \qquad (2)$$

Les conditions de stabilité de ce filtre (pôles de la fonction de transfert à l'intérieur du cercle unité sont /3/ :

$$b > -1 \qquad (3)$$
$$b < 1 - a$$
$$b < 1 + a$$

Un système est stable si et seulement si le point de coordonnées (a, b) est intérieur au triangle ABC (fig. 1a)

Soient s_{n-2} , s_{n-1} , s_n , trois échantillons successifs du signal. s_n ne peut être la sortie d'un système linéaire du deuxième ordre, l'entrée e_n étant nulle seulement si la droite (D)

$$s_n - as_{n-1} - bs_{n-2} = 0 \qquad (4)$$

coupe le triangle ABC (fig. 1a et 1b).

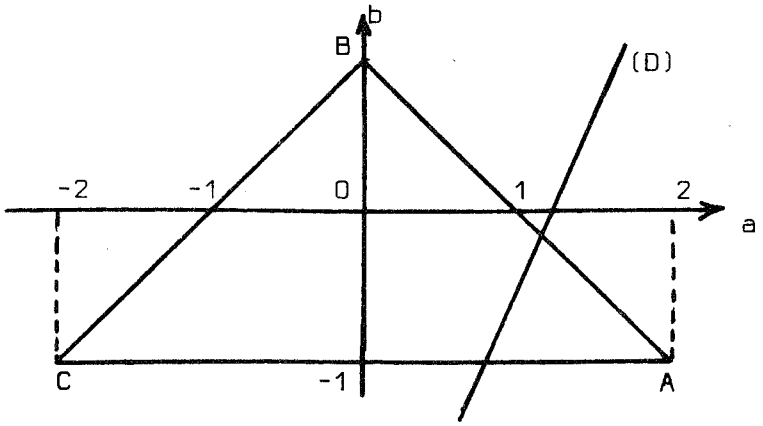


Fig. 1a - POSSIBILITE DE STABILITE AVEC ENTREE NULLE

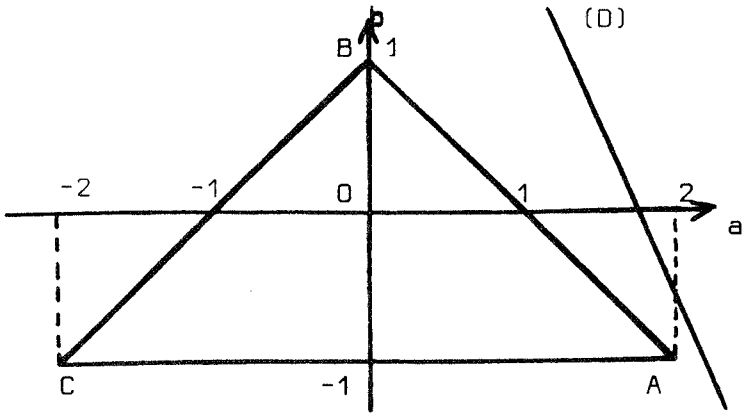


Fig. 1b - DETECTION D'ENTREE OU D'INSTABILITE

La détection d'un système instable ou d'une entrée se ramène aux inégalités :

$$|s_n| > |s_{n-2}| \tag{5a}$$

$$|s_n + s_{n-2}| > 2\,|s_{n-1}| \tag{5b}$$

3 - IDENTIFICATION DES SYSTEMES LINEAIRES EVOLUANT AU COURS DU TEMPS PAR UNE METHODE DU GRADIENT ADAPTATIF

On cherche à identifier un système linéaire du deuxième ordre vérifiant l'équation (1) sur un signal donné.

Si à l'échantillon n-1, le modèle est représenté par les paramètres a_{n-1} et b_{n-1}, l'estimation de s_n est (s_n, s_{n-1} et s_{n-2} étant les mesures réelles) :

$$\hat{s} = a_{n-1}\, s_{n-1} + b_{n-1}\, s_{n-2} \tag{6}$$

L'erreur de prédiction est :

$$\varepsilon_n = s_n - \hat{s}_n \tag{7}$$

La correction des paramètres a_{n-1} et b_{n-1} par la méthode du gradient peut prendre la forme /4/ :

$$a_n = a_{n-1} + \frac{\varepsilon_n\, s_{n-1}}{s_{n-1}^2 + s_{n-2}^2 + \sigma^2} \tag{8a}$$

$$b_n = b_{n-1} + \frac{\varepsilon_n\, s_{n-2}}{s_{n-2}^2 + s_{n-2}^2 + \sigma^2} \tag{8b}$$

où σ^2 est un coefficient permettant de diminuer l'effet du bruit.

A chaque instant, on calcule, dans l'espace des paramètres a, b, le modèle le plus proche du précédent vérifiant la nouvelle série de mesures. Cette correction s'interprète géométriquement sur la figure 2 par la projection du point représentatif du système à l'instant (n-1) sur la droite D_n d'équation :

$$as_{n-1} + bs_{n-2} = s_n$$

(Voir la figure 2 en page suivante)

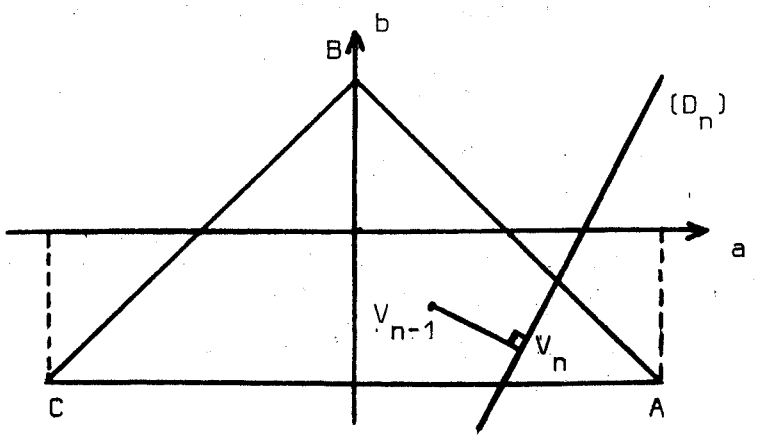


Fig. 2 - CORRECTION DU VECTEUR V_n DE COMPOSANTS (a_n, b_n)

Si l'évolution des paramètres du modèle représentant le signal est suffisamment lente -ce qui est le cas pour les phonèmes autres que les plosives- les paramètres calculés de cette manière arrivent à suivre cette évolution.

On a ainsi à chaque instant les paramètres du filtre du deuxième ordre représentant le système et l'erreur de prédiction ε_n.

4 - APPLICATION DE LA DETECTION DU FONDAMENTAL DANS LE SIGNAL VOCAL

Le signal est filtré dans la bande 300 Hz-700 Hz et échantillonné à 2KHz. Le filtrage donne un signal dans la bande de fréquences du premier formant. Il a l'allure de la réponse à une suite d'impulsions d'un système du deuxième ordre variant au cours du temps (sons voisés) (fig. 3a).

C'est sur ce signal qu'est appliquée la méthode précédente.

L'impulsion du fondamental étant toujours dans le même sens (signal électrique négatif pour nos expériences), on rajoute aux conditions d'instabilité la condition :

$$s_n < 0 \qquad (9)$$

et l'on ne considère donc que le cas où l'erreur de prédiction ε_n définie en (7) est négative. On peut considérer qu'alors l'erreur de prédiction est l'entrée e_n du système définie en (1)

$$\varepsilon_n < 0 \qquad (10)$$

Les 4 conditions de détection du fondamental sont (5), (9), (10) :

$$s_n < 0 \qquad (11a)$$

$$\varepsilon_n < 0 \qquad (11b)$$

$$|s_n| > |s_{n-2}| \qquad (11c)$$

$$|s_n + s_{n-2}| > 2|s_{n-1}| \qquad (11d)$$

Il arrive que cet ensemble de conditions détermine deux impulsions au cours d'une période fondamentale. Mais pour la seconde impulsion détectée, l'amplitude $|\varepsilon_n|$ de l'erreur de prédiction est nettement plus petite (inférieure à la moitié de l'amplitude principale), ce qui n'est pas le cas lorsque l'amplitude des impulsions principales varie avec la puissance du signal. Cette remarque permet donc d'éliminer ces fausses détections.

La séparation des signaux bruités (fricatives ou silences) des signaux voisés se fait par un test de régularité portant sur les intervalles de temps séparant les détections d'entrée.

Si l'on a détection d'instabilité aux instants t_{i-1} , t_i , t_{i+1} , t_{i+2} , on décidera que le signal est voisé si :

$$0,9(t_i - t_{i-1}) < (t_{i+1} - t_i) < 1,1(t_i - t_{i-1}) \qquad (12a)$$

ou

$$0,9(t_{i+2} - t_{i+1}) < (t_{i+1} - t_i) < 1,1(t_{i+2} - t_{i+1}) \qquad (12b)$$

Remarques

1- Les détections ainsi obtenues sont synchrones et en bon accord avec celles de la méthode proposée par A. MAISSIS /5/.

2- Pour obtenir une meilleure précision dans la mesure de la période du fondamental, on échantillonne le signal à une fréquence plus élevée (par exemple : 10 KHz) et on applique l'algorithme aux échantillons s_n , s_{n-5} et s_{n-10} . On conserve comme instant de détection celui où l'erreur de prédiction a la plus grande amplitude à l'intérieur d'une fenêtre de 10 points.

3- L'amplitude de l'impulsion détectée est $|\varepsilon_n|$. On a ainsi une mesure précise de l'évolution instantanée de l'amplitude du signal.

5 - RESULTATS

La figure 3 montre sur une partie de phrase le signal filtré (3a), et les impulsions (avec leur amplitude) du fondamental détectées.

La figure 4 montre l'évolution de la période fondamental détectée par cet algorithme sur une phrase entière (l'absence de tracé correspond à l'absence de fondamental).

Pour les phrases que nous avons étudiées le taux de détection est élevé pour un locuteur masculin : plus de 95 % et souvent 100 % de détection correcte (aucune impulsion manquante et aucune supplémentaire).

La méthode ne permet pas d'obtenir de résultats pour certains phonèmes prononcés par une voix aigüe lorsque la fréquence du fondamental se confond avec celle du premier formant. Dans ce cas, le signal filtré a une allure sinusoïdale, sa période étant celle du fondamental.

L'algorithme ne détecte alors aucune impulsion (par opposition au bruit où il en détecte de façon irrégulière).

Dans ce cas, la période du fondamental est le temps qui sépare deux passages du signal d'une valeur positive ou nulle à une valeur négative.

L'algorithme demande peu de calculs (on peut d'ailleurs simplifier la méthode d'identification du modèle). Il a été programmé sur un calculateur hybride PACER 500. Il permet d'obtenir à la fois une détection synchrone en temps réel du début des périodes fondamentales avec un taux d'erreur minime, et une estimation instantanée de la puissance du signal.

BIBLIOGRAPHIE

/1/ J. N. MAKSYM : Real time pitch period extraction by adaptative prediction of the speech wave form

Conference on speech communication and processing, Newton (Massachusetts), 1972

/2/ C. GUEGUEN, G. CARAYANNIS : Analyse de la parole par filtrage optimal de Kalman

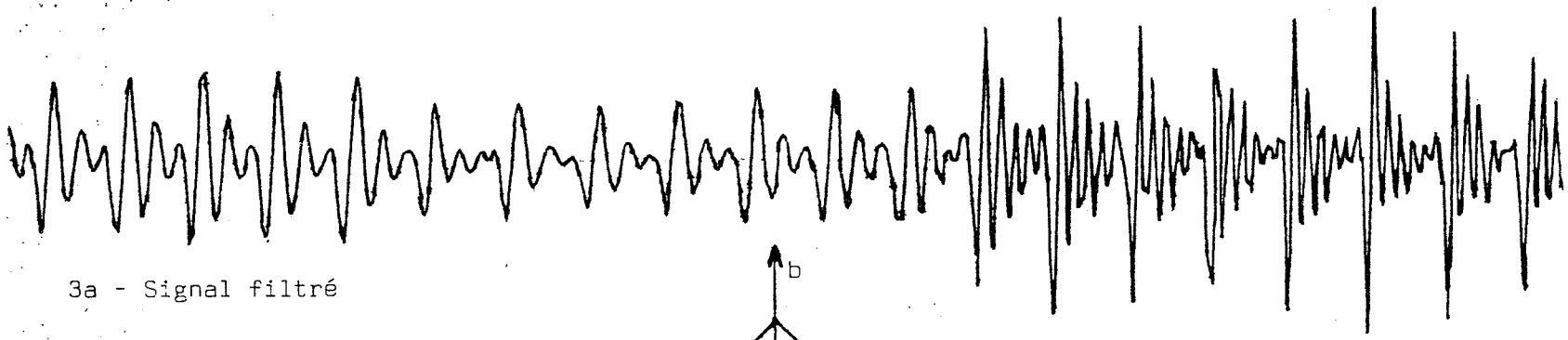
Automatisme (Dunod Ed.), Tome 18, n° 3, 1973

/3/ P. MERON, A. A. SEKEY, E. ZEHEB : Design method for stable second order digital filters

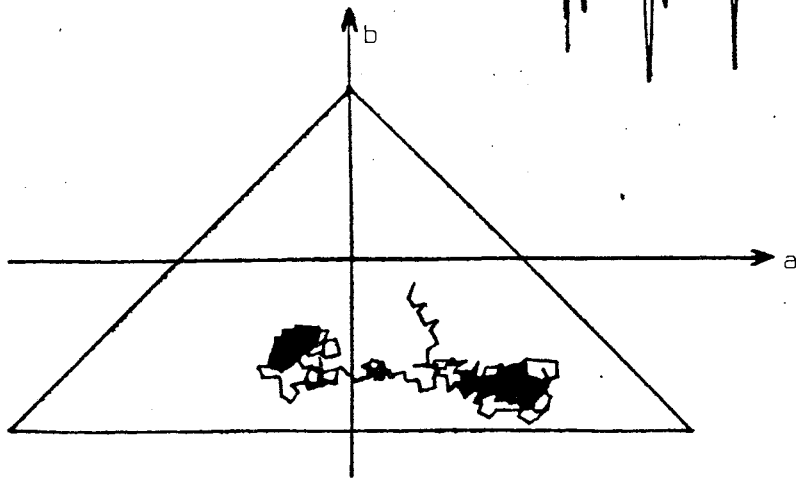
IEEE Trans. on Acoustics, Speech and Signal Processing, Vol ASSP-22, n° 3, June 1974

/4/ Y. Z. TSYPKIN : Adaptation and learning in automatic systems
Academic Press (N.Y. and London), 1971

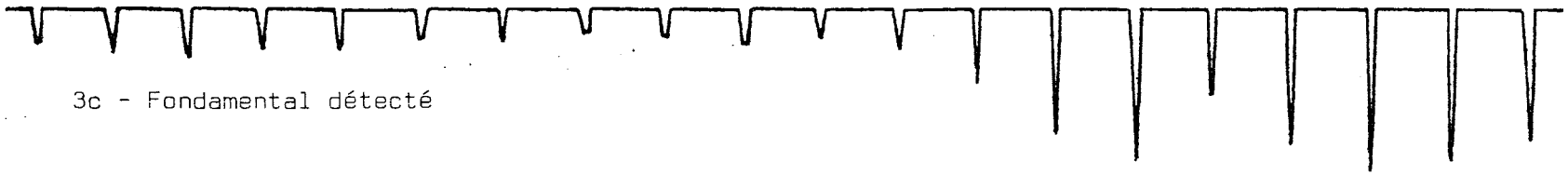
/5/ A. MAISSIS : Une méthode d'extraction du fondamental
L'Onde Electrique, Vol 53, Fasc. 3, mars 1973



3a - Signal filtré



3b - Evolution des paramètres du modèle



3c - Fondamental détecté

Fig. 3 - DETECTION DES IMPULSIONS DU FONDAMENTAL

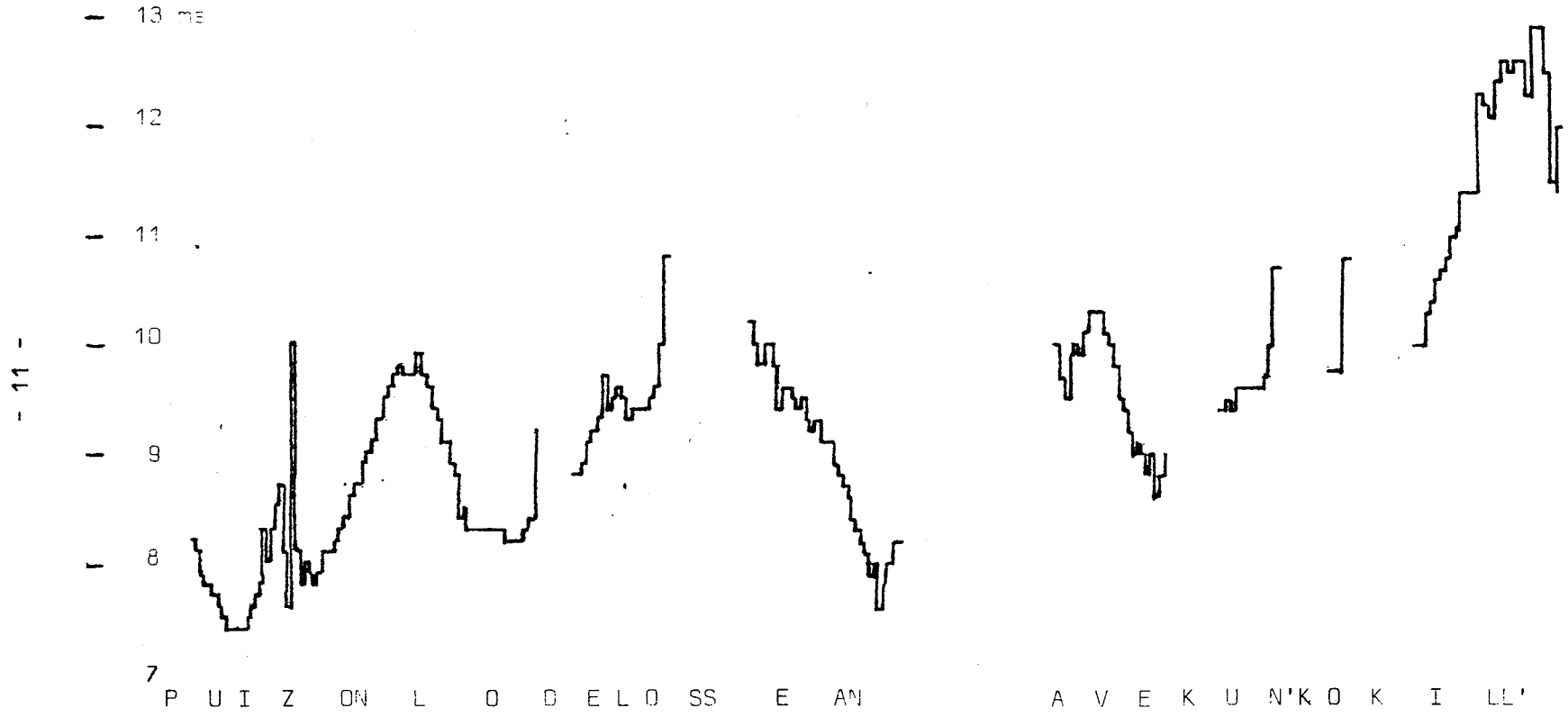


Fig. 4 - EVOLUTION DE LA PERIODE DU FONDAMENTAL SUR UNE PHRASE

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

DETECTION ET MESURE DU FONDAMENTAL

Jean - Frédéric ZURCHER , Michel CARTIER et Louis-Jean BOE

C.N.E.T. - LANNION

INSTITUT DE PHONETIQUE
DE GRENOBLE

RESUME

On décrit un détecteur de fondamental :

La décision de voisement est prise en fonction de la répartition spectrale de l'énergie. La période est obtenue par la mise en évidence de crêtes significatives dans des circuits analogiques de mise en forme, et par un algorithme câblé de correction et de mesure. Le dispositif fonctionne automatiquement dans des conditions variées d'utilisation.

SUMMARY

A pitch-detector is described :

Voicing decision is a function of spectral distribution of energy. Pertinent peaks are elaborated in a preprocessing part. Fundamental frequency is determined by correction and measurement circuits. This device automatically provides satisfactory results in a wide range of conditions and speakers.

DETECTION ET MESURE DU FONDAMENTAL

Jean-Frédéric ZURCHER, Michel CARTIER
C.N.E.T. LANNION

Louis-Jean BOE
Institut de phonétique
GRENOBLE

/INTRODUCTION/

Cet exposé présente une réalisation de détection et de mesure du fondamental étudié pour un vocodeur à canaux. Un tel dispositif doit satisfaire aux contraintes suivantes :

- fonctionner en temps réel.
- être indépendant du locuteur.
- accepter un fondamental affaibli et une dynamique importante.
- être peu sensible au bruit de fond.

Le dispositif décrit résulte d'améliorations successives apportées à un même principe - la détection de crêtes - dont une réalisation précédente a été décrite (1).

D'une façon générale il est possible de classer les méthodes de mesure du fondamental de la façon suivante :

1/ - Mesure de l'intervalle de temps entre :

- les passages par zéro du signal filtré (un ou plusieurs filtres).
- des crêtes d'amplitude du signal, de la puissance à très court terme etc.

2/ - Appréciation de la périodicité :

- méthode des différences.
- fonction d'auto-corrélation.

3/ - Traitement dans le domaine fréquentiel :

- recherche du fondamental, éventuellement d'harmoniques.
- cepstre.

Ces opérations de mesure et de décision sont généralement précédées d'une mise en forme dont le but est de renforcer ou isoler le fondamental (traitements linéaires), et /ou le régénérer (traitements non linéaires).

---

(1) J.F. ZURCHER - Dispositif de détection et de mesure du fondamental de la voix humaine.
Analyse et Synthèse de la parole - CNET - LANNION
1972/73, 1, pp. 7 - 15.

.../...

Pour plus de précisions nous renvoyons aux ouvrages de Mac KINNEY (2) et de PIROGOV (3).

L'appareillage qui a été réalisé se situe dans la première catégorie ; la mise en forme consiste en un filtrage suivie d'une compression d'enveloppe.

La détection des crêtes significatives utilise la dissymétrie et l'évolution à court terme de l'amplitude de tous les extrémum du signal. Le résultat final est calculé par un algorithme câblé de mesure et de correction. La décision de voisement est indépendante de la mesure du fondamental : elle ne dépend que de la répartition spectrale de l'énergie du signal de la parole (4).

/DETECTION DU VOISEMENT/

La détection du voisement est effectuée en application des propriétés de la répartition spectrale des énergies des différents sons. On sait que celle des sons voisés est centrée vers les basses fréquences du spectre du signal de parole (4).

Sur ce principe général, nous avons procédé à deux réalisations différentes. Dans la première on compare les énergies pondérées en deça de 700 Hz et au-delà de 1500 Hz. Si la première l'emporte sur la seconde, c'est la décision voisement qui est prise.

Un circuit à seuil impose le non-voisement si l'énergie totale est trop faible.

Cette réalisation a été testée à l'Institut de Phonétique de Grenoble (4). Les tests ont mis en évidence des "manques" pour certaines consonnes voisées. A titre indicatif, la durée de ceux-ci est de l'ordre de 72 % pour [z], 59 % pour [Z], 21 % pour [d] et décroit pour [R, g, b, v, j, n, l] jusqu'à 3 % pour [m]. Ces résultats nous ont conduits à une nouvelle réalisation dont le schéma de principe est indiqué figure n° 1.

---

(2) N.P. Mac KINNEY - Laryngeal Frequency Analysics for Linguistic. Research communic. Sci. Lab. Univ. Michigan Report 14. Contract 1224 (22) NR 049/122.1968.

(3) A.A. PIROGOV - Vokodernaja Telefonija. Metody i Problemy.Ed. Svjaz , Moscou.

(4) C. ABRY, L-J.BOE & J-F-ZURCHER - La détection du voisement par les propriétés phsiques résultant de l'excitation périodique du conduit vocal : comparaison de trois procédés . 6 ème J.E. du Groupe Communication Parlée du G.A.L.F. - Toulouse 1975.

.../...

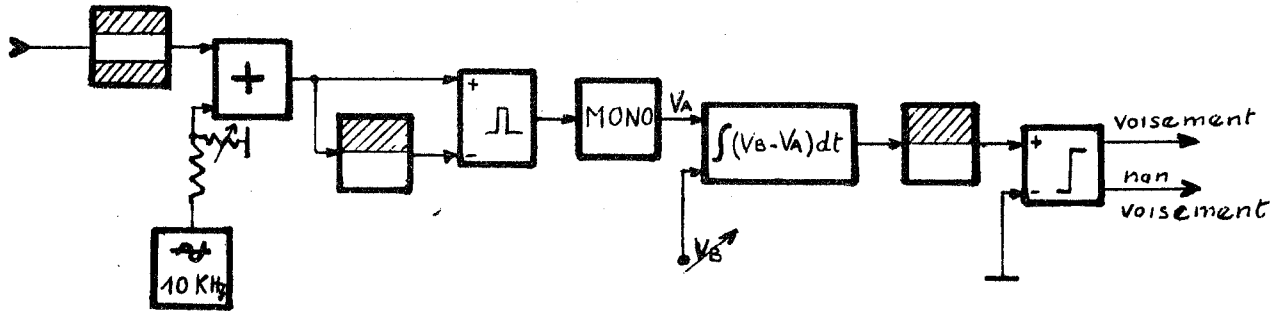


Figure 1

Le signal subit d'abord un filtrage dont la fonction de transfert est schématisée sur la figure 2.

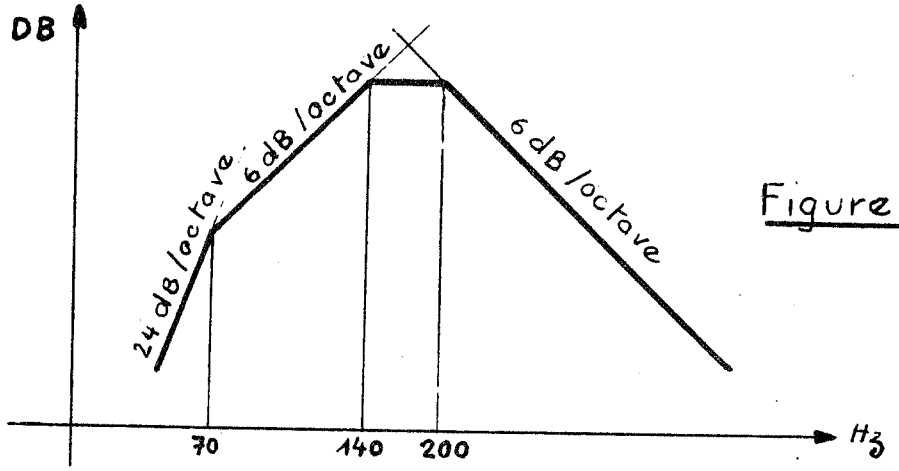
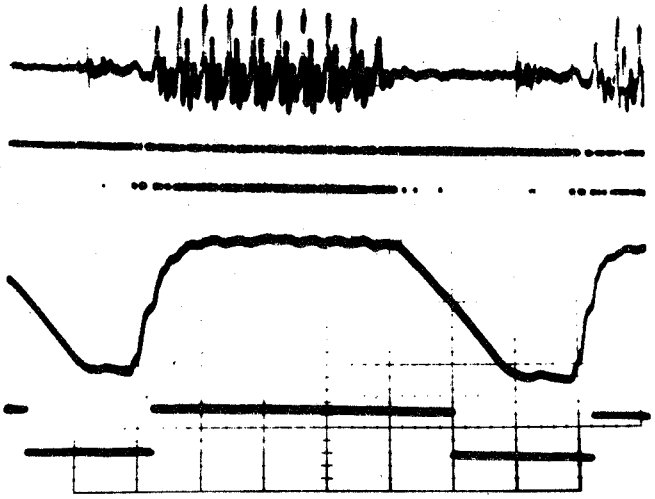


Figure 2

On lui additionne alors, à un niveau réglable, une fréquence pure de 10 KHz. L'ajustage de ce niveau détermine le seuil d'énergie en-dessous duquel tous les sons sont reconnus comme non voisés. Nous trouvons ensuite un dispositif de détection des passages par zéro, ou plus précisément des passages par la valeur moyenne obtenue dans un filtre passe-bas dont la fréquence de coupure est égale à 250 Hz. Ce dispositif déclenche une bascule monostable qui délivre des impulsions calibrées VA, de 0,8 ms, sur les passages par la moyenne d'un même sens.

Les impulsions VA attaquent un intégrateur qui délivre une tension $Vs = k \int (VB - VA)\, dt$, VB étant une tension réglable. En l'absence d'impulsions VA, Vs sera positif. VB st ajustée afin que Vs devienne négatif pour tout son non voisé, c'est-à-dire quand les passages par la moyenne sont nombreux. Vs subit un filtrage passe-bas dans un circuit RC dont la fréquence de coupure est égale à 35 Hz. Ce filtrage supprime les rebondissements indésirables. Un comparateur prend la décision voisement ou non selon le signe de Vs.

.../...



Signal de parole

Passages par la moyenne (après le mono.)

Sortie filtrée de l'intégrateur

Fonction de voisement.

Cette deuxième réalisation a été testée dans les mêmes conditions que la première. Les décisions erronnées sont inférieures à 1%. Elles se produisent dans des zônes de faible énergie et il n'est pas possible de dégager d'erreur systématique.

MESURE DE LA PERIODE FONDAMENTALE

Prétraitement du signal de parole

Afin de faciliter la détection des crêtes significatives le signal subit un filtrage dont la fonction de transfert est schématisée sur la figure 3.

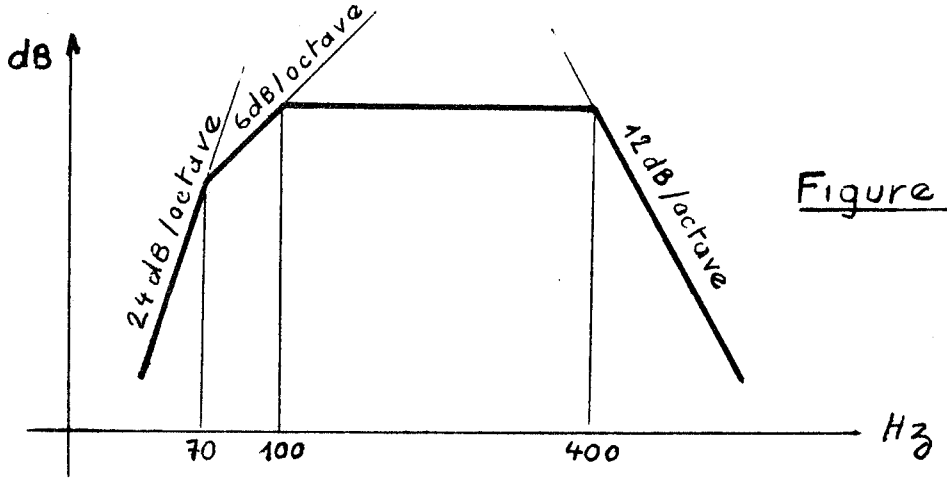


Figure 3

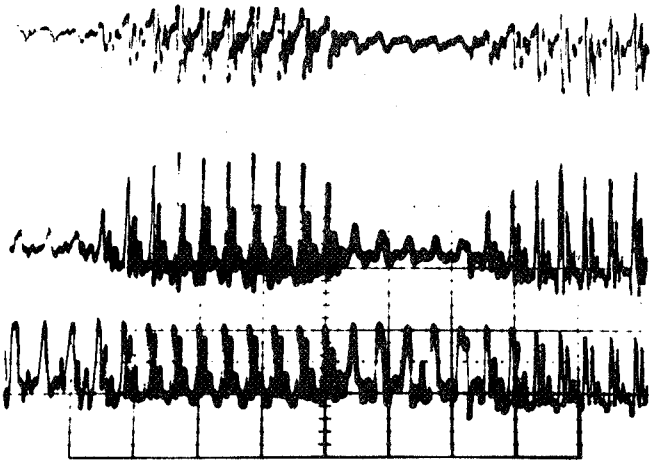
.../...

Ce filtrage comporte :

- un filtrage passe-haut (Fc = 70 Hz, 24 dB/octave).
- un filtrage passe-bas (Fc = 400 Hz ; 12 dB/octave).

et de plus un filtrage réjecteur de bande à 50 Hz.

Le signal ainsi filtré va subir une compression symétrique de l'enveloppe des crêtes maximum . Cette compression est essentielle pour la détection des crêtes significatives de la périodicité. Il est très difficile de détecter celles-ci sans discontinuité lorsque l'amplitude varie rapidement. Le compresseur utilisé se caractérise par un très faible temps de réaction à la montée et un temps de réaction à la descente d'autant plus court que l'énergie du signal décroît plus rapidement. Ceci est obtenu en commandant la compression par une fonction "mémoire de crête" dont la constante de temps de perte de mémoire varie en fonction de la dérivée de l'énergie du signal (5). De plus, la commande de la compression résulte de l'addition des signaux résultants des mémoires de crêtes positive et négative, ce qui assure une compression symétrique, propriété qui sera utilisée dans les circuits de repérage des crêtes. Le taux de compression est très important : 2 dB de variation en sortie pour une excursion de 60 dB à l'entée.



Repérage des crêtes significatives

Si, à l'intérieur d'une même période de son voisé, se présentent plusieurs crêtes d'amplitudes voisines, le circuit de repérage des crêtes doit être en mesure de délivrer une impulsion unique par période, deux à la rigueur. (Figure 4).

---

(5) J.L. COURBON - Régulation automatique de niveau - Etude et réalisation de deux compresseurs de dynamique.
Analyse et synthèse de la parole, CNET-LANNION, 1972/73, 1 pp. 47-59.

.../...

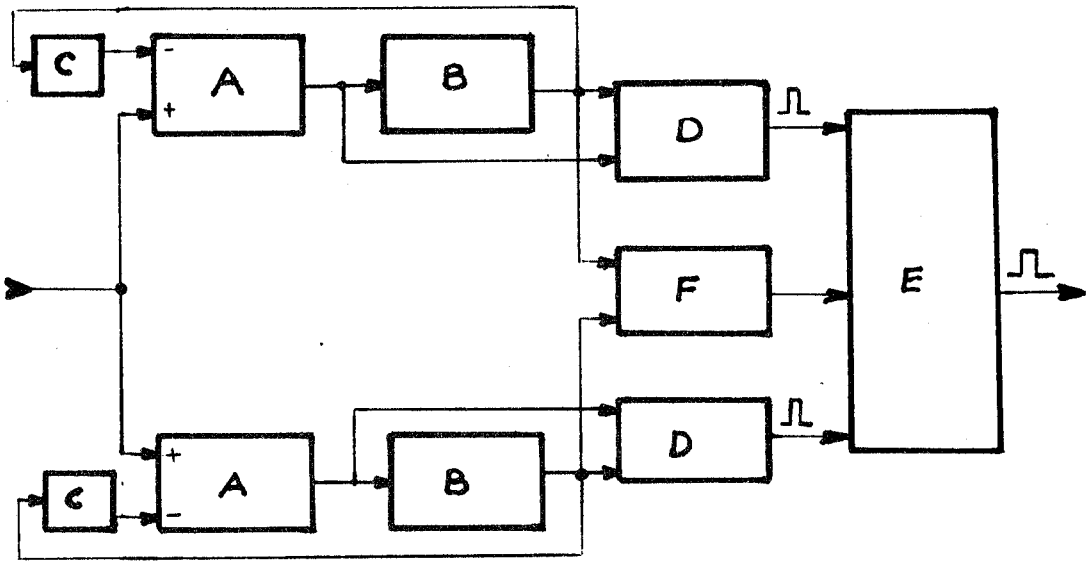


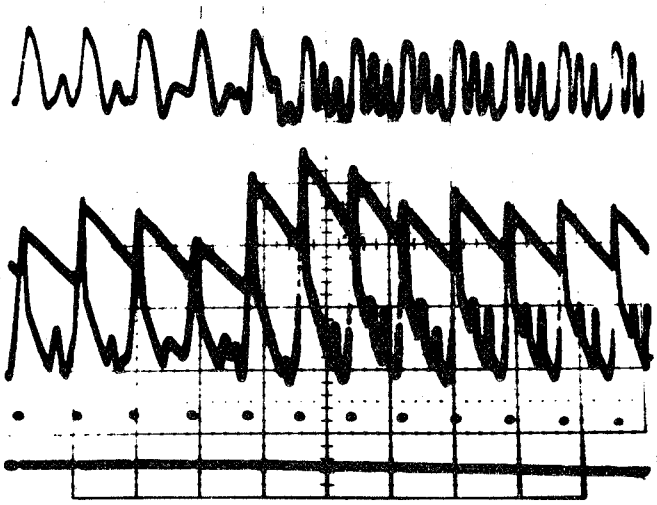
Figure 4

Les crêtes positives et négatives sont détectées dans deux chaînes pratiquement identiques.

Considérons par exemple la chaîne positive.

Le signal est appliqué sur l'entrée directe d'un amplificateur détecteur simple alternance A. La sortie A attaque un circuit de mémoire de crête B. La perte de la mémoire a un profil qui évite au maximum l'intersection avec des crêtes non significatives et sa constante de temps est fonction de l'amplitude des crêtes détectées. La tension de la mémoire B est appliquée avec un retard dû au filtre passe-bas C, sur l'entrée inverseuse de l'amplificateur A. Cette contre-réaction a pour effet d'affaiblir le signal après détection d'une crête importante, donc de mettre en valeur les crêtes significatives. La coïncidence des sorties de l'amplificateur A et de la mémoire de crête B est détectée dans le comparateur D qui délivre des impulsions à l'emplacement des crêtes significatives de la périodicité :

.../...



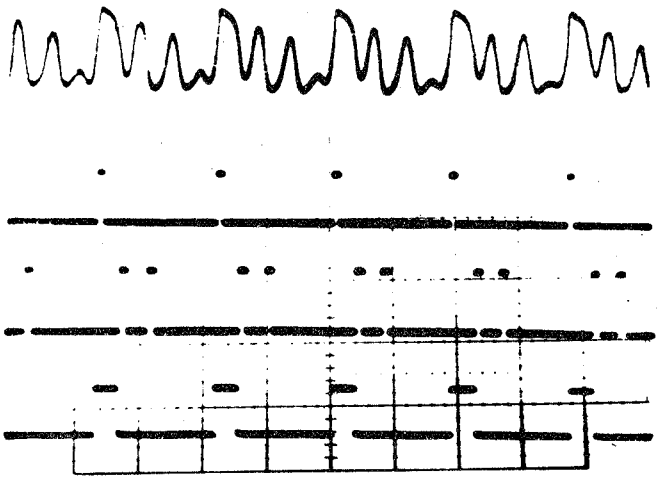
Signal de parole prétraité

Mémoire de crête B

Sortie de l'amplificateur A

Impulsion délivrée par D

Les sorties des deux chaînes sont couplées par un circuit logique E piloté par un dispositif F qui reconnaît la chaîne la plus favorable à une bonne détection des crêtes significatives.



Signal prétraité

Sortie de la chaîne favorable

Sortie de la chaîne défavorable

Sortie du dispositif de couplage

Algorithme de correction et de mesure

Cet algorithme a pour but d'éliminer les impulsions non significatives et de déterminer et mesurer la période du fondamental (figure 5).

Considérons une séquence voisée. Soit Imp 1 la première impulsion, Imp n la dernière. Nous appelons Tn le temps qui sépare Imp n - 1 et Imp n et t' n l'intervalle entre les deux dernières impulsions non supprimées.

.../...

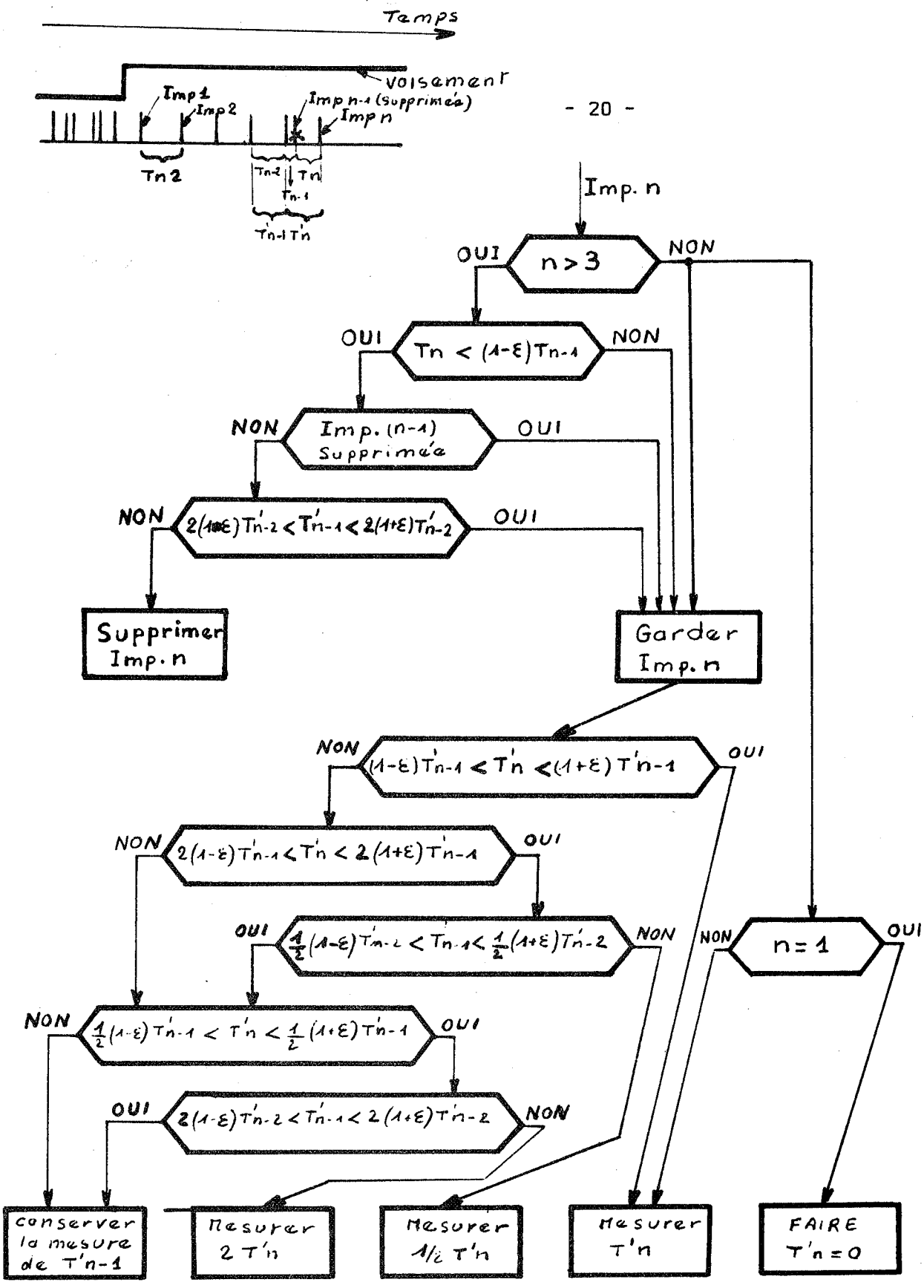
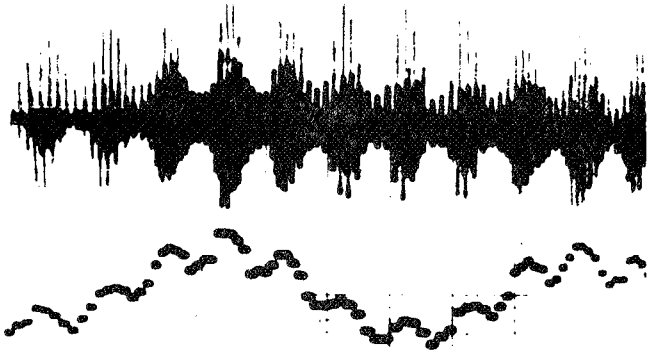


Fig. 5 : Algorithme de Correction et de Mesure

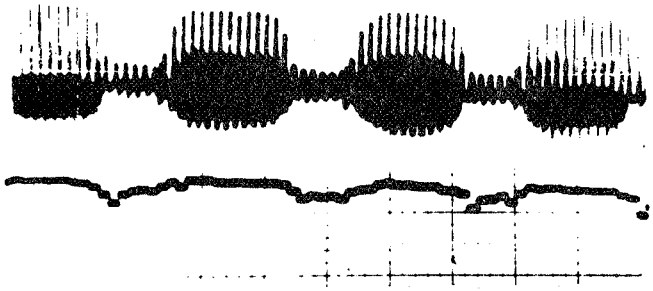
La réalisation de l'algorithme fait appel pour les comparaisons à des techniques analogiques, ce qui permet de garder une tolérance relative constante de l'ordre de 20 %, et à des techniques logiques pour la détermination de la valeur de la période. Néanmoins celle-ci est aussi restituée en analogique sous la forme d'une tension $V = K \log \frac{1}{T} + C^{te}$.



Signal de parole
(R vibré chanté)

Visualisation de la fréquence du fondamental

On notera la micro-mélodie dûe aux vibrations.



Signal de parole

a d a d a d a

Visualisation de la fréquence du fondamental

RESULTATS OBTENUS - CONCLUSIONS

Dans une ambiance de laboratoire, avec un microphone dynamique, aucune erreur notable n'a été décelée sur un large éventail de voix féminines et masculines. Ce résultat ne nécessite aucune adaptation à la voix du locuteur.

En attendant des essais plus complets de ce dispositif, on peut évaluer à 5-10%le temps moyen d'erreur sur une voix d'homme et une voix de femme distordues par un poste téléphonique. Par rapport au système cité en référence (1), la qualité d'écoute a été sensiblement améliorée bien que les tracés d'évolution du fondamental ne présentent pas de différences notables.

Le traitement décrit dans cet exposé est effectué en temps réel, sans retard, avec une mémoire de deux périodes. A partir de ces résultats, un traitement ultérieur, adapté à des conditions d'exploitation sévères, pourrait être développé dans deux optiques :

- celle du temps réel par la prise en compte de tout le passé du signal.
- celle du temps différé où interviendraient les contraintes d'un modèle d'évolution.

-=-=-=-=-=-

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UNE METHODE DE DESCRIPTION PROSODIQUE

Y. GRENIER

RESUME

La description de l'évolution de paramètres prosodiques (fondamental, enveloppe énergétique....) a déjà été abordée par des approches grammaticales. On propose ici de considérer la courbe temporelle à décrire comme la sortie d'un système dynamique excité par une entrée connue. L'entrée peut être considérée comme une suite de commandes visant une cible e_i, survenant aux instants t_i. Le système (décrit par une relation récurrente d'ordre 2) rend compte de la dynamique des organes phonatoires (amortissement Z_i, fréquence propre w_i.)

SUMMARY

The description of the evolution of prosodic parameters (Fundamental, Frequency, energy....) has already been investigated by grammatical approach. In the present approach, the temporal curve is considered as the output of a linear system, with a known input. The input may be viewed as a sequence of commands aiming a target e_i, at instants t_i. The system, described by a recurrent equation of order 2, specifies the dynamic of articulatory organs.

UNE METHODE DE DESCRIPTION PROSODIQUE par Y. GRENIER

I - Description de la méthode

I-1. On rencontre souvent en reconnaissance de la parole ou en reconnaissance des locuteurs le problème de décrire et modéliser des courbes représentant une évolution temporelle de paramètres, pour lesquelles les méthodes classiques d'analyse (spectre ...) ne peuvent s'appliquer. Ce peuvent être par exemple l'évolution du fondamental, des Formants, de l'enveloppe énergétique sur l'ensemble d'une phrase, ou simplement au cours d'une transition phonémique.

Une approche grammaticale de ce problème est possible (De Mori).

La méthode envisagée ici consiste à considérer la courbe temporelle à étudier comme la sortie d'un système excité par une entrée donnée (suite de créneaux de hauteur et de durée variable). Le système est décrit par une relation récurrente dont l'ordre sera choisi minimum (ordre 2) permettant d'ajuster l'amortissement Z_i et la fréquence propre w_i.

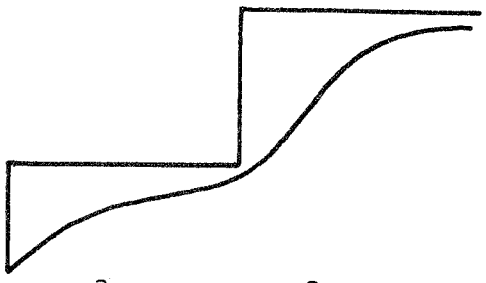


Fig. 1

$a_{o,i}$ $a_{o,i+1}$

$a_{1,i}$ $a_{1,i+1}$

$a_{2,i}$ $a_{2,i+1}$

Le problème est de déterminer les instants t_i et entre ces instants les valeurs $\{e_i, a_o, a_1, a_2\}$

I-2. Détermination des t_i

Deux méthodes sont envisageables :

- les calculer à priori et indépendamment de $\{e_i, a_o, a_1, a_2\}$
- les calculer à partir de l'erreur de prédiction du modèle linéaire (on fixe alors t_{i+1} lorsque l'erreur de prédiction sur le modèle $\{t_i, e_i, a_o, a_1, a_2\}$ devient supérieure à un certain seuil, par exemple

$$|S_n - \hat{S}_n| > \varepsilon \quad ,$$

Il est apparu tout aussi efficace (et plus simple) de déterminer les t_i par la première méthode, en repérant certaines formes de la courbe à modéliser, et certains points particuliers.

../..

I.2.1. Modèle d'ordre 1

Les t_i sont les points où la dérivée première s'annule (car les segments $[t_i, t_{i+1}]$ doivent correspondre à des portions monotones de la courbe).

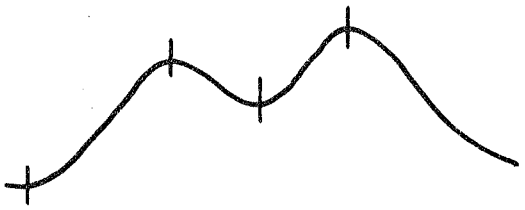
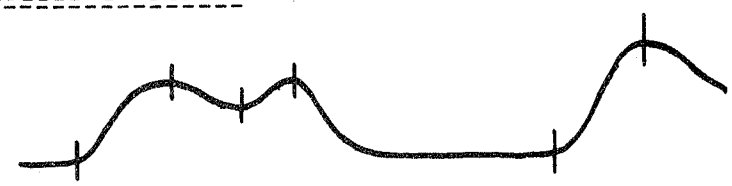


Fig. 2

L'ordre 1 a dû être abandonné à cause de la discontinuité de la courbe estimée, aux points t_i.

I.2.2. Modèle d'ordre 2



Il apparaît que les instants où la dérivée première s'annule peuvent être conservés (au moins ceux qui correspondent à des maxima). Il a fallu en outre choisir les instants où la pente de la courbe est maximum, et positive, soit les instants où la dérivée seconde du signal devient négative.

La dérivée première (resp. seconde) est estimée par régression linéaire sur 4 points de la courbe (resp. 5 points de la dérivée première) afin d'éliminer l'influence d'un bruit éventuel sur la courbe temporelle (dû par exemple à l'influence du pitch sur le calcul de l'énergie sous une fenêtre étroite) ce qui évite de lisser la courbe temporelle.

Les expressions de la dérivée première d_1 et deuxième d_2 sont :

$$d_{1,j} = -\frac{3}{10} S_{j-2} - \frac{1}{10} S_{j-1} + \frac{1}{10} S_j + \frac{3}{10} S_{j+1}$$

$$d_{2,j} = \frac{6}{100} S_{j-2} + \frac{5}{100} S_{j-1} - \frac{1}{100} S_j - \frac{1}{10} S_{j+1} - \frac{1}{10} S_{j+2} - \frac{1}{100} S_{j+3} + \frac{5}{100} S_{j+4} + \frac{6}{100} S_{j+5}$$

avec S_j = échantillon du signal à l'instant $j \Delta t$

(le déplacement des indices j dans $d_{2,j}$ a été introduit afin de créer une avance dans les points repérés par $d_{2,j} = o$, pour placer les t_i en début de flanc montant du signal).

Ces deux types de points ont été choisis à partir de considérations sur la réponse du système. Il a été supposé que l'amortissement du système était assez élevé pour qu'un régime oscillant apparaisse peu.

../..

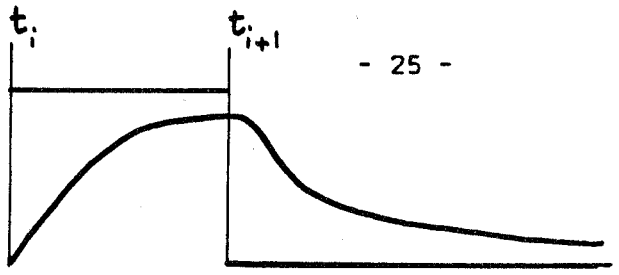


Fig. 4

I.2.3. Modèle d'ordre supérieur à 2

Lorsque l'ordre croît, le modèle devient trop souple, et englobe trop facilement les courbes à étudier, l'entrée en créneaux perdant son sens. Nous ne considérerons donc plus que des modèles d'ordre 2.

I-3. Détermination des e_i et a_o, a_1, a_2

Notations :

S_n = valeur du signal à l'instant $t_{i,n} = t_i + n \Delta t$

Δt = période d'échantillonnage de la courbe

$\hat{S}n$ = valeur estimée du signal à l'instant $t_{i,n}$

S_o = valeur du signal à l'instant t_i

S_{-1} = " " " " " $t_i - \Delta t$ conditions initiales

Le modèle choisi est de la forme $e_i = a_o S_n + a_1 S_{n-1} + a_2 S_{n-2}$

La modélisation se fait en minimisant un critère qui est :

$$\Delta_i = \sum_{ti}^{t_{i+1}} (S_n - \hat{S}_n)^2$$

Ce critère peut être compris de deux manières différentes suivant la manière de calculer $\hat{S}_n$.

I.3.2. $\hat{S}_n = \frac{1}{a_o} (e_i - a_1 S_{n-1} - a_2 S_{n-2})$

C'est la manière habituelle de calculer $\hat{S}n$. Elle conduit à une résolution algébrique complète du problème, à condition de s'imposer, soit $a_o = 1$.

soit $a_o + a_1 + a_2 = 1$.

Néanmoins cette méthode, valable pour des segments comportant un nombre élevé de points, ne l'est plus sur des segments pouvant comporter parfois 5 points. Il est alors préférable de choisir la méthode suivante.

I.3.2. $\hat{S}_n = \frac{1}{a_o} (e_i - a_1 \hat{S}_{n-1} - a_2 \hat{S}_{n-2})$

où $\hat{S}_n$ ne dépend plus que de $\{S_{-1}, S_o, e_i, a_o, a_1, a_2\}$.

Malheureusement la complexité des calculs empêche de déterminer (à ma connaissance) les paramètres optimaux. La méthode retenue est une combinaison des deux. ../..

α - détermination de a_o, a_1, a_2 par la méthode 1, après s'être fixé une valeur de e_i

$$e_i = \frac{11}{10} S_{n_i} - \frac{1}{10} S_o \qquad n_i = \frac{1}{\Delta t} (t_{i+1} - t_i)$$

β - détermination de e_i optimum, a_o, a_1, a_2 étant fixés.

Δ_i est un polynome de degré 2 en e_i, il est donc aisé de calculer la valeur de e_i qui minimise Δ_i, bien qu'on ne connaisse pas l'expression de Δ_i en fonction de e_i.

On peut alors itérer le processus en revenant au 1° et déterminer les (a_o a_1 a_2) optimums pour cet e_i etc... En fait l'amélioration obtenue est faible, même à la seconde itération. On se contentera donc d'un seul pas d'itération.

I-4. Introduction des conditions initiales

Le calcul des e_i se fait à partir des $\hat{S}_n$ qui ne dépendent que de a_o, a_1, a_2 et S_o S_{-1}

Pour le premier segment $[t_o, t_1]$, S_o et S_{-1} sont estimés par régression linéaire sur les 4 premiers points de la courbe.

Pour les segments suivants, on prend pour $S_{-1,i}$ et $S_{o,i}$ les deux derniers points estimés du segment précédent

$$S_{o,i} = \hat{S}_{n(i-1)}, \; i-1$$

$$S_{-1,i} = \hat{S}_{n_{(i-1)}-1}, \; i-1 \qquad \text{avec } n_{i-1} = \frac{1}{\Delta t} (t_i - t_{i-1})$$

Ceci conduit à modifier l'expression de Δ_i, en effet il est important que S_o et S_{-1} représentent bien les valeurs initiales de la courbe à estimer, sur le segment étudié, sinon des erreurs de tangente ou de position se répercuteront d'un segment à l'autre sur toute la longueur de courbe étudiée. Il a fallu privilégier les évènements les plus récents du segment afin de contraindre la courbe estimée $\hat{S}_n$ de s'approcher au mieux, en fin de segment de la courbe réelle S_n. La distance Δ_i a été remplacée par :

$$\Delta_i' = \sum_{n=1}^{n=n_i} \alpha^n (S_n - \hat{S}_n)^2 \qquad \alpha \text{ légèrement supérieur à 1.}$$

II - Analogie physique

Il est plus parlant de considérer les paramètres d'amortissement et de fréquence propre, aussi chaque segment sera-t-il modélisé par (t_i, e_i, Z_i, ω_i)

Z_i = amortissement

ω_i = fréquence propre.

Bien que l'analogie ne soit pas complète, on pourra considérer ces paramètres comme décrivant la dynamique des organes phonatoires, soumis aux instants t_i, à des commandes visant les cibles e_i. ../..

III - Applications et développement futur

III-1. On trouvera Fig. 5 un exemple de courbe modélisé par cette méthode. Il s'agit de l'énergie sur le début d'une phrase. Les applications de la méthode sont nombreuses :

- en reconnaissance des locuteurs pour modéliser l'évolution du pitch, l'enveloppe énergétique, les Formants sur une phrase code
- en reconnaissance de la parole pour modéliser la sortie par exemple d'un vocodeur.

III-2. Il serait intéressant de pouvoir implanter ce système en ligne et non plus avec un retard, pour cela il faudrait revenir sur le mode de détermination des t_i. (par. 1/2, 1er alinéa) afin de déterminer simultanément t_{i+1}, e_i, Z_i et ω_i avec t_i fixé. Il sera souhaitable également de revenir sur les méthodes exposées au 1/3/1/ et 1/3/2/ afin d'expliciter la seconde méthode qui semble plus performante.

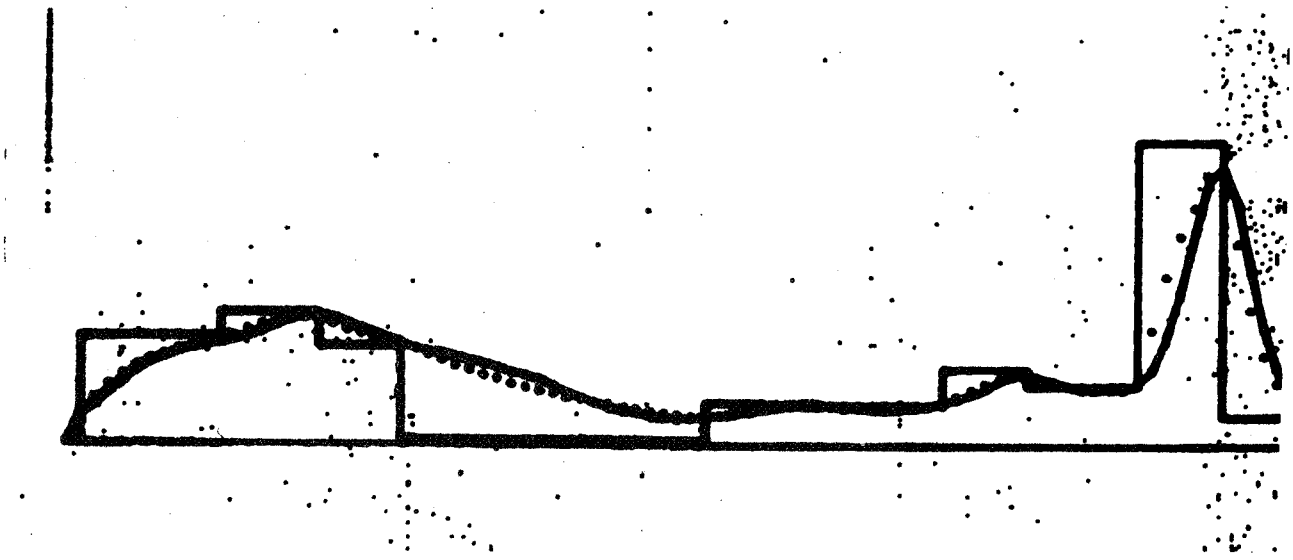


Fig. 5

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

RECONNAISSANCE DE PATRONS INTONATIFS

E. LHOTE

Laboratoire de Phonétique
BESANCON

M. FILLEAU

Lab. d'Automatique
de l'ENSCM BESANCON

M. F. GRANGE

Sercice médico
psychologique
du Jura (Suisse)

RESUME

Les auteurs utilisent des glottogrammes, considérés comme la forme globale de la source vocale, pour tester la reconnaissance de phrases et la reconnaissance de locuteurs à partir des seuls patrons mélodiques.

Les résultats montrent qu'il y a interaction entre les indices de reconnaissance de phrases et les indices de reconnaissance de voix.

SUMMARY

The aim of this present study is to use glottograms considered as the global form of the vocal source to test how it is possible to recognize sentences or speakers from melodic patterns only.

The results show some interaction between cues to recognize sentences and cues to recognize voice.

RECONNAISSANCE DE PATRONS INTONATIFS

E. LHOTE

Laboratoire de Phonétique
BESANCON

M. FILLEAU

Lab. d'Automatique
de l'ENSCM BESANCON

M. F. GRANGE

Service médico
psychologique
du Jura (Suisse)

-o-

Le travail que nous présentons s'inscrit dans l'ensemble des études orientées vers la connaissance des faits laryngés au cours de la parole.

Dans un travail antérieur nous avons constaté sur des sons tenus, que la forme de l'onde captée par un glottographe varie peu avec les locuteurs, mais varie avec les catégories de sons, et nous avons émis l'hypothèse que si le glottogramme contient des informations caractéristiques sur la voix du locuteur, c'est vraisemblablement dans la forme globale du signal qu'il faut les chercher. (6)

Nous avons donc choisi des phrases et nous considérons que le glottogramme de ces phrases constitue la forme globale de la source vocale pour ces énoncés : nous appelons cette forme un patron intonatif.

Les tests de reconnaissance effectués ensuite à partir des seuls glottogrammes des phrases ont eu pour objet :

- de comparer le taux de reconnaissance des phrases en fonction de leur contenu, de leur mouvement mélodique, de la modification apportée par le déplacement d'un élément, et en fonction des locuteurs ;

- de faire apparaître des probabilités de reconnaissance des locuteurs à partir de leurs patrons mélodiques propres.

I - CONDITIONS EXPERIMENTALES

I-I- LE CORPUS :

Nous avons tout d'abord élaboré un corpus de IO phrases répondant à certaines conditions.

I-I-I Les énoncés ne comportent que des sons sonores (sauf 2 exceptions) afin d'éviter que la reconnaissance ne se fasse à l'aide des discontinuités des patrons.

I-I-2 Les énoncés sont brefs, ce qui permet de les réaliser en principe dass un seul groupe de souffle. Ils ont à peu près la même durée (7 ou 8 syllabes).

I-I-3 Nous avons peu différencié le choix des phonèmes afin d'atténuer les effets de la micromélodie.

I-I-4 Dans le choix des phrases nous avons évité de faire intervenir des nuances de sens par l'intonation.(5)

CORPUS :

I- Demain il ira à Venise
2- Il ira à Venise demain
3- Il ira demain à Venise
4- S'il va mieux il ira demain
5- Il va venir à Lille demain
6- Va les voir à Lille demain
7- Il aimerait bien voyager
8- Il a envie de voyager
9- Il aime lire et voyager
IO- Voyager est un grand rêve.

I-2- LES SUJETS - LOCUTEURS :

Ces phrases ont été dites par 8 locuteurs différents, tous masculins, de même classe d'âge, travaillant dans le même laboratoire, et ne possédant pas de connaissances spéciales en phonétique.

Ils ont été enregistrés à l'aide d'un magnétophone NAGRA IV S:

- Au niveau de la bouche par un microphone A. K. G. (phonogrammes)
- Au niveau du larynx par un électroglottographe (glottogrammes)

Il a été recommandé aux sujets d'éviter les pauses,mmais de donner la priorité au caractère naturel et spontané de leur parole.

...

I-3- LES TESTS DE RECONNAISSANCE :

Ces tests ont été élaborés à partir des IO phrases de 4 locuteurs Nous avons combiné les phrases d'abord, les phrases et les locuteurs ensuite.

I-3-I Test I : Les phrases.

3 triplets ont été constitués :
Les phrases I-2-3
Les phrases 4-5-6
Les phrases 7-8-9

et à l'intérieur de chaque triplet, 5 combinaisons différentes ont été effectuées sur chacun des 4 locuteurs.

Ayant l'indication du locuteur et du triplet, les auditeurs devaient reconnaître les phrases dans les triplets.

I-3-2 Test 2 : Phrases et locuteurs.

Nous avons choisi certaines phrases de chaque locuteur et leur avons donné une distribution aléatoire. Les auditeurs devaient reconnaître la phrase et le locuteur.

I-4- LES AUDITEURS :

Nous avons demandé d'abord aux sujets locuteurs de reconnaître la voix de ceux qu'ils connaissent. Nous avons ensuite soumis les 2 tests à un groupe de 25 étudiants participant à un même cours de phonétique, groupe homogène par le niveau de connaissances, mais hétérogène par les origines linguistiques (5 français et 20 étrangers : 9 hispanophones d'Amérique latine, 3 Thaïlandais, I Vietnamien, 2 Zambiens, I Japonais, I Indonésien, I Allemand, 2 Brésiliens).

Ces étudiants ne connaissaient pas les locuteurs ; de plus ils n'ont pas entendu les phonogrammes avant les épreuves de reconnaissance qui ne portent que sur des glottogrammes.

I-5- Une analyse sonagraphique de l'ensemble des glottogrammes a été réalisée. Elle sera en partie utilisée pour l'interprétation des résultats.

2- RESULTATS DES EXPERIENCES

2-I- RECONNAISSANCE DES LOCUTEURS PAR CEUX QUI LES CONNAISSENT :

D'une façon générale il est très rare que l'on reconnaisse le locuteur si on ne sait pas d'abord parmi quels locuteurs il faut reconnaître. La reconnaissance par l'audition de glottogrammes est en effet très difficile ; il ne faut pas négliger l'importance de l'apprentissage dans tout fait de reconnaissance, et personne n'est a priori, entraîné à enten-

tendre des patrons intonatifs privés de leur support phonémique.

Si l'on indique l'ensemble des locuteurs, la reconnaissance est grandement améliorée :

- Certains auditeurs (très peu) reconnaissent bien et vite les locuteurs ; d'autres ne reconnaissent qu'un seul des 4 locuteurs.

- Le locuteur 3 est bien reconnu par tout le monde. C'est celui que tous s'accordent à dire reconnu immédiatement au téléphone.

- Les locuteurs fumeurs sont reconnus par cet indice.

- Les indices de reconnaissance que chaque auditeur fournit pour un locuteur donné sont extrêmement variables d'un auditeur à l'autre.

2-2- RECONNAISSANCE DES PHRASES (Test I)

Par 25 auditeurs.

Les résultats sont très différents selon les triplets :

Tableau :

Phrase I	Phrase 2	Phrase 3
44,6 %	4I,8 %	39,8 %
Phrase 4	Phrase 5	Phrase 6
78,6 %	58,4 %	60 %
Phrase 7	Phrase 8	Phrase 9
42,4 %	53,6 %	62,4 %

Tableau : Pourcentage d'identification correcte de chaque phrase dans un triplet.

Remarques :

Les réponses correctes de l'ordre de 50 % (ou même de 40 %) ne sont pas imputables au hasard, car il y a ici un choix ternaire ; d'autre

...

part nous avons constaté que souvent l'auditeur s'accroche à une phrase du triplet qu'il reconnaît bien dans toutes les combinaisons.

Il faut aussi noter que la reconnaissance des phrases se fait bien chez certains locuteurs et pas chez d'autres.

2-3- RECONNAISSANCE DES PHRASES ET DES LOCUTEURS : (Test 2) par 22 auditeurs.

Ce 2ème test ne peut se faire qu'après avoir fait subir le premier, en raison de sa difficulté : il nécessite en effet un très gros effort et un apprentissage par l'entraînement réalisé lors du premier test.

Précisons aussi que les auditeurs doivent choisir parmi les IO phrases.

Nous nous attendions en fait à un échec total, c'est-à-dire à constater qu'il était impossible de reconnaître à la fois le locuteur et la phrase (Rappelons que les auditeurs ne connaissent des locuteurs que les glottogrammes du Ier test).

2-3-I Certains auditeurs ont reconnu à la fois phrase et locuteur dans quelques cas. Les meilleurs résultats sont obtenus par un allemand (musicien) et une thaïlandaise : I6 % des occurrences.

2-3-2 La plupart des auditeurs reconnaissent :

- soit la phrase - soit le locuteur - mais pas les deux à la fois.

20 % de l'ensemble des phrases est correctement identifié. La phrase la mieux reconnue est la phrase 4, ce qui était attendu.

26,7 % de l'ensemble des réalisations des locuteurs ont été attribués au bon locuteur, le locuteur 3 étant toujours le mieux reconnu (cf. 2-I-)

3- INTERPRETATION DES RESULTATS

La reconnaissance des phrases et celle des locuteurs ne relèvent pas des mêmes processus : dans le premier cas l'auditeur effectue une démarche linguistique et fait appel à ses propres habitudes intonatives quand il parle français, qu'il s'agisse de sa langue maternelle ou d'une langue seconde ; dans le second cas l'auditeur sollicite toutes ses réactions personnelles et cherche à "ficher" ses sensations à l'audition des voix.

La qualité des résultats obtenus tient à la fois :

- au contenu lexical et syntaxique des phrases choisies

- à la nature des voix individuelles

...

- et à la réalisation personnelle de chaque phrase par chaque locuteur.

3-I- RECONNAISSANCE DES PHRASES :

Si la phrase 4 (conditionnel), caractérisée par un schéma montant suivi d'un schéma descendant, est la mieux reconnue, c'est parce que c'est la seule du groupe, et qu'elle est très nettement différenciée des autres.

Par contre si les 3 premières phrases ont obtenu à peu près les mêmes pourcentages, faibles, c'est que la différence très faible (dans la signification et dans la mélodie) rend difficile la discrimination dans le triplet.

La IOème phrase a été la phrase la mieux reconnue (2ème test) après la phrase 4 ; ceci tient en partie à son contenu et en partie à sa position dans le corpus : en effet c'est la seule phrase se prêtant à l'emphase et chaque locuteur l'a interprêtée comme un groupe final dans un long énoncé.

Pour les phrases 7 et 8,

7- Il aimerait bien voyager
8- Il a envie de voyager

la distinction se fait par le changement de place de la montée du mouvement mélodique montant, c'est-à-dire par la mise en relief de "bien" en 7 et de "envie" en 8 (figure).

3-2- RECONNAISSANCE DES LOCUTEURS :

Nous n'avons pas analysé tous les éléments permettant de justifier ou d'expliquer la reconnaissance des locuteurs.

Si l'on s'en tient à l'avis des auditeurs, on a l'impression que chaque auditeur a son propre système de traits de reconnaissance, et que ce qui est pertinent pour l'un ne l'est pas pour l'autre.

La reconnaissance de ces voix masculines se fait à peu près de la même façon, quel que soit le sexe de l'auditeur ; mais elle semble varier avec "l'oreille" de l'auditeur ; il y a très nettement prédominance de la reconnaissance de la voix sur la reconnaissance des phrases chez certains individus musiciens ou d'origine étrangère, en particulier chez des sujets dont la langue maternelle est une langue tonale.

Il faut aussi faire remarquer que très souvent, la reconnaissance est dirigée avec acuité sur un paramètre, inhibant la faculté de reconnaissance du 2ème paramètre : ceux qui reconnaissent bien les phrases ne reconnaissent pas bien les locuteurs, et inversement.

CONCLUSION :

Ces travaux sur la reconnaissance de certains patrons glottographiques de phrases françaises doivent se poursuivre, car il ne portent que sur 4 des 8 locuteurs enregistrés, et essayer de faire progresser la notion de reconnaissance de locuteurs à partir de patrons mélodiques.

Quoique utilisant une technique personnelle, ce rapport est un prolongement de bien des travaux antérieurs, et en particulier ceux de :

- ABBERTON (I), AUTESSERRE et DI CRISTO (2), BOE et LARREUR (3) CARRE (4), STEVENS (7), SUZUKI et NAKATSUI (8), VAISSIERE (9), VIVES - BUISSON - GRESSER - MERCIER - QUERRE (IO).

Il fait apparaître que le glottogramme considéré comme la forme globale de la source vocale apporte d'une part des informations sur la phrase -ce qui n'est pas nouveau - d'autre part sur la voix du locuteur.

Les résultats des tests ont également montré qu'il y a interaction entre les indices perceptuels qui participent à la reconnaissance de phrases et ceux qui participent à la reconnaissance de voix.

BIBLIOGRAPHIE

I- ABBERTON E. : Listener identification of speakers from larynx frequency. 8th Int. Cong. on Acoustics, London I974

2- AUTESSERRE D. , DI CRISTO A. : Recherches sur l'intonation du français traits significatifs et non significatifs. Actes du 7ème Cong. Int. Sciences Phon., Montréal. Mouton I972,842 - 859.

3- BOE L. J. , LARREUR D. : Les caractéristiques intrinsèques de la fréquence laryngienne : production, réalisation et perception : 5èmes journées d'Etude du GALF, ORSAY I974.

4- CARRE R. : Identification des locuteurs ; exploitation des données relatives aux fréquences formants.7th Int. Cong. on Acoustics, Budapest I97I.

5- DELATTRE P. : La nuance de sens par l'intonation.French Review 4I, I967, 326 - 339.

6- LHOTE E. : Contribution à l'étude de la fonction linguistique du larynx. Phonetica 28, I973, 26-4I

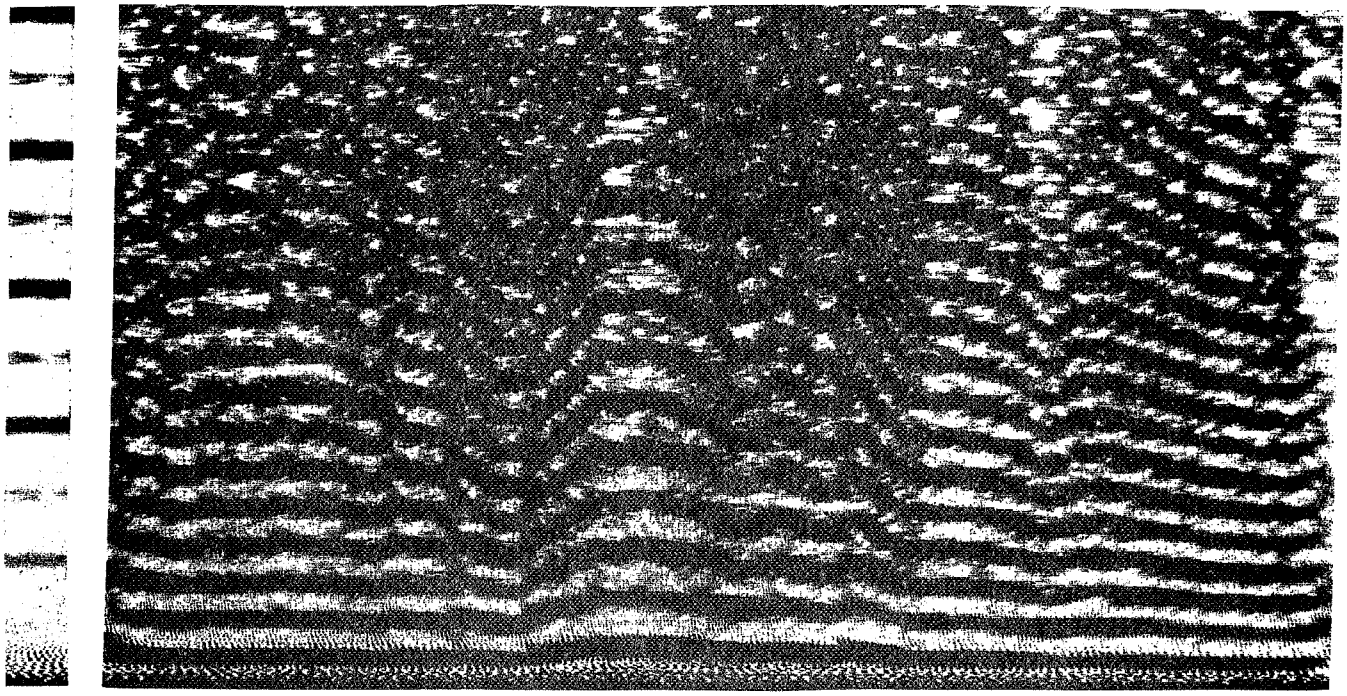
7- STEVENS K. : Source of inter and intra-speaker variability on the acoustic properties of speech sounds.7ème Cong Int. Sci. Phon., Montréal,Mouton I972, 206 - 232.

8- SUZUKI J., NAKATSUI M : Information of individuality conveyed by vowels. 7th Int. Cong. on Acoustics, BUDAPEST I97I.

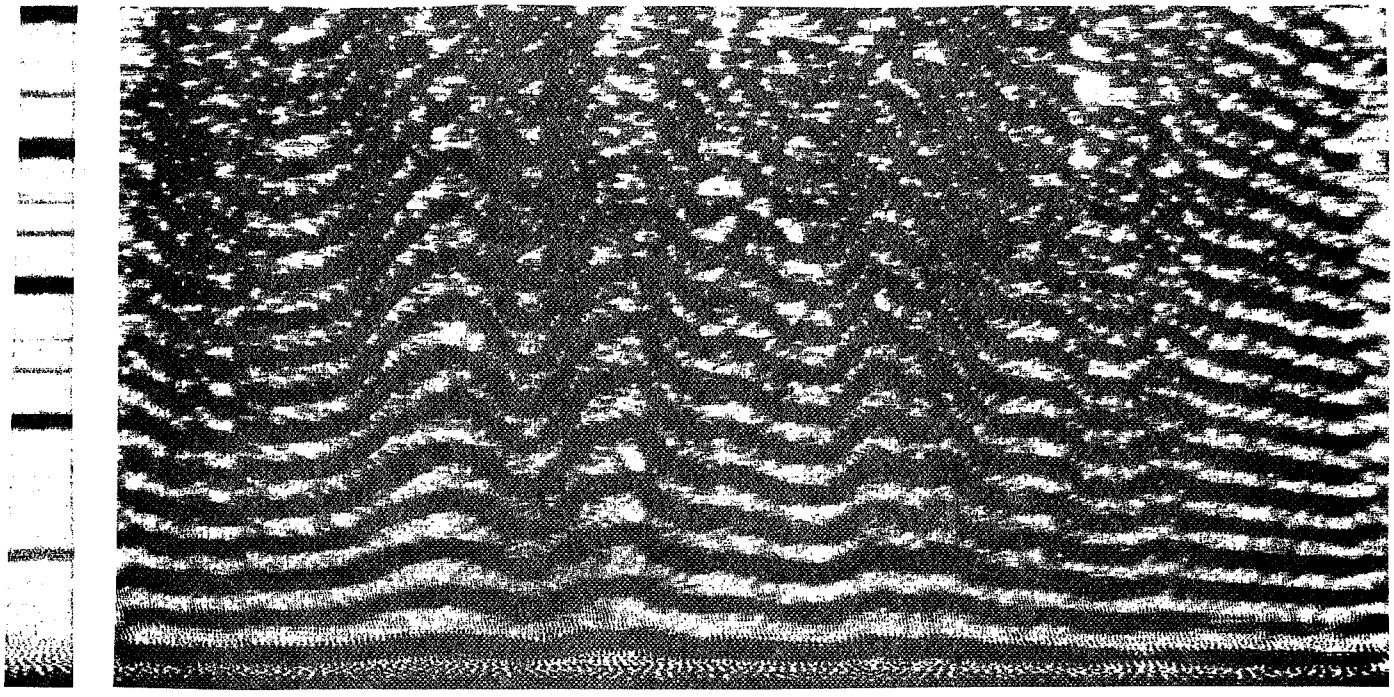
9- VAISSIERE J. : - Contribution à la synthèse par règle du français. Thèse, Grenoble I97I.

- Fréquence fondamentale des phrases déclaratives en français. 5èmes journées d'Etude du GALF, ORSAY I974.

IO - VIVES, BUISSON, GRESSER, MERCIER, QUERRE :Reconnaissance des grands dictionnaires prononcés par plusieurs locuteurs. 5èmes journées d'Etudes du GALF, ORSAY I974.



IL AI ME RAIT RIEN V O ' A ' E R



IL A EN VIE DE VOYA GER

FIGURE : Analyses sonagraphiques des glottogrammes des phrases 7 et 8 du locuteur I.

THEME 1B

ROLE DE LA PROSODIE EN RECONNAISSANCE

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

CARACTERISATION DES VARIATIONS DE LA FREQUENCE DU FONDAMENTAL DANS LES PHRASES FRANCAISES

Jacqueline Vaissière

Laboratoire de Recherche en Electronique
M.I.T.

RESUME

Cet exposé traite des rapports entre la structure grammaticale de la phrase française et l'évolution dans le temps des valeurs prises par la fréquence laryngienne.

Le contour de fréquence larygienne des mots intégrés dans une phrase est representé par l'un de quatre contours typiques et les rapports entre syntaxe et fréquence laryngienne sont décrits à l'aide de ces quatre contours typiques.

SUMMARY

This paper concerns the relationships between grammatical structures of the French sentences and their fundamental frequency contours.

Any fundamental frequency contours of the words are schematized using one of four typical patterns and the relationships are described in terms of these patterns.

CARACTERISATION DES VARIATIONS DE LA FREQUENCE DU FONDAMENTAL DANS LES PHRASES FRANCAISES

INTRODUCTION

Cet exposé présente deux points importants se rapportant à l'étude des paramètres prosodiques dans les phrases françaises. Il concerne d'une part le système de notation utilisé pour décrire l'évolution actuelle des valeurs prises par la fréquence du fondamental(F_o) au cours de la phrase, et d'autre part, l'interprétation du rôle et les limites de la quantité d'information apportée par les variations de F_o dans les phrases prononcées sans insistance particulière sur certains mots.

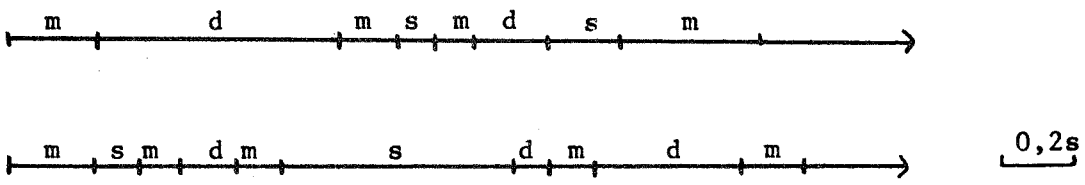
1. SYSTEME DE NOTATION:

Nous allons présenter dans un premier temps le système de notation que nous avons adopté pour la description des courbes de F_o. Ce système est inspiré des expériences perceptuelles conduites sur le hollandais par Cohen et n'Hart, système qui a été développé et appliqué ensuite à la description de l'anglais américain par Maeda dans son étude physiologique: la courbe de F_o est décomposée en mouvements successifs et est décrite de facon très simple par des attributs représentant la direction des mouvements réalisés (telle que "montée", "descente", ou "plateau"). Nous avons adapté ce système à la description des phrases francaises et introduit une notion fondamentale supplémentaire: en établissant une différence entre une montée des valeurs de F_o percue par l'auditeur comme une intonation montante (et qui concerne essentiellement une voyelle) d'une part, et d'autre part, une montée (ou augmentation des valeurs) qui est percue comme un simple changement de hauteur entre deux syllabes (consécutives ou non).

1.1. Les trois mouvements de base:

La description des courbes réelles part de l'idée fondamentale que le locuteur ne peut finalement réaliser que trois mouvments de F_o, malgré la complexité des mouvements de vibration des cordes vocales et du mécanisme de controle sous-jacent): nous pouvons soit augmenter volontairement le rythme de vibration durant certaines zones de la phrase, soit les diminuer, ou encore essayer de garder des valeurs de Fo à peu près constantes d'une syllabe à l'autre. Désignons par les lettres "m" (montée), "d" (descente) et "s" (soutenir) les trois mouvements.

Par exemple, les courbes de Fo représentées sur la figure 1 (voir page suivante) peuvent être décrites en fonction du temps de la façon suivante:



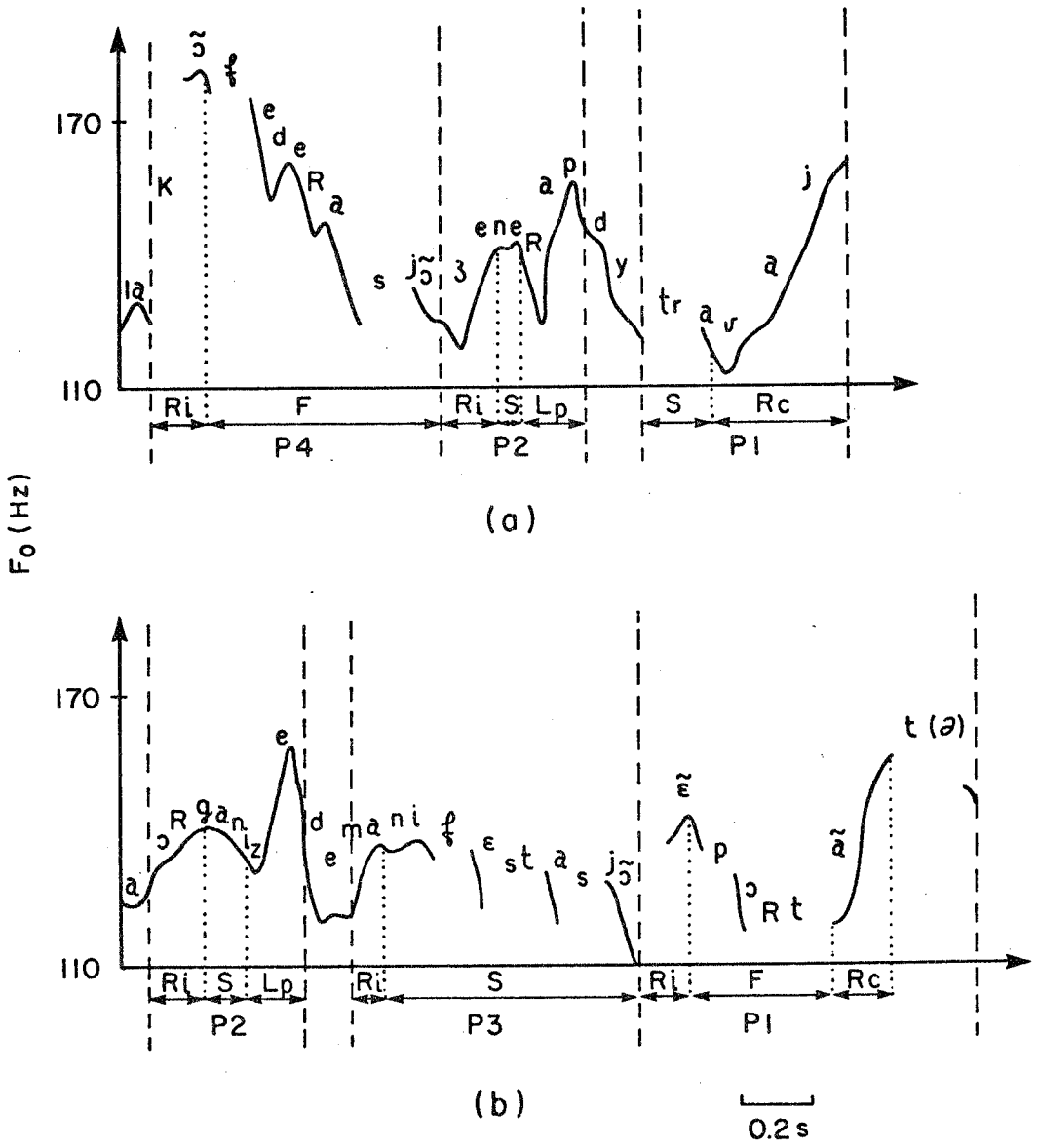


Figure 1: Fréquence du fondamental de la proposition indépendante:
"La confédération générale du travail a organisé des manifestations importantes..."
prononcée par un locuteur masculin.

Rappelons que nous ne tenons pas compte des déviations dues à la nature des consonnes sous-jacentes (micromélodie), étant donné que les variations rapides de Fo durant ces consonnes ne sont pas le résultat d'un contrôle conscient par le locuteur.

1.2. Adaptation des trois mouvements de base à la description de phrase francaise; les six attributs.

Les trois mouvements précédents ont été subdivisés selon la longueur de la chaine et l'identité des phonèmes qui leur correspondent. Six attributs nous ont paru nécessaires pour une description adéquate des courbes de Fo en français (dans les phrases prononcées sans insistance).

α) Montée initiale et montée finale: Ri et Rf

Il y a une différence essentielle entre la montée qui caractérise le début du mot (comme par exemple dans les mots "confédération", "générale", "organisé", "manifestations" et "importantes") et la montée à la fin d'un mot (comme par exemple sur la syllabe "vail" du mot "travail"). Dans le premier cas, l'oreille interprète l'augmentation des valeurs de Fo comme un changement de hauteur (comme par exemple, entre le mot grammatical "le " et la première syllabe du mot lexical "confédération"); dans le second cas, l'oreille interprète la montée des valeurs comme une intonation montante sur la syllabe (cf: la montée de continuation de Delattre).

La première montée est notée Ri (de l'anglais "initial rise") et la montée de second type est désignée par Rf (de l'anglais "final rise"). La nature et le nombre des phonèmes correspondant à Ri ont été décrit en détail dans une publication récente (Vaissiére, 1975). Quant à Rf, il concerne au moins la voyelle de la dernière syllabe et éventuellement la ou les consonnes qui la precèdent dans cette même syllabe.

β) Descentes: F, L et Lp

Le troisième attribut est F (de l'anglais "fall"). F désigne un mouvement faisant directement suite à Ri, comme dans le mot "confédération").

Les quatrième et cinquième attributs sont L (de l'anglais "lowering") et Lp (de l'anglais "lowering associated with peak"). Par exemple, L concerne la syllabe "tions" du mot "manifestations" et Lp concerne les syllabes "rale" et "sé" des mots "générale" et "organisé".Lp caractérise une chute rapide des valeurs de Fo, précédée par une montée rapide de ces mêmes valeurs au début de la syllabe: la succession de ces deux mouvements forme un pic caractéristique. Quant à L, il succède à S, sans être prècédé par une montée.

γ) Plateau: S

Le sixième et dernier attribut est S (de "soutenir"). S désigne une zone où les valeurs restent à peu près stationnaires et S fait suite en général à un mouvement Ri.

En résumé, les courbes de la figure 1 peuvent être representées de la facon indiquée sur la partie inférieure de la figure. Nous voudrions enfin noter que les symboles Ri, Rf, F, Lp, L et S, dérivés de mots

anglais n'ont pas été "traduits" en français (par "Mi", "Mf", "D" etc..) afin de faciliter des comparaisons (que nous ferons ultérieurement dans un prochain article) du français avec d'autres langues.

2. CONTOURS DE BASE:

L'utilisation systématique des attributs décrits précédemment ont permis de découvrir quatre contours caractéristiques au niveau des mots et deux contours au niveau des groupes de mots.

2.1. Contours de mots

Reprenons l'exemple de la figure 1. Nous pouvons remarquer que:

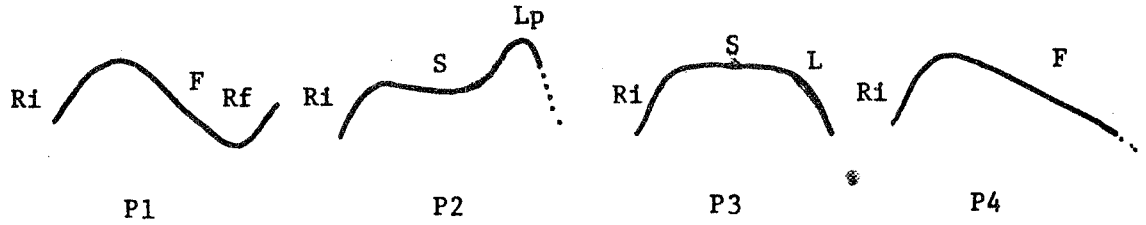
confédération	:	Ri	+	F		
générale	:	Ri	+	S	+	Lp
travail	:			S	+	Rf
organisé	:	Ri	+	S	+	Lp
manifestations	:	Ri	+	S	+	L
importantes	:	Ri	+	F	+	Rf

L'étude des phrases isolées et des textes lus par tous nos locuteurs (actuellement au nombre de 14) nous ont conduite à la conclusion que seul un nombre limité de combinaisons de mouvements est permise en francais, tout au moins pour les mots longs (nous appelons "mot long" un mot de trois syllabes au moins).

Ces combinaisons sont (en allant du contour le plus montant au plus descendant):

contour 1	:	Ri	+	F	+	Rf	(ex: importantes)
contour 2	:	Ri	+	S	+	Lp	(ex: organisé, générale)
contour 3	:	Ri	+	S	+	L	(ex: manifestations)
contour 4	:	Ri	+	F			(ex: confédération)

La figure suivante schématise les quatre contours de base (P1, P2, P3 et P4):

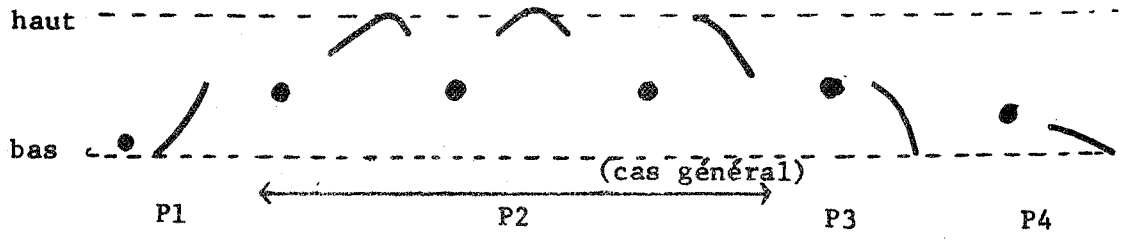


P1 et P4 sont les deux schémas situés aux extrèmes, le premier étant perçu comme un mot à intonation montante et le second comme un mot à intonation descendante. Le schéma P4 est le plus simple des quatre (car il ne contient que deux attributs). Nous l'utilisons pour donner une réponse à une question, ou encore á la fin d'une phrase (cf l'intonation appelée "finalité" par Delattre) et le schéma est également attesté à l'interieur d'une phrase (comme pour le mot "confédération" de la figure 1).

P1 est également très courant. Nous l'utilisons souvent pour poser une question, et à la fin de nos groupes de mots en position non finale dans une phrase (cf l'intonation appelée "continuation majeure de Delattre).

P2 et P3 appartiennent au domaine de la phrase et ne peuvent être prononcés de facon consciente aussi aisement que P1 et P4. P2 s'oppose à P1 par les valeurs de Fo sur la syllabe précédant Lp ou Ri: dans le contour 1, les valeurs diminuent vers le registre bas du locuteur avant de remonter, alors que dans le contour 2, Lp (Lp = m + d) n'est pas précédé d'une chute vers les graves. Dans le contour P2, les valeurs restent à peu près stationnaires après Ri, ou elles déclinent legerement. De plus, Rf dans le contour P1, affecte obligatoirement la voyelle. Dans le contour P2, la montée m de Lp s'effectue en général durant la première consonne de la dernière syllabe, la voyelle étant affectée par un mouvement descendant.

La figure suivante résume les observations précédentes et illustre les valeurs de fréquence du fondamental au niveau de l'avant-dernière syllabe (notée par "●"), et durant la voyelle de la dernière voyelle.



P3 diffère de P2 par le fait que la dernière syllabe ne contient pas de valeurs de Fo supérieures à celles de l'avant dernière syllabe. P3 s'oppose à P4 par le fait que dans P4, les valeurs commencent à decliner dès le début du mot (apres Ri), alors que dans P3, les valeurs tendent à rester stationnaires après Ri.

La substitution de mots courts à des mots longs dans des phrases type nous a permis d'étudier le phénomène de réduction de ces schémas de base dans les mots courts. Dans les mots courts, les schémas P1, P2 , P3 et P4 (dans les monosyllabiques) dégénèrent en contours simplifiés où certains attributs ont été supprimés. Dans le mot "travail" représenté à la figure 1, par exemple, Ri a été supprimé par le locuteur. D'autres locuteurs conservent cette montée Ri et nous avons la distribution suivante:

/tr a vaj/ (transcription phonétique)
Ri F Rf

En résumé, les courbes de la figure 1 peuvent être representées de la facon suivante:

$$\{P4+P2+P1+pause+P2+P3+P1+pause\}$$

2.2 Contours de groupes de mots

A un niveau supérieur, celui des groupes de mots, on peut également schématiser le contour général de Fo.

Deux schémas seulement ont été attestés: le premier P1g dont l'allure générale s'apparente à P1, et le second P4g, qui s'apparente à P4.

$P1^g$ est composé des attributs suivants:

$$P1^g = Ri^g + F^g + Rf^g$$

la valeur maximale de Fo durant le groupe étant située à la frontière de Ri^g et de F^g. La valeur mininale est située à la jonction des mouvements F^g et Rf^g. Les figures de la page suivante illustrent les contours des groupes de mots non situés en fin de phrase et prononcés par deux locuteurs differents. Ces figures permettent d'observer des variations sensibles entre les valeurs des maximas et des minimas d'un groupe à l'autre chez un même locuteur.

Rf^g est équivalent au Rf du dernier mot de groupe (qui a un contour P1). Quant a Rig, il est égal ou supérieur au Ri du premier mot dans le groupe. Si le premier mot a un pattern P2, Rig est dans ce cas égal à la somme de Ri et Rf du premier mot. Si le premier mot a un contour P4 ou P3 (P3 est rare pour un premier mot), Rig est équal dans ce cas au Ri du premier mot et on observe que le Ri du premier mot du groupe est plus important que les autres Ri dans les mots suivants et ne peut en aucun cas être supprimé. Un mot de deux syllabes aura donc une tendance très nette à recevoir un P2 en début de groupe, les valeurs de Fo augmentant du début du mot jusqu'à un point situé dans la derniere syllabe.

On peut noter que plus la distance entre la valeur maximale du groupe et la valeur minimale est longue, et plus le locuteur aura la possibilité d'insérer une série de montées et de descentes. Si cette distance devient minimale du fait d'un nombre de syllabes peu élevé ou d'un rythme trop rapide, tout mouvement intermédiaire (entre le premier sommet de Fo et la dernière syllabe) devient impossible. Les informations prosodiques livrées par les variations de Fo à l'intérieur d'un groupe sont en quelque sorte inversement proportionnelles à la longueur du groupe: ceci est compensé par le fait que plus les groupes sont courts et moins l'auditeur a besoin d'être "aidé" dans sa segmentation du groupe en mots.

$P4^g$ est composé des attributs suivants:

$$P4^g = Ri^g + F^g$$

Une phrase prononcée sans pause est affectée d'un contour $P4^g$. Les valeurs de l'attaque, du maximum de Fo et du minimum ont été étudiés sur un corpus de 45 phrases enregistrées par 10 locuteurs (Voir Larreur et Boé). La valeur maximale est située en général sur le premier mot du groupe. Le dernier mot du groupe a un contour P4, et l'avant dernier mot a en général un contour P2: la combinaison des contours P2 et P4 donne à la fin de la phrase française sa forme caractéristique en demi cercle.

La description des deux contours des groupes précédents est apparentée à celle des "tunes" 1 et 2 de Armstrong et Ward (pour l'anglais), et aux "groupes de souffle" marqués (P1g) et non marqués (P4g) de Lieberman.

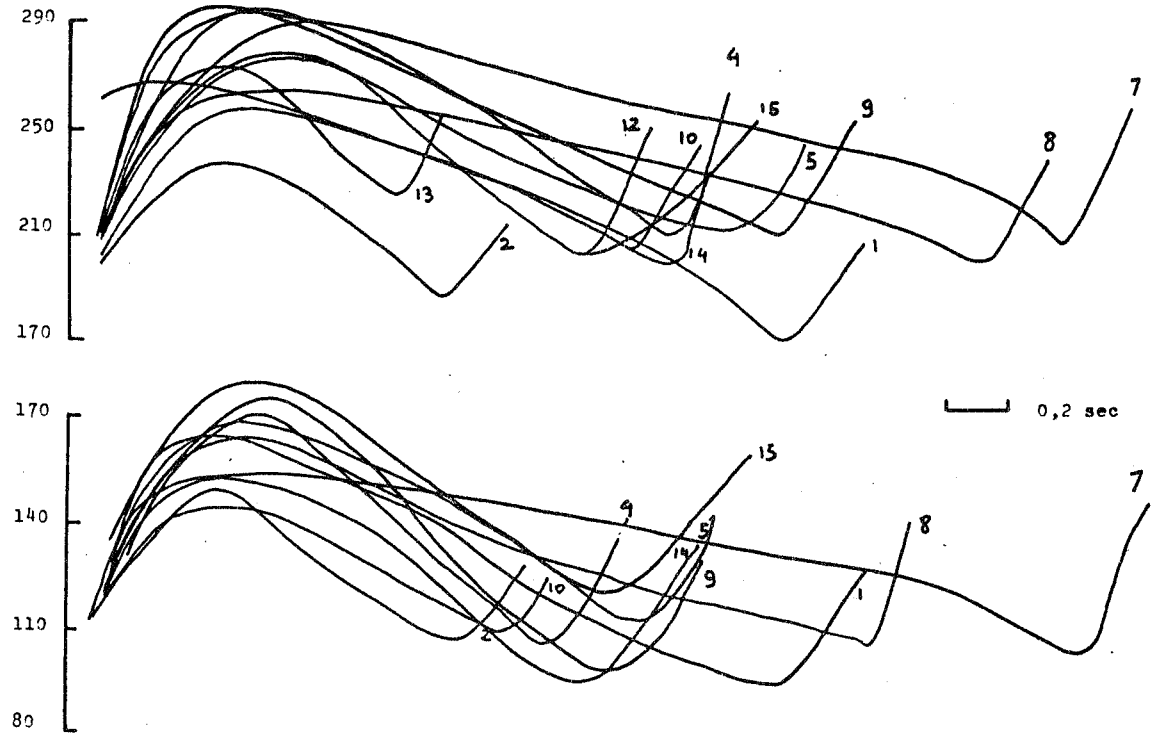


Figure 2: Contours généraux de fréquence du fondamental superposés à des groupes de mots. Ces groupes de mots sont extraits du début d'un texte lu par deux locuteurs. Les chiffres réfèrent aux groupes de mots indiqués ci-dessous.

DEBUT DU TEXTE:

On a depuis longtemps essayé de faire parler des machines. (1 Ce courant a été inauguré à la fin du XVIIIième siècle)(2 par les essais de Van Kempelen) et de l'Abbé Rousselot. (4 Par des mécanismes souvent astucieux)(5 il fut possible de reproduire certains sons), puis certains enchaînements de sons.

(7 Grâce aux récents progrès dans les techniques de l'analyse du signal) (8 et surtout grâce à l'apparition des gros ordinateurs numériques), (9 la reconstitution automatique de la parole)(10 a réalisé des progrès considérables) depuis une dizaine d'années. (12 La technique devance notre savoir)(13 et la synthèse par règles) (14 n'a pas encore eu d'applications concrètes) (15 car on n'a pas encore su donner à la machine) tous les éléments qui lui sont nécessaires pour bien parler.

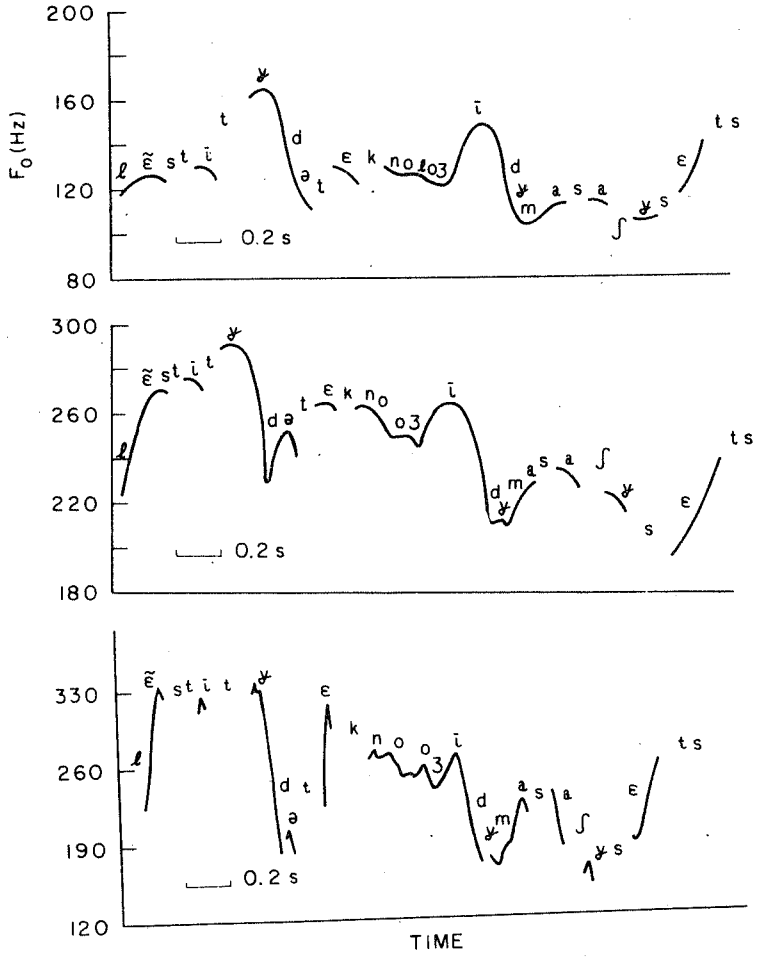
3. INTERPRETATION

La question est de savoir pourquoi un locuteur utilise un contour plutot qu'un autre pour un mot à un endroit précis dans la phrase et pourquoi tous les locuteurs n'utilisent pas le même contour pour le même mot dans la même phrase?.

C'est cette double question qui a été l'objet de nos recherches au cours de ces derniers mois et nous sommes arrivée aux deux conclusions suivantes:

3.1. Variantes individuelles.

En analysant une suite de mots peu ambigüe telle que :"L'institut de technologie du Massachusetts" (sujet)dans la première phrase du premier texte que nous avons fait lire par 14 locuteurs, nous avons eu la surprise de découvrir que les locuteurs utilisaient des contours différents. La figure suivante illustre trois réalisations différentes de ce groupe de mot, les deux premières pouvant être représenter par (P2 +P2 + P1) (noter les différences de rapport entre Ri et Rf dans les deux cas) et la dernière par (P3 +P3 +P1). La suite (P4 +P4 +P1) a également été attestée.



En analysant la suite du texte, il fut aisé de découvrir que chaque locuteur emploie un contour plutôt qu'un autre (mis a part naturellement les contours P1 et P4 qui sont définis par la position du mot dans le groupe: le locuteur n'a pas le choix entre plusieurs contours pour un mot situé en fin de groupe, avant une pause, sauf si ce mot est situé en fin d'indépendante en position non finale dans la phrase, où il a le choix entre P1 et P4). Le texte ayant été prononcé dans les mêmes conditions par tous les locuteurs, nous désignons ce genre de variations comme des variantes stylistiques ou individuelles.

3.2. Nuclear Stress Rule en Francais.

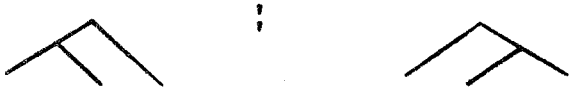
Imaginons un groupe composé de trois mots et représentons par Px, Py et Pz les contours des premier, second et troisième mot, respectivement. Nous savons déjà que Pz est déterminé par la position du groupe dans la phrase: Pz = P1 dans le cas d'un groupe terminé par une intonation montante et Pz = P4 dans le cas d'un groupe terminé par une intonation descendante.

Deux cas peuvent se produire. Soit x = y (comme dans la figure précédente), ou soit x ≠ y (comme dans la figure 1). Dans le premier cas, nous parlerons d'une structure parallèle, et dans le second cas, d'une structure marquée.

Une structure prosodique parallèle est en général utilisée pour les suites de mots désignant une institution (telle que "l'institut de technologie du Massachusetts") ou encore une organisation, c'est-à-dire une suite de mots usuelle ("la confédération générale travail" a été generalement prononcée avec une structure prosodique parallèle).

Quant aux structures prosodiques marquées, nous pouvons les diviser en deux groupes. Dans le premier groupe, x est supérieur a y (l'intonation du premier mot est plus "descendante" que celle du second mot) et dans le second cas, y est supérieur à x. Le schéma suivant illustre cette division:

x = y	x ≠ y	
Structure prosodique parallèle.	Structure prosodique marquée	
	x > y	x < y
	a gauche	a droite
P4 P4 Pz	P4 P3 Pz	P3 P4 Pz
P3 P3 Pz	P4 P2 Pz	P2 P4 Pz
P2 P2 Pz	P3 P2 Pz	P2 P3 Pz



L'examen des structures grammaticales sous-jacentes aux structures prosodiques marquées nous a permis de constater que dans le premier cas (x > y), la structure grammaticale avait un noeud à gauche (comme dans le premier groupe de mot représenté à la figure 1a); dans le second cas (x < y) le noeud se trouve à droite (comme dans le deuxième groupe de mot répresente sur la figure 1b).

La combinaison des contours (et non les contours eux-mêmes) ont donc un rôle grammatical à assumer. Néanmoins le schéma indiqué à la page precédente n'est pas suivi régulièrement par tous les locuteurs. Certains ont recours à un cinquième contour, PØ (P zéro) , où Ri est supprimé, pour differencier des paires de phrases du type:

Un professeur de géographie du Canada

qui peut être interprété de deux facons différentes (le professeur enseigne la géographie du Canada, d'une part, et le professeur vient du Canada , d'autre part).

Trois locuteurs français ont prononcé cette phrase, en essayant de marquer une différence entre les deux interprétations. Deux de ces locuteurs ont suivi les règles décrites précédemment, alors que le troisième a utilisé PØ pour le second mot dans le second cas (le professeur vient du Canada). La comparaison de cette phrase prononcée par ce dernier locuteur avec cette même phrase prononcée par un locuteur employant P3 P2 P1 indique à l'auditeur que le troisième locuteur considère le mot "géographie" comme non important (sa phrase est equivalente a : un professeur du Canada) (cf la notion de "parenthèse" de Delattre), alors que l'autre locuteur regroupe les deux premiers mots "professeur" et "géographie" en une unité se terminant par une sorte d'accent secondaire sur la syllabe "phie". Cet exemple illustre la remarque de Faure qui considère que le "relief" prosodique d'un relief est proportionnel à la quantité d'information convoyée par ce mot dans la phrase. Nous espérons pouvoir éclaicir dans les mois à venir la question du rôle exact joué par la syntaxe de la phrase, sa sémantique et le contenu informationnel des mots qui la constituent.

CONCLUSION

Comme le note Chomsky, le problème pour le linguiste est de déterminer à partir des données de la performance (c'est-à-dire l'utilisation actuelle que font les locuteurs de leur langue - au sens de language- dans des situations concrètes) le systéme sous-jacent de règles possédées par les auditeurs-locuteurs et qu'ils utilisent dans leur performance. En appliquant cette théorie à la prosodie, nous espérons pouvoir définir une grammaire générative de la prosodie (rendant compte des courbes de variations de la fréquence du fondamental et des variations de durée entre les phonèmes). Cette grammaire décrirait l'utilisation actuelle que font les français de la possibilité intrinsèque qu'ils ont de faire varier constamment le rythme de vibrations de leurs cordes vocales durant l'acte de parole (Fo) et d'accomplir les mouvements articulatoires correspondant aux phonèmes plus ou moins rapidement.(durée des phonèmes).

Cette grammaire ne peut être complète et même valide que si elle comprend également une composante perceptuelle. Nous allons donc mener dans les mois prochains une étude perceptuelle systématique sur les

attributs intonationnels décrits dans cet exposé. Cette étude sera réalisée à l'aide de synthèse partielle et peut-être à l'aide de synthèse complète, comme nous l'a aimablement proposé H. Tzeuschler (qui a realisé un programme de synthèse complète du français). Une étude perceptuelle parallèle à nos travaux (et décrivant également les variations de la fréquence laryngienne en termes d'attributs et de contours de mots) sera également menée au sein du Speech Communication Group par B. Delgutte.

Références:

Armstrong L.E. et Ward I.C., Handbook of English Intonation, Cambridge, Heffner (2ieme edition), 1926.

Chomsky N., Aspects of the Theory of Syntax, The M.I.T. Press, 1965.

Chomsky N. et Halle M., The Sound Pattern of English, New York, Harper & Row, 1968.

Delattre P., La lecon d'intonation de Simone de Beauvoir, étude d'intonation déclarative comparée, French Review, Vol.35,1961.

Delattre, P., Les dix intonations de base du Francais, French Review, Vol.41, 1966.

Faure G., Accent, Rythme et Intonation, dans la Grammaire du Francais parlé, A. Rigault, 19

Hart J.T. et Cohen A., Intonation by Rule, a perceptual quest, Manuscript 242/11, Institut de Recherche sur la perception, Eindhoven, 1973.

Larreur D., et Boé J.L., Synthèse paramétrique de la phrase énonciative en Français, 5iemes Journees d'Etudes sur la Parole, Orsay, 1974.

Lieberman P., Intonation, Perception and Language, The M.I.T. Press, 1967.

Maeda S., An Electromyographic Study on Intonational Attributes, Quaterly Progress Report, M.I.T., Research Laboratory of Electronics, January 1975.

Tzeutschler H.S., Terminal Analog Synthesis of French, J.A.S.A., S56, Vol. 55, Spring 1974.

Vaissière J., On French Prosody, Quaterly Progress Report, Research Laboratory of Electronics, M.I.T., Juin 1974.

Vaissière J., Further Note on French Prosody, Quaterly Progress Report, Research Laboratory of Electronics, M.I.T., Janvier 1975.

6èmes JOURNEES D'ETUDE SUR LA PAROLE
Toulouse 28 au 30 Mai 1975

INTONATION ET RECONNAISSANCE AUTOMATIQUE DE LA STRUCTURE SYNTAXIQUE

Philippe MARTIN
Laboratoire de phonétique expérimentale
Université de Toronto

RESUME

On décrit ici différents algorithmes pour la reconnaissance automatique de la structure syntaxique de l'énoncé utilisant totalement ou partiellement les indications contenues dans la mélodie de la phrase. L'élaboration de ces algorithmes s'appuie sur une théorie nouvelle présentée sous forme de grammaire générative de l'intonation.

SUMMARY

Pattern recognition algorithms for syntactic structures, using sentence melody totally or partially, are described. These algorithms are based on a new theory presentes here in the form of an intonation generative grammar.

INTONATION ET RECONNAISSANCE AUTOMATIQUE DE LA STRUCTURE SYNTAXIQUE

Ph. MARTIN

0. A leurs débuts, les méthodes de reconnaissance automatique de la parole ont été basées essentiellement sur l'identification séquentielle des éléments segmentaux de la chaîne parlée. La nécessité d'intégrer de tels processus, relativement simples, dans un cadre plus large tenant compte de la syntaxe des énoncés à reconnaître s'est rapidement fait sentir.

Les éléments suprasegmentaux, déjà partiellement mis en œuvre pour l'identification des phonèmes, s'avèrent alors particulièrement intéressants s'ils sont considérés dans leur rôle d'indicateur de la sructure syntaxique de la proposition. Il s'agit alors, dans cette perspective, d'une part d'en établir par analyse linguistique la fonction de corrélation avec les caractéristiques syntaxiques de l'énoncé, et, d'autre part, d'utiliser ces propriétés pour l'élaboration d'algorithmes de reconnaissance de la structure syntaxique.

1. L'hypothèse de base qui préside à l'utilisation des faits prosodiques dans la reconnaissance automatique de la parole porte sur l'existence d'une corrélation entre ces faits et la structure syntaxique de l'énoncé. Cette corrélation a été pressentie depuis longtemps (cf. par ex. Gleason, 1967), sans cependant être précisée de manière convaincante.

On s'appuiera alors ici sur une théorie nouvelle proposée récemment (Martin, 1974, 1975a, 1975b). Cette théorie fait apparaître une corrélation entre la séquence de contours mélodiques de la phrase et le classement hiérarchique des unités minimales de sens de la proposition, phrase et proposition étant respectivement l'expression et le contenu de l'énoncé. Ces contours manifestent des marques de présupposition des éléments minimaux appartenant à la phrase, et constituent la contrepartie des marques du contenu à même fonction que sont les indicateurs syntaxique tels que flexions, prépositions, etc. Plutôt que d'exposer, dans leur cadre théorique original (à orientation glossématique), les résultats obtenus à partir d'une telle hypothèse, on en donnera ici une formulation concise utilisant les notations habituelles en phonologie générative.

2. En termes génératifs, une grammaire de l'intonation vise à dériver une séquence de contours mélodiques à partir d'une repré-

sentation de la structure superficielle de l'énoncé.

Structure superficielle ——→ Séquence de contours mélodiques
Grammaire

Soit donc un certain énoncé, pourvu d'une structure superficielle décrite par le parenthésage:

(((le)(gros)(chien))(blanc))((mange)((le)(fromage)))

Il faut souligner ici qu'un parenthésage donné est censé rendre compte de la structure superficielle voulue par l'usager de l'énoncé qu'est la locuteur, et non telle que conçue par un linguiste particulier.

Une première règle d'accentuation de mot (AM) repère par une marque o les syllabes accentuées dans la séquence syllabique de l'énoncé (la discussion qui suit s'applique à l'intonation du français):

AM: (((le)(gros)(chien))(blanc))((mange)((le)(fromage)))
o o o o o

Une règle d'ajustement (AJ) élimine ensuite des couples de parenthèses)(de manière à n'avoir qu'une et une seule syllabe accentuée (donc une seule marque o) par séquence située entre deux parenthèses correspondantes (...). Dans l'exemple, ((le)(gros)) devient (le gros) par suppression des parenthèses internes)(de façon à intégrer la séquence (le) dans un groupe plus grand pourvu d'une marque accentuelle. Il en va de même pour ((le)(fromage)). On supposera ici que la sructure superficielle donnée est telle que des cas d'introduction de nouvelles parenthèses)(pour séparer des marques accentuelles situées dans une même séquence ne puisse pas se présenter.

L'application de la règle d'ajustement à l'exemple donne:

AJ: (((le gros)(chien))(blanc))((mange)(le fromage))

On applique ensuite récursivement une règle d'attribution de niveaux (NIV), qui, à chaque cycle, modifie une des marques accentuelles o de départ par élimination d'un ou de plusieurs couples de parenthèses intérieures appartenant à un même niveau:

NIV: ((...V)(...V) . . . (...V)) ——→ (...V...V...V...V)
x x y x+1 x+1 x+1 y

V représente la syllabe accentuée du groupe;
x représente le numéro d'ordre de la marque accentuelle.

Sur l'exemple:

(((le gros)(chien))(blanc))((mange)(le fromage))
o o o o o

on a successivement:

NIV:	cycle 0	(((		o	)(		o	))(		o	))((	o	)(			o	))
	cycle 1	((		1		o	)(		o	))(	1				o	)
	cycle 2	(		2		1		o	)(	1				o	)	
	cycle 3			3		2		1		2				o		

A chaque cycle, donc à chaque suppression des parenthèses les plus intérieures, toutes les marques accentuelles intérieures à la nouvelles séquence constituée sont augmentées d'une unité, sauf la dernière.

Enfin, un groupe de règles d'attribution des contours mélodiques (CMEL) assigne à chaque syllabe repérée par une marque accentuelle déterminée un contour spécifique choisi parmi les 8 contours mélodiques suivants:

C_d ⟍ long descendant (contour final de l'énoncé corrélatif de la déclaration).

C_i ⟋ long montant (contour final de l'énoncé corrélatif de l'interrogation).

C_n ⟍⟋ long neutre (syncrétisme de C_d et de C_i).

C_1 ⟋ ou ⟍ court, ample et α montant (α = + ou - suivant le sens de la pente du contour final C_d ou C_i respectivement).

C_2 ⟍ ou ⟋ court, ample et α descendant ($-\alpha$ montant).

C_3 ⟋ ou ⟍ court, restreint et α montant.

C_4 ⟍ ou ⟋ court, restreint et $-\alpha$ montant.

C_s — court et neutre (syncrétisme de C_1, C_2, C_3 et C_4).

La description de ces 8 classes de contours par une matrice phonologique utilisant les traits ± montant, ± ample et ± long est donc:

	C_d	C_i	C_n	C_1	C_2	C_3	C_4	C_s
Montant	-	-	±	$-\alpha$	$+\alpha$	$-\alpha$	$+\alpha$	±
Ample	(+)	(+)	(+)	+	+	-	-	±
Long	+	+	+	-	-	-	-	-

L'utilisation du trait α montant pour caractériser les contours non terminaux permet de rendre compte du renversement de pente de ces contours lorsqu'un même énoncé passe d'une modalité déclarative à une modalité interrogative et inversément.

Les règles d'attribution des contours sont:

CMEL:

(1) $o \rightarrow C_o$

$C_o \rightarrow C_d$ / énoncé déclaratif

C_i / énoncé interrogatif

C_n / énoncé déclaratif ou interrogatif dont la modalité est indiquée par une marque du contenu

(2) $1 \rightarrow C_1$

(3) $2 \rightarrow \begin{cases} C_2 \ / \ \text{——} \ C_1 \\ C_3 \ / \ \text{——} \ C_o \end{cases}$

(4) $3 \rightarrow \begin{cases} C_3 \ / \ \text{——} \ C_2 \\ C_4 \ / \ \text{——} \ C_1 \\ C_s \ / \ \text{——} \ C_o \end{cases}$ (saturation)

(5) $4 \rightarrow \begin{cases} C_4 \ / \ \text{——} \ C_3 \\ C_s \ / \ \text{——} \ C_x \end{cases}$ ($x \neq 3$, saturation)

A ce groupe s'ajoute une règle d'attribution d'un contour C_o supplémentaire, non terminal, et indiquant la division de l'énoncé en propos et thème:

(6) $C_o \rightarrow C_o \ C_n$

Elle s'applique par exemple dans:

(c'est le château)(que j'ai acheté)

o o

C_d C_n

qui s'oppose à l'énoncé dépourvu de thème:

(c'est le château)(que j'ai acheté)

1 o

C_1 C_o

L'exemple " le gros chien blanc mange le fromage " soumis aux règles CMEL donne: 3 2 1 2 o

Marques o: C_d

1: C_1

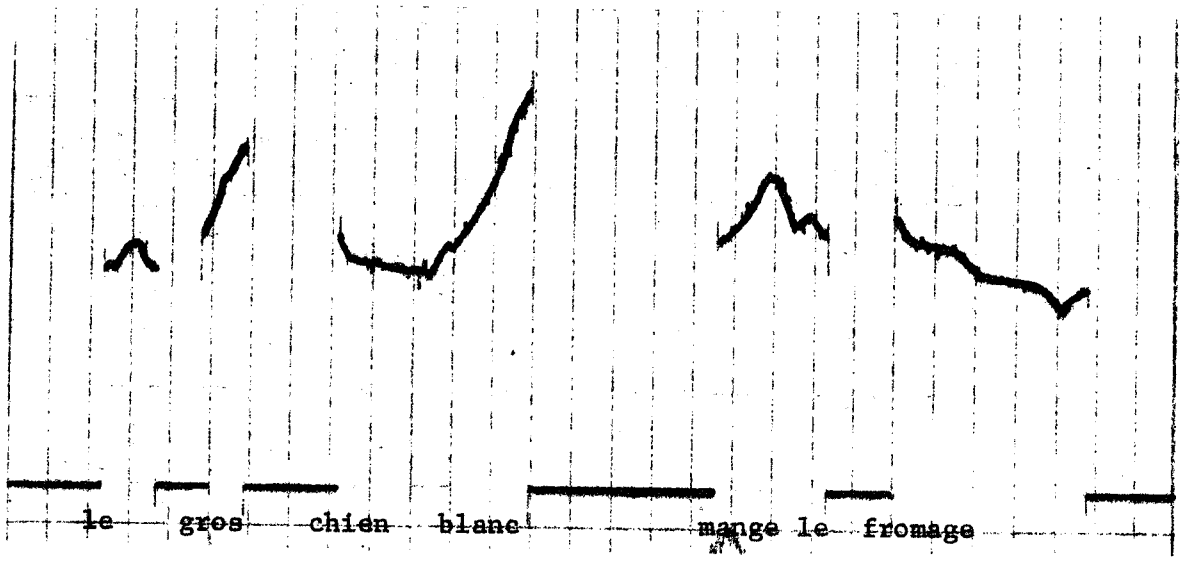
2: C_2 C_3

3: C_3

et la séquence de contours est:

le gros chien blanc mange le fromage
C_3 C_2 C_1 C_3 C_d

La courbe mélodique expérimentale qui correspond à cet énoncé est:



(Ce tracé a été obtenu par l'analyseur de mélodie décrit dans Léon et Martin, 1969).

En présence d'autres indicateurs de la strucure superficielle en un point de l'énoncé, un contour déterminé entre en syncrétisme:

$C_x \rightarrow C_s$ / —— autre marque syntaxique

L'utilisation des contours dans une procédure de reconnaissance automatique implique donc 1) soit que ces syncrétismes soient tous résolus, c'est-à-dire que les contours, même redondants, soient réalisés; 2) soit que les marques du contenu indiquant la structure syntaxique puissent être utilisés en même temps que les marques constituées par les contours mélodiques de manière à les écarter du processus lorsqu'ils entrent en syncrétisme.

4. La reconnaissance automatique de la structure syntaxique de l'énoncé revient à classer hiérarchiquement les éléments d'une chaîne à partir des indications contenues dans ces éléments:

chaîne ⟶ structure syntaxique

Ces indications peuvent être déterminées à partir 1) du seul contenu de l'énoncé (cas de l'analyse traditionnelle), 2) de l'expression seulement (c'est-à-dire par la séquence de contours mélodiques), 3) ou encore à partir du contenu et de l'expression. On envisagera successivement les deux derniers cas.

D'une manière générale, le problème de la reconnaissance des contours mélodiques de l'énoncé pour obtenir une représentation de sa structure superficielle s'articule de la façon suivante:

a. analyse de la mélodie à partir du signal de parole;
b. localisation des contours dans la chaîne;
c. reconnaissance des contours;
d. établissement de la structure superficielle.

Le premier point, qui revient à la mesure de la fréquence fondamentale du signal de parole en fonction du temps, a reçu diverses solutions plus ou moins satisfaisantes (pour une étude comparative, voir Léon et Martin, 1969).

La localisation des contours dans l'énoncé revient à identifier les syllabes accentuées, porteuses de ces contours. Un algorithme simple a été proposé par Lea (1974) pour l'anglais, et se base presque exclusivement sur la détection des sommets d'intensité. En fait, il semble qu'un processus plus élaboré doive être utilisé, en tenant compte de la prépondérance du contour mélodique lui-même dans la perception de l'accentuation d'une syllabe, à côté des facteurs prosodiques traditionnels que sont l'intensité et la durée. Un mécanisme de reconnaissance de type perceptron avec procédure d'apprentissage et utilisant comme paramètres d'entrée l'énergie intégrée, combinant intensité et durée, ainsi que la variation de fréquence fondamentale, devrait remplir ce rôle.

La reconnaissance des contours, assimilés à un tronçon de droite dans un plan fréquence fondamentale-temps, peut se faire en utilisant les surfaces de séparation simples définies par des traits phonétiques résultant d'une procédure d'apprentissage et correspondant aux traits phonologiques binaires Long, Ample et Montant. Le trait d'amplitude est représenté par la projection du contour sur l'axe des fréquences, lessens de la pente définit le quadrant où se place le contour, enfin, le trait de longueur détermine une circonférence limite à l'intérieur ou à l'extérieur de laquelle est située l'extrémité du contour.

5. Au lieu d'identifier chaque contour mélodique séparément pour .tablir la structure superficielle de l'énoncé, on peut envisager des algorithmes opérant à partir des contrastes de pente, d'amplitude et de longueur repérés dans toute l'étendue de la chaîne et non plus individuellement.

Le premier qui sera exposé ici est du type "breadth first". Il établit la structure syntaxique par balayage selon des niveaux de plus en plus profonds.

Les opérations successives sont:

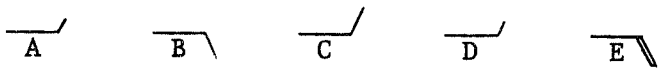
1. Identifier le premier contour à droite qui soit +Long; la pente + ou -Montant correspondant à C_i ou C_d, corrélatifs de ladernière unité

d'un énoncé interrogatif ou déclaratif (les syncrétismes sont supposés résolus et le propos et le thème sont traités de la même manière);

2. Pour chacune des séquences situées à gauche des contours C_d ou C_i (et s'arrêtant à un autre contour C_o), de pente α Montant:
   a. parmi les contours de pente $-\alpha$, identifier les contours amples C_1 par opposition aux contours restreints C_3;
   b. parmi les contours de pente $+\alpha$, identifier les contours amples C_2 par opposition aux contours restreints C_4.

3. Déterminer le parenthésage de la structure superficielle correspondant à la séquence de contours reconnue: pour chacun des contours C_i ($i = 1,2,3,4$), et en commençant par C_1, inscrire 2 parenthèses)(immédiatement à droite du symbole du contour considéré, les parenthèses de fermeture correspondantes à gauche et à droite se plaçant immédiatement à droite ou à gauche de la première parenthèse rencontrée (au début ou à la fin de la séquence si aucune parenthèse ne peut être rencontrée).

Appliqué à la séquence suivante, l'algorithme donne successivement:

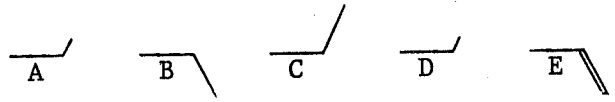


Contours -Long					+Long	$\alpha = -$
Contours $-\alpha$ Mt	$-\alpha$		$-\alpha$	$-\alpha$		
Contours +Amp			+Amp			
Contours -Amp	-Amp			-Amp		
Contours $+\alpha$ Mt		$+\alpha$				
Contours +Amp		+Amp				
Identification des contours:	C_3	C_2	C_1	C_3	C_d	
Parenthésage: 1	(		)(		)	
2	(	)(	)			
3	( )(	)	(	)(	)	
Structure obtenue :	(((A)(	B))(	C))((	D)(	E))	

6. Le deuxième algorithme, de type "depth first" se base sur la comparaison de deux contours successifs d'une séquence de manière à établir la structure superficielle au fur et à mesure de l'analyse, effectuée de gauche à droite dans la chaîne de l'énoncé.

Plus spécifiquement, les opération à réaliser sont:

- déterminer le premier contour +Long à droite (C_o), et identifier α, le signe de la pente mélodique;
- ensuite, et pour chaque couple de contours successifs C_x et C_y:
  - si les contours sont de même pente et
    - si $C_x > C_y$, c'est-à-dire si l'amplitude ou la longueur de C_x est supérieure à celle de C_y, l'unité X et les unités qui y sont déjà rattachées ne forment pas d'unité plus grande avec Y à ce stade. Donc:
      $(...X)(Y) \longrightarrow (...X)(Y)$
    - si $C_x \leqslant C_y$, l'unité X et les unités déjà rattachées forment avec Y et avec une des unités suivantes une unité plus grande. Donc:
      $(...X)(Y) \longrightarrow ((...X)(Y)$
      (inscription d'une parenthèse d'ouverture qui sera fermée plus loin).
  - si les contours sont de pente opposée et
    - si la pente de C_x est la même que celle du contour final C_o, l'unité X et les unités déjà rattachées forment avec Y une unité plus grande. Donc:
      $(...X)(Y) \longrightarrow ((...X)(Y))$
    - si les pentes de C_x et de C_o sont différentes et
      - si $C_x > C_y$, il n'y a pas de formation d'unité plus grande à ce stade:
        $(...X)(Y) \longrightarrow (...X)(Y)$
      - si $C_x \leqslant C_y$, il y a formation d'une unité plus grande incluant les éléments déjà rattachés à X, ou les éléments situés à droite d'une parenthèse non fermée:
        $(...X)(Y) \longrightarrow ((...X)(Y))$
        $((...U)(...X)(Y) \longrightarrow ((...U)(...X)(Y))$

Appliqué à la séquence 

l'algorithme réalise les opérations suivantes:

- C_o est descendant, $\alpha = -$
- C_A et C_B sont de pente différente, C_A et C_o également, et $C_A < C_B$. Donc $(A)(B) \rightarrow ((A)(B))$.
- C_B et C_C présentent des pentes différentes, C_B et C_o également. Donc on a $(((A)(B))(C))$.
- C_C et C_D sont de même pente et $C_C > C_D$. On a donc: $(((A)(B))(C))(D)$.

- C_D et C_E sont de pente opposée, C_D et C_o également, et $C_D < C_E$. Donc : (...)((D)(E)).

La structure superficielle finalement obtenue est donc:
(((A)(B))(C))((D)(E)).

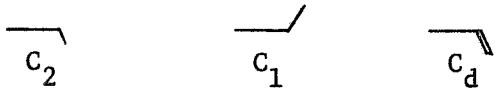
7. On évoquera pour terminer quelques aspects de la reconnaissance automatique de la structure syntaxique à partir d'indicateurs appartenant aussi bien au contenu (flexions, prépositions, articles, etc.)qu'à l'expression (les contours mélodiques). Les cas évoqués permettront de mettre en lumière le mécanisme par lequel des contours constituent un facteur de désambiguation de structures dans lesquelles les marques du contenu habituellement utilisées n'assument plus leur fonction. Les contours impliqués se révèlent alors comme dominants et, partant, se réalisent selon les règles vues plus haut sans entrer en syncrétisme.

Un premier exemple d'ambiguité est celui présenté par la séquence (A)(B)(C) lorsque les deux derniers éléments B et C contiennent des marques de sélection à gauche. L'unité C peut alors former une unité plus grande aussi bien avec B seul qu'avec A et B. Les deux structures syntaxiques correspondantes sont ((A)(B))(C) et (A)((B)(C)).

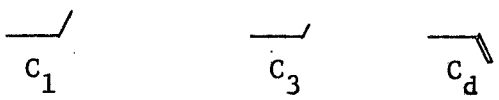
Ainsi dans (un professeur)(de football)(américain), il peut s'agir d'un professeur américain enseignant le football (première structure) ou d'un professeur enseignant le football américain (deuxième structure). De même, dans l'énoncé (je vais acheter)(le manteau) (que tu as essayé)(pour Pierre), le manteau en question est destiné à Pierre (premier cas), ou est qualifié par le fait qu'il a été essayé pour Pierre.

En l'absence de marques appartenant aux circonstances, l'ambiguité sera levée par des contours mélodiques correspondant à la structure voulue par le locuteur. Dans l'exemple ci-dessous, les contours des unités A et B correspondant à la position du groupe dans l'énoncé décideront de la structure à choisir. Si le groupe constitue tout l'énoncé, on aura respectivement:

1. ((un professeur)(de football))(américain)



2. (un professeur)((de football)(américain))



Si le groupe occupe une autre position dans l'énoncé:

1. (((un professeur)(de football))(américain))(est venu)

2. ((un professeur)((de football)(américain)))(est venu)

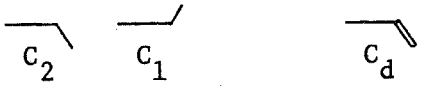
En fait, et puisqu'il n'y a ambiguité qu'entre deux structures, un seul contour sera dominant: le contour porté par B dans le premier cas, et le contour porté par A dans le second.

Un autre exemple est fourni par la séquence (A)(B)(...), dans laquelle la marque de sélection à gauche de B peut être neutralisée par un contour mélodique approprié (qui du reste se traduit par la ponctuation dans la graphie). Les deux structures correspondantes sont:
((A)(B)(...) et (A)(B)(...).

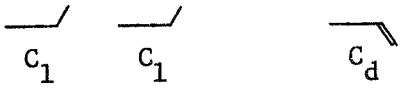
Ainsi dans (le chat)(malade)(a disparu) et en l'absence de circonstances qui lèveraient l'ambiguité, il peut s'agir d'un chat malade parmi tous les autres chats (cas 1: détermination), ou du seul chat dont il peut être question, et qui est malade (cas 2: qualification).

Cette fois, les séquences de contours appropriées sont:

1. ((le chat)(malade))(a disparu)



2. (le chat)(malade)(a disparu)



L'ambiguité est donc levée par le premier contour.

Des cas plus rares, et quelque peu artificiels, du type " la belle ferme le voile " ou " l'obstination de ces hommes braves la tourmente " dans lesquels les marques de l'unité verbale entrent également en syncrétisme et peuvent être confondues avec des marques nominales, sont également envisagables.

Un algorithme de désambiguation de la structure syntaxique, qui n'a pas accès aux indication contenues dans les circonstances, opèrera donc en 1) déterminant selon les règles de la grammaire les contours mélodiques correspondant aux syntagmes concernés par l'ambiguité structurale; 2) comparant les traits de ces contours avec ceux des contours observés dans la courbe de fréquence fondamentale et décider ainsi de la structure syntaxique à retenir. Si le choix doit être fait entre deux structures, un seul contour est pertinent pour lever l'ambiguité. Ce contour devrait alors être choisi de manière à ce que l'opposition paradigmatique impliquée porte sur une différence de pente.

8. En s'appuyant sur une théorie nouvelle permettant de dériver une séquence de contours mélodiques à partir de la structure superficielle de l'énoncé, on a présenté des algorithmes de reconnaissance automatique de la structure syntaxique opérant sur la courbe de fréquence fondamentale du signal de parole. L'utilisation d'autres paramètres phonétiques (intensité, durée des pauses, etc.) complémentaires de la

mélodie dans la manifestation des marques de l'expression, et l'incorporation de ces données dans un algorithme général exploitant les marques syntaxiques du contenu doit permettre l'élaboration de processus de reconnaissance de la structure syntaxique à la fois simples et sûrs.

Références bibliographiques

Gleason, H.A. (1967) An Introduction to Descriptive Linguistics, New-York, Holt, Rinehart and Winston.

Léon, P.R. et Martin, Ph. (1969) Prolégomènes à l'étude des structures intonatives, Montréal, Didier.

Lea, W.N. (1974) Prosodic Aids to Speech Recognition IV, Technical Report PX 10791, Univac Corp., St. Paul.

Martin, Ph. (1974) Phonologie de l'intonation de la phrase, à paraître dans les Actes du XIVème Congrès International de Linguistique et Philologie Romanes.

Martin, Ph. (1975a) Analyse phonologique de la phrase française, Linguistics, (sous presse).

Martin, Ph. (1975b) Théorie pour l'Intonation de la Phrase, à paraître.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

Effet de l'intonation sur la mémoire immédiate de messages verbaux.

Roger Deschamps.

RESUME

Des petits discours ont été interrompus à des endroits variables pour tester le rappel mot à mot des séquences qui précèdent les arrêts. Celles-ci contiennent une séquence de mots pouvant être placées dans deux environnements syntaxiques et ont été enregistrées avec trois intonations différentes.

La courbe de rappel ne semble pas dépendre des liens syntaxiques entre les propositions. Par contre, la proposition apparaîtrait comme une unité de représentation syntaxique au niveau de la mémoire immédiate, indépendante de l'intonation.

SUMMARY

Spoken connected discourses were interrupted for testing immediate recall where speech just presented contained an identical sequence of words in one of two syntactic configurations. These discourses had been recorded with three types of intonation.

Verbatim measures of recall do not support the immediately heard sentence as a retrievable unit in memory. Only the clauses seem to be organized as speech processing structures in memory not linked with the intonation.

Notre propos est d'examiner si les propriétés acoustiques d'un discours influence la manière dont des adultes le structurent en suites d'unités de représentation.

Nos arguments expérimentaux proviennent d'une situation étudiée en 1971 par Jarvella, où on propose à l'écoute de sujets, un texte interrompu à différents endroits et où la tâche expérimentale consiste à reproduire le plus fidèlement possible, les séquences qui précèdent les arrêts. Il est implicitement admis que la forme de la courbe de rappel témoigne des mécanismes de décodage du discours. Celle-ci, en ce qui concerne les vingt derniers mots, semble être liée à des facteurs syntaxiques: on observe une plus grande homogénéité des taux de mémorisation pour les mots qui appartiennent à une même phrase et à l'intérieur des phrases, les taux de mémorisation sont plus proches lorsque les mots sont inclus dans une même proposition. Un tel résultat, en supposant qu'il ne soit pas lié à des facteurs parasites, suggère une hypothèse intéressante sur la façon dont des sujets traitent l'information contenue dans un discours (Jarvella 1971).

La perception du discours se ferait selon un mécanisme de décodage proposition par proposition inclu dans un mécanisme de décodage phrase par phrase.

Toute proposition pourrait, en tant qu'unité syntaxique, être reconnue en temps réel et c'est représentée mot à mot qu'elle serait par la suite placée dans une mémoire rapide. La durée de la mise en réserve de la proposition dans la mémoire rapide, dépendrait de sa position dans la phrase. Deux cas peuvent se présenter: ou bien la proposition n'appelle aucune suite, ou bien son interprétation dépend de constituants ultérieurs. Dans le premier cas, la représentation mot à mot de la proposition se dégrade tout de suite après avoir été recodée en une forme plus abstraite qui conserve de la forme originale, les relations sémantiques entre les constituants. Dans le second cas, là où la proposition ne termine pas une phrase, elle reste maintenue dans la mémoire rapide sous sa forme superficielle.

L'hypothèse dont il vient d'être question, implique pour le sujet parlant, la capacité de reconnaître en temps réel, les propositions d'une part et d'autre part, les phrases. La reconnaissance de ces structures se ferait en outre sur la base d'indices essentiellement syntaxiques.

L'unité fonctionnelle que constitue la proposition, a été reconnue à la suite de divers travaux expérimentaux (Fodor & Bever 1965, Bever, Lackner & Stoltz 1969, Holmes & Forster 1970, Chapin, Smith & Abrahamson 1972, Seitz & Weber 1974). Certains des effets seraient par ailleurs indépendants des propriétés acoustiques (intonation et pauses) du signal (Garrett & Fodor 1968).

L'importance théorique de la phrase en tant qu'unité de représentation munie d'un ensemble de propriétés structurales, est apparue quant à elle, dans des études psychologiques de la grammaire dont le domaine ne s'étend pas au-delà de la manipulation de phrases présentées isolément

(Miller & Chomsky 1963, Miller 1969, Bever 1970, 1972).

La question qui se pose est de savoir si les mécanismes impliqués dans la perception de phrases isolées et les mécanismes de décodage du discours sont identiques. Une réponse affirmative supposerait au moins que la phrase constitue indépendamment de ses propriétés acoustiques, une unité de représentation facilement isolable.

Le matériel expérimental dans l'expérience de Jarvella, avait été enregistré avec une intonation visiblement très monotone: l'expérimentateur avait veillé à ce que le texte soit lu à un rythme à peu près constant de quatre syllabes par seconde. Néanmoins, les syllabes ayant été produites dans leur ordre normal, il est peu probable que toute intonation de phrase ait été complètement effacée. Il se pourrait alors que la forme de la courbe de rappel ne soit liée à la grammaire que par l'intermédiaire de l'intonation, la structure intonative du discours et sa structure grammaticale étant en effet corrélées (Lieberman 1966).

Un effet notable de l'intonation sur la segmentation du discours en propositions courtes, semble à première vue peu probable. Dans certaines conditions expérimentales, des propositions courtes peuvent en effet être isolées indépendamment de leurs propriétés acoustiques (Garrett & Fodor 1968). Le seul effet que nous attendons concerne le découpage du discours en unités formées d'une suite de propositions.

Dans la présente recherche, les vingt mots qui précèdent les interruptions du texte, composent trois propositions réparties dans deux phrases. Deux cas ont été prévus: la première phrase est soit formée par les deux premières propositions, soit par la première uniquement.

Les textes ont été enregistrés de trois manières différentes: une fois en respectant la correspondance entre la syntaxe et l'intonation et une fois en suggérant pour les items expérimentaux, des regroupements en propositions non conformes à la syntaxe; le troisième enregistrement a été construit en collant bout à bout des syllabes dites dans un ordre quelconque.

Deux prédictions sont à envisager: ou bien c'est la grammaire qui détermine la courbe de rappel ou bien c'est l'intonation. Si les effets sont liés à la grammaire, il faut que les intéractions entre la position du mot et la dépendance syntaxique, soient du même type quelles que soient les propriétés acoustiques de la séquence. Par contre, si le facteur critique est l'intonation, ces intéractions doivent varier en fonction de l'enregistrement et s'annuler au total.

Matériel expérimental .

On considère deux fois huit paires d'items comprenant trois propositions réparties dans deux phrases. Les membres de chaque paire se différencient par les liens de dépendance syntaxique entre les propositions:

la première phrase du premier membre est formée des deux premières propositions (item 2-1); dans le second membre, la première phrase coïncide avec la première proposition (item 1-2). La première et la troisième proposition comprennent sept mots tandis que la deuxième, six.

Dans huit paires, l'item 1-2 est construit en permutant la première et la troisième proposition de l'item 2-1. La deuxième proposition reste par conséquent identique dans ces paires d'items.

-exemple

item 2-1 . Les individus les plus résistants titubaient encore en voulant quitter la zone dangereuse. Le quartier était entouré de fil électrifié.

item 1-2 . Le quartier était entouré de fil électrifié. En voulant quitter la zone dangereuse, les individus les plus résistants titubaient encore.

Dans les huit paires restantes, on veille à ce que le lien de dépendance entre la subordonnée et la principale soit plus marqué. La présence d'une subordonnée était facultative dans les huit premiers items, ici, elle est obligatoire. Cependant, on s'est arrangé pour que la combinaison de la deuxième et de la troisième proposition de tout item 2-1 forme une phrase. Les items 1-2 sont alors construits en remplaçant la principale de chaque item 2-1 par une proposition indépendante et en regroupant les deux dernières propositions en une seule phrase. Dans ce cas, les deux dernières propositions restent identiques dans chaque paire.

-exemple

item 2-1 . On a déjà évoqué dans trois articles, le problème des petites exploitations agricoles. Reste toujours le problème des minorités nationales.

item 1-2 . On y a déjà consacré trois articles. Le problème des petites exploitations agricoles reste toujours le problème des minorités nationales.

Chaque item est inclu dans un texte d'environ cent cinquante mots. Huit des items sont placés dans la première moitié de leur texte, tandis que les huit restants, dans la seconde moitié.

Les textes sont enregistrés de trois manières différentes: une fois en respectant la correspondance entre la syntaxe et l'intonation (IC), une fois en donnant à l'item expérimental, l'intonation de l'autre item de la paire (IF) et une troisième fois en collant bout à bout des syllabes produites dans un ordre quelconque (SI). Seule subsiste dans ce dernier enregistrement, une intonation d'énumération de syllabes; leur rythme est de deux syllabes par seconde.

Huit combinaisons des textes ont été prévues, compte tenu de l'intonation et des propriétés syntaxiques des items expérimentaux. Chaque combinaison comprend huit textes avec une intonation SI, quatre avec une intonation IC et quatre avec une intonation IF. Le nombre d'items 1-2 et d'items 2-1 est identique dans chaque condition d'intonation et les combinaisons sont établies de telle sorte qu'au total, les items apparaissent le même nombre de fois dans chaque condition expérimentale: une fois en IC et en IF et deux fois en SI.

Sujets .

Il s'agit d'étudiants en sciences économiques dont l'âge est compris entre 18 et 23 ans et dont la langue maternelle est le français. Il y avait 80 sujets répartis au hasard dans huit groupes correspondant aux huit versions du matériel. Chaque goupe comprend dix sujets.

Procédure .

L'expérience s'est déroulée au laboratoire de langue à un moment où les étudiants auraient dû normalement faire des exercices d'anglais ou de néerlandais. La tâche consiste à écouter attentivement les textes. Lorsque le sujet entend un signal, c'est-à-dire, lorsque l'expérimentateur arrête le magnétophone, il est invité à reproduire le plus fidèlement possible la séquence qu'il vient d'entendre en la retranscrivant dans un livret destiné à cet effet. On insiste dans les instructions sur la nécessité de bien reproduire mot à mot la portion du message qui vient d'être entendue. On demande aussi aux sujets de veiller à ce que les mots qui précèdent le plus directement les arrêts soient bien retranscrits.

Le test commence par une série d'essais au cours desquels l'expérimentateur vérifie si les instructions ont bien été comprises.

Présentation des résultats .

La figure 1 représente les taux de mémorisation des mots en fonction de leur position dans les phrases.

On remarque , là où l'intonation et la grammaire sont compatibles (IC), une meilleure mémorisation de la deuxième proposition si elle appartient à la deuxième phrase (IC 1-2) et des maxima au début des trois propositions dans les deux conditions.

Dans la condition IF, la courbe de mémorisation des phrases 1-2 est plus proche de la courbe relative aux phrases 2-1 dites correctement que de la courbe se rapportant aux phrases IC 1-2. De même, les phrases

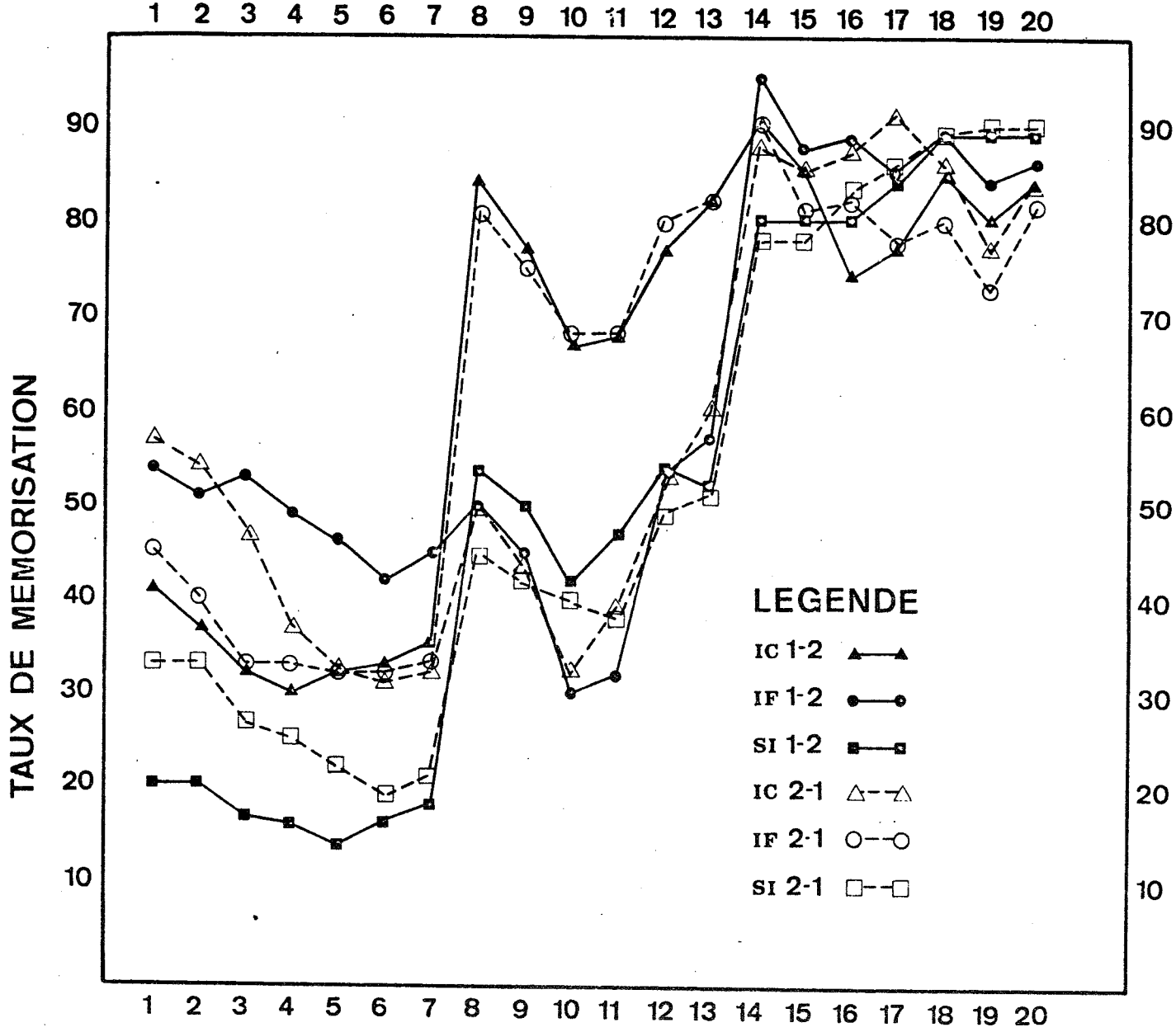


FIG.I. TAUX DE MEMORISATION DES MOTS EN FONCTION DE LEUR POSITION DANS LES PHRASES.

IF 2-1 présentent une courbe de mémorisation qui rejoint davantage celle des phrases IC 1-2 que des phrases IC 2-1. Notons que les courbes IC 1-2 et IF 2-1 sont indissociables. Par contre, ce n'est qu'à partir du premier mot de la deuxième proposition que les courbes IC 2-1 et IF 1-2 tendent à se confondre, la première proposition semblant être légèrement mieux retenue lorque l'intonation et la syntaxe sont compatibles.

Les deux courbes se rapportant aux conditions SI sont quant à elles, assez bien mélangées. Remarquons toutefois les scores apparemment meilleurs pour le cas de la première proposition dans la condition SI 1-2. Enfin, notons qu'à partir du premier mot de la deuxième proposition (le huitième dans la séquence), les courbes relatives aux deux conditions SI, aux conditions IC 2-1 et IF 1-2 se superposent.

Après avoir remplacé les valeurs des proportions par la fonction 2 arc Sin$\sqrt{x}$ pour rendre les variances des observations plus ou moins identiques, nous avons soumis nos données à une analyse de la variance. Trois facteurs ont été retenus:

- la position du mot
- la dépendance syntaxique entre les propositions (1-2 et 2-1)
- l'intonation de l'item (IC, IF et SI).

Reprenons chacun de ces points.

La position du mot est hautement significative (F.01 = 1.91, Fcalc. = 117.4).

Le type de dépendance syntaxique ne semble pas avoir d'effet (F.05 = 3.84, Fcalc. = .338).

L'intonation est hautement significative (F.01 = 4.61, Fcalc. = 56.6). C'est la condition SI qui est responsable de cet effet. Les scores y sont en moyenne moins bons.

Les intéractions entre le type de dépendance syntaxique et la position du mot ne sont pas statistiquement différentes de zéro (F.05 = 1.59, Fcalc. = .82); par contre, les intéractions entre l'intonation et la position du mot sont significatives à P = .01 (F.01 = 1.60, Fcalc. = 2.49). Sont aussi hautement significatives, les intéractions de deuxième ordre entre la position du mot, le type de dépendance syntaxique et l'intonation (F.01 = 1.60, Fcalc. = 6.69).

Dans les courbes SI, les discontinuités apparaissant à la huitième et à la quatorzième position, sont significatives à P = .01.

Discussion .

Des résultats antérieurs avaient suggéré que la mémoire de séquences grammaticales dépend de leur structure de surface. En particulier, il avait été montré que les taux de mémorisation des mots varient moins s'ils appartiennent à une même phrase (Jarvella 1971). Ce fait ne se vérifie ici que si l'intonation du matériel est compatible avec la syntaxe.

Les intéractions de premier ordre entre la position du mot et la dépendance syntaxique ne sont pas significatives alors que les intéractions de deuxième ordre entre la position du mot, la dépendance syntaxique et l'intonation, le sont. En d'autres termes, s'il est vrai, lorsque la syntaxe et l'intonation sont compatibles, que les courbes de rappel relatives aux deux types de dépendance syntaxique ne sont pas parallèles et sont liées au découpage du discours en phrases grammaticales, cet effet s'annule dès que l'on donne aux textes, une intonation d'énumération de syllabes et s'inverse lorsque les items sont lus avec une intonation qui suggère une inversion au niveau des regroupements des propositions. Cette dernière manipulation a en moyenne le même effet qu'une manipulation similaire de la syntaxe.

L'identité des deux courbes de rappel dans la condition SI et les effets inverses obtenus en IC et en IF, suggèrent que ce ne sont pas les groupes syntaxiques délimités par la phrase grammaticale qui constituent des unités directement accessibles à la mémoire rapide, mais les séquences de propositions munies d'une intonation de phrase.

On avait noté dans les courbes SI, deux discontinuités correspondant chacune à une frontière entre propositions.Ainsi, même dans la condition où il ne subsiste plus qu'une intonation d'énumération de syllabes, le repérage des propositions resterait possible.

En conclusion, la seule unité syntaxique de représentation qui apparaît dans nos résultats, est la proposition. La segmentation en unités plus vastes semble quant à elle, liée à des facteurs acoustiques.

En ce qui concerne la phrase grammaticale, les résultats expérimentaux discutés jusqu'ici, ne suggèrent pas qu'elle constitue une unité facilement isolable dans un discours. Ceci n'implique toutefois pas que la phrase vue comme unité fonctionnelle de représentation, ne joue aucun rôle au niveau de la structuration de portions de discours dans la mémoire rapide.

Reprenons le mécanisme de décodage mentionné plus haut et, raisonnant de la même manière que Jarvella (1971), admettons que la perception du discours se fasse selon un mécanisme de décodage proposition par proposition inclu dans un mécanisme de décodage séquence munie d'une intonation de phrase par séquence munie d'une intonation de phrase. Supposons en outre, que les séquences munies d'une intonation de phrase soient structurées au niveau de la mémoire rapide par rapport à un modèle de phrase correcte du

point de vue grammatical. Si cette hypothèse est exacte, il est bien évident que là où la séquence ne correspond pas à une phrase grammaticale, sa structuration au niveau de la mémoire rapide, doit être plus difficile. On devrait par conséquent s'attendre à ce que les taux moyens de mémorisation dans la condition IF, soient moins élevés que dans la condition IC. Or, ceux-ci ne sont pas statistiquement différents.

Une autre hypothèse qui fait intervenir la phrase comme unité fonctionnelle de représentation, consiste à admettre que les échanges au niveau de la mémoire rapide des séquences munies d'une intonation de phrase, sont gérés partiellement en fonction d'indices syntaxiques: la probabilité qu'une séquence de propositions reste maintenue dans la mémoire rapide serait par exemple d'autant plus grande que la structure syntaxique de la séquence s'éloigne d'une structure de phrase. Dans la condition IF 1-2, la combinaison des deux premières propositions - combinaison qui est lue avec une intonation de phrase - ne forme jamais une phrase grammaticale. Elle devrait par conséquent rester plus longtemps dans la mémoire rapide que la première phrase des items IC 2-1 et être de ce fait mieux rappelée. Un examen des courbes de rappel semble suggérer qu'un léger effet de ce type existe, effet qui se concentre essentiellement sur la première proposition. Il n'est toutefois pas statistiquement significatif.

De ce qui vient d'être dit, il résulte que les seuls indices syntaxiques dont semblent témoigner nos résultats, concernent le découpage et la représentation du discours en propositions comportant six ou sept mots. Les analyses syntaxiques au niveau de la mémoire rapide, seraient par conséquent essentiellement locales.

Quant à l'intonation, son rôle se limiterait à optimiser les échanges dans les unités d'entrée et de sortie: grâce à la structure rythmique qu'elle confère aux énoncés, elle augmenterait la capacité des unités périphériques à appréhender de plus grandes portions du message en une fois. Les séquences définies par l'intonation seraient ensuite prises en charge par le mécanisme d'analyses syntaxiques locales dont nous avons parlé plus haut.

La question qui se pose à présent est de savoir comment l'écouteur intègre les courtes unités syntaxiques en une représentation plus globale. Ce problème dépasse le cadre de notre travail. Des suggestions intéressantes peuvent être trouvées dans un ensemble d'études publiées récemment par Anderson & Bower (1973).

Bibliographie .

Anderson, J.R. & Bower G. H. Human associative memory. Washington: Winston, 1973.

Bever, T. G. The cognitive basis for linguitic structures. in J. R. Hayes (ed.), Cognition and the development of language. New York: Wiley, 1970.

Bever, T. G. Perceptions tought and language. in R. O. Freedle & J. B. Carroll (Eds.), Language comprehension and the acquisition of knowledge. Washington: Winston, 1972.

Bever, T. G., Lackner, J. & Stolz, W. Transitional probability is not a general mechanism for the segmentation of speech. Journal of Experimental Psychology, 1969, 70, 387-394.

Chapin, P. G., Smith, T. S. & Abrahamson, A. A. Two factors in percptual segmentation of speech. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 164-173.

Fodor, J. & Bever, T. G. The psychological reality of linguistic segments. Journal of Verbal Learning and Verbal Behavior, 1965, 4, 414-420.

Garrett, M. & Fodor, J. Psychological theories and linguistic contructs. in T. R. Dixon & D. L. Horton (Eds.), Verbal behavior and general behavior theory. Englewood: Prentice-Hall, 1968.

Holmes, V. M. & Forster K. I. Detection of extraneous signals during sentence recognition. Perception & Psychopysics, 1970, 7, 297-301.

Jarvella,R. J. Syntactic processing of connected speech. Journal of Verbal Learning & Verbal Behavior, 1971, 10, 409-416.

Lieberman, P. Intonation, perception and language. Cambridge: MIT Press, 1966.

Miller, G. A. Quelques études psychologiques de la grammaire, Langages, 1969, 16, 61-82.

Miller, G. A. & Chomsky, N. Finitery models of language users. in R. D. Luce, R. R. Bush & E. Galanter (Eds.), Handbook of mathematical psychology, Vol II, New York: Wiley, 1963

Seitz, M. R. & Weber, B. A. Effects of response requirements on the location of clicks surimposed on sentences. Memory & Cognition, 1974, 2, 43-46.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UNE APPROCHE SYNTAXIQUE DE RECONNAISSANCE DE PHRASES DANS UN CONTEXTE DONNE

Jean-Marie PIERREL et Jean-Paul HATON

Université de NANCY 1

RESUME

Cet article décrit une approche analytique de la reconnaissance du discours continu. Le système en cours de réalisation comprend : un analyseur spectral (câblé), un étage de segmentation-identification de phonèmes, un analyseur syntaxique et un vérificateur acoustique de mots. On donne les résultats obtenus dans le cadre d'une application concernant l'automatisation d'un standard téléphonique.

SUMMARY

This paper describes an analytical approach toward the recognition of connected sentences. The system which is being implemented is made up of a hardware spectral analyzer, a phoneme segmentation and identification program, a syntactic parser, and an acoustic word-verification program. This system has been tested on a specific application and some results are given.

UNE APPROCHE SYNTAXIQUE DE LA RECONNAISSANCE DE PHRASES DANS UN CONTEXTE DONNE

Jean-Marie PIERREL - Jean-Paul HATON

Université de NANCY 1

1 - *INTRODUCTION*.

Jusqu'à présent, la plupart des recherches en reconnaissance automatique de la parole concernaient l'identification de mots isolés extraits d'un vocabulaire donné. Des résultats intéressants ont été enregistrés dans ce domaine, même s'il reste encore beaucoup à faire pour la généralisation des systèmes existants. Quoiqu'il en soit, la reconnaissance de mots prononcés isolément ne peut être utilisée que dans un nombre limité d'applications pratiques. Dans le cadre du dialogue homme-machine, il est nécessaire de pouvoir communiquer par phrases, prononcées continûment, avec la machine, et plus seulement par mots isolés. C'est ainsi qu'une grande partie des recherches actuelles est maintenant dirigée vers la reconnaissance du discours continu, en particulier depuis le lancement d'un important projet américain de l'ARPA sur le sujet [1] .

Depuis Vicens [2] , les systèmes de reconnaissance utilisant des contraintes syntaxiques ont été peu nombreux et toujours très limités : entrée orale de programmes [3] [4] , commande en temps réel de machines [5] , [6] . De plus, ces systèmes n'admettaient le plus souvent que des phrases à mots isolés. Cependant, ces premières expériences ont mis en évidence l'amélioration que les contraintes de niveaux hiérarchiques supérieurs (morphologiques, syntaxiques, sémantiques) pouvaient apporter aux performances d'un système acoustique de reconnaissance. Il reste encore de nombreux points à éclaircir dans l'application de ces contraintes aux diverses étapes du traitement. Sans arriver aux très grosses réalisations, de type intelligence artificielle, du projet ARPA, il est intéressant d'étudier le fonctionnement d'un analyseur syntaxique appliqué aux données d'un système de reconnaissance et son insertion dans un tel système. Nous présentons dans cet article une des approches que nous utilisons pour tenter de résoudre ce problème. Il s'agit d'une approche analytique dans laquelle on utilise un étage acoustique de segmentation-identification, transformant une phrase en une chaîne phonémique à

réponses multiples. Le fonctionnement du niveau acoustique et les conditions expérimentales sont également précisées ici.

Pour mettre en œuvre l'analyseur syntaxique, nous avons choisi un premier exemple d'application, l'automatisation d'un standard téléphonique. Il ne s'agit là que d'une application parmi biend'autres possibles et qui est surtout destinée à tester de façon commode nos hypothèses de départ. Dans cette application, nous envisageons des phrases constituant un sous-ensemble restreint du français parlé, car il est de toute façon impossible pour l'instant de concevoir un système automatique capable de traiter une langue naturelle dans son ensemble.

Après avoir expliqué le principe de l'analyse syntaxique, on donne des exemples de résultats obtenus en temps différé à partir des transcriptions phonémiques de phrases.

3 - LE SYSTEME ACOUSTIQUE.

Le système de reconnaissance acoustique est construit autour d'un minicalculateur NOVA 2 de DATA GENERAL, avec 16K mots de mémoire centrale. Il comprend :

- un étage (hardware) d'analyse du signal vocal, avec un analyseur spectral à 32 canaux, couvrant la bande de fréquences 100Hz - 7000Hz et un détecteur de mélodie (Mélographe CNET/ETA). Les données spectrales numérisées sont acquises en temps réel par le calculateur chaque 10ms (fréquence d'échantillonnage de 100Hz) ;

- un programme de segmentation, permettant de scinder le mot ou la phrase prononcée en unités élémentaires assimilables en général aux phonèmes. Ce programme tient compte principalement des zones transitoires (détectées en étudiant les variations d'énergie dans les canaux de l'analyseur spectral) et des indications de voisement fournies par le détecteur de mélodie. La segmentation obtenue est toujours plus ou moins entachée d'erreurs correspondant à des omissions ou, au contraire, à des insertions de phonèmes (segmentations multiples) ;

- un programme d'identification des segments fournis par l'étage précédent. Aux erreurs de segmentation viennent s'ajouter des erreurs d'identification des phonèmes, de telle sorte que la chaîne de phonèmes fournie par le système acoustique est fortement erronée. Pour diminuer ces erreurs, on prend en compte, pour chaque segment, plusieurs réponses possibles, affectées chacune d'un "score" acoustique. L'analyseur syntaxique accepte ainsi en entrée deux types de données :

  - des chaînes de phonèmes à réponses multiples, entachées d'erreurs,
  - les données spectrales fournies par l'analyse acoustique.

Actuellement, les chaînes de phonèmes sont traitées en temps différé par l'analyseur syntaxique, car le système complet n'a pas encore été implémenté sur NOVA.

3 - *PRINCIPE DU SYSTEME D'ANALYSE SYNTAXIQUE* [7] .

Supposons que les phrases prononcées fassent partie d'un langage L engendré par une grammaire G. Dans ce cas, un algorithme d'analyse syntaxique permet de déterminer si la phrase $\alpha_1 \alpha_2 \ldots \alpha_n$ fait partie du langage L(G) et le système acoustique intervient alors à un étage de vérification ou de décision en cas d'ambiguïté.

3.1. *Définition d'une grammaire (grammaire Context free)*

Une grammaire G, associée à un tel langage, est formée d'un quadruplet (T,N,:: =, X) où :

- T et N sont des ensembles finis, disjoints, appelés respectivement alphabet terminal et alphabet non terminal (leurs éléments s'appellent symboles terminaux et symboles non terminaux) .

V = T U N est l'alphabet de G.

- X est un élément de N appelé Axiome de G.

- :: = est une relation binaire appelée relation de production entre T et V^* (V^* = ensemble de mots sur V). Un couple (A,α) tel que $A \in T$, $\alpha \in V^*$ et A::= α s'appelle une règle de la grammaire.

3.2. *Langage engendré par une grammaire*

A la grammaire G(T,N,::=,X) on associe la relation $\succ$ dans V^* définie par :

$$\alpha \succ \beta \Leftrightarrow (\exists A \in N, \lambda \in V^*, \lambda' \in V^*, \mu \in V^*)$$

$$(\alpha = \lambda A \lambda' \qquad \beta = \lambda \mu \lambda' \text{ et } A::=\mu).$$

Lorsqu'on a $\alpha \succ \beta$ on dit que α se réécrit β. La fermeture transitive de la relation $\succ$ s'écrit $\succ^*$. Si $\alpha \succ^* \beta$ on dit que de α dérive β .

On montre que le langage L(G) engendré par une grammaire G(N,T,::=,X) est l'ensemble des mots de T^* qui dérivent de l'Axiome X.

3.3. *Analyse syntaxique généralisée. Définitions*

- Définition de l'analyse syntaxique.

Etant donné un mot $\alpha \in T^*$, trouver un chemin de dérivation de X à α s'appelle faire l'analyse syntaxique de α pour la grammaire G.

Si $\alpha \notin$ L(G), il n'existe pas de tel chemin.
Si $\alpha \in$ L(G), il existe au moins un tel chemin.

- Définition de la relation initiale (J^*) dans V.

Λ désigne le mot vide.
Soit la relation J dans V définie par

$$A\ J\ B \Leftrightarrow (\exists\ \emptyset \in V^*\ ,\ \exists\ \psi \in V^*)\ (A ::= \emptyset\ B\ \psi \text{ et } \emptyset \overset{*}{\succ} \Lambda).$$

Notons J^* la fermeture transitive de J.
Si $A\ J^*\ B$, on dit que B est initial de A.
Cette relation permettra, au cours de l'analyse, d'émettre des hypothèses sur les mots, qui seront ensuite vérifiées au niveau acoustique.

3.4. Algorithme d'analyse syntaxique

L'analyse syntaxique descendante (top down) consiste à appliquer un algorithme qui, partant de l'Axiome X, construit de proche en proche la suite de dérivation conduisant à la donnée α .

Cet algorithme comprend essentiellement une procedure ANALYSE (A) qui, appliquée aux éléments $A \in V$, construit la suite de dérivation de A à une partie β de la donnée α :

si $A \in T$ il s'agit de vérifier que β commence par A ;
si $A \in N$ la procédure doit choisir la règle, parmi les règles de premier membre A, qui conduit par dérivation à β ou à une partie initiale de β (soit K le numéro de cette règle $A_K = B_{K1} B_{K2} \cdots B_{Kq_K}$ tq $A_K = A$), et appelle ensuite les procédures "analyse" associées aux symboles du membre droit de la règle.
C'est une procédure indéterministe et récursive que l'on traitera par un automate à pile.

On peut décrire ainsi ANALYSE (A) :

Pour $A \in T$

début si γ (1ère partie de β) = A alors sortir (A)
sinon erreur fin

Pour $A \in N$

début CHOIX (d'un numéro de règle K tq A_K = A) ;
analyse (B_{K1}) ; analyse (B_{K2}) ;
.....
analyse (B_{Kq_K}) ;
fin.

3.5. *Etage de vérification acoustique : procedure CHOIX*

a) Elaboration des hypothèses :

Pour chaque règle K telle que A_K = A, on recherche les initiaux terminaux de A_K que l'on conserve associés au rang K de la règle.

On obtient ainsi un ensemble d'hypothèses, chaque hypothèse étant formée d'un couplet (A,K) c'est-à-dire (mot terminal du langage, numéro de règle associé).

b) Vérification des hypothèses :

Un algorithme de reconnaissance acoustique de mots infirme ou confirme chacune des hypothèses et classe les différents résultats suivant un ordre décroissant de probabilité de reconnaissance (certaines hypothèses sont purement et simplement rejetées).

c) Poursuite du processus :

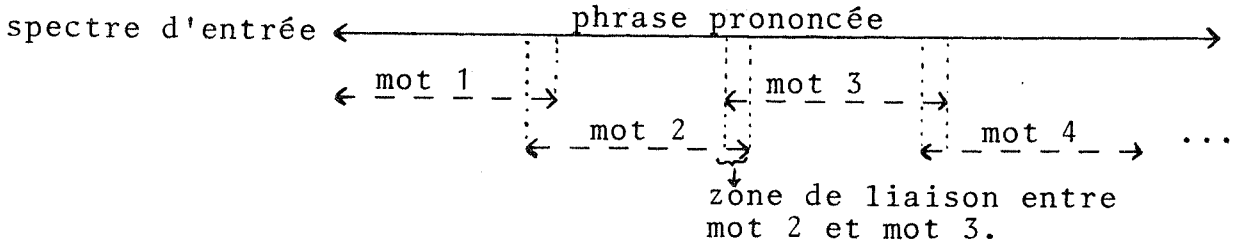
Les résultats (n° de terminal, n° de règle, probabilité, pointeur dans la liste de données d'entrée) de l'étape précédente sont rangés sur une pile et on réitère le processus à partir des données fournies par le sommet de pile.

3.6. *Algorithme de reconnaissance acoustique des terminaux*

Les terminaux que nous utilisons sont en général de type "mot" et nous nous servons de deux algorithmes de reconnaissance.

a) Reconnaissance globale :

Cette méthode permet de comparer le spectre obtenu dans la chaîne d'entrée à des spectres théoriques (ou de référence) correspondant aux terminaux recherchés. Cette comparaison se fait par une méthode de comparaison dynamique. Il est nécessaire de faire des petits recouvrements des zones de la chaîne d'entrée pour rendre compte des jonctions de mots.



b) Reconnaissance analytique :

Cette méthode utilise la transcription phonétique fournie par le système acoustique. Nous avons donc une chaîne de phonèmes d'entrée et un dictionnaire donnant pour chaque terminal la chaîne de phonèmes de référence. Pour chaque phonème, nous prendrons en compte les 3 réponses les plus probables (ex. pour t : p,t,K, cf. § II). La méthode de comparaison se rapproche de la recherche d'un chemin optimal dans un graphe et tient compte des possibilités d'omission, d'insertion ou de substitution de phonèmes [8] .

A titre d'exemple, la Fig. 1 schématise la comparaison du terminal "Je voudrais" avec une sous-chaîne d'entrée contenant ce mot.

3.7. Schéma général du système

La Fig. 2 donne un schéma simplifié de fonctionnement du système complet.

4 - APPLICATION A UN LANGAGE DONNE.

Le système que nous avons élaboré est un système général dont les données sont de 3 types :

- la grammaire du langage,
- le dictionnaire de référence des terminaux (spectre + phonèmes),
- la liste d'entrée représentant la phrase (spectre + phonèmes).

Nous l'avons testé sur une application particulière : un langage d'appel téléphonique en entrée d'un standard. Ce langage comporte 7104 x n phrases de 2 à 13 mots (n étant le nombre de personnes ou de postes que l'on peut appeler).

Actuellement, pour les essais, nous avons 6 personnes et 6 postes différents. Le système peut donc reconnaître 42624 phrases différentes de 2 à 13 mots. Le vocabulaire utilisé se compose de 30 mots (18 terminaux vrais du langage et 12 mots constituant "l'annuaire" du standard). La grammaire utilisée est donnée Fig. 3.

Pour l'instant, les tests ont été fait en temps différé sur le CII 10070 de l'I.U.C.A. de NANCY et avec un corpus de phrase dont nous donnons quelques exemples. Ces phrases sont correctement reconnues après un temps moyen de traitement inférieur à une demi-seconde par phrase. Il ne s'agit là que de temps approximatifs ; seule l'implémentation

Chaîne de référence : ʒ ə v ɯ d [ɾ] e
v a ʒ o b

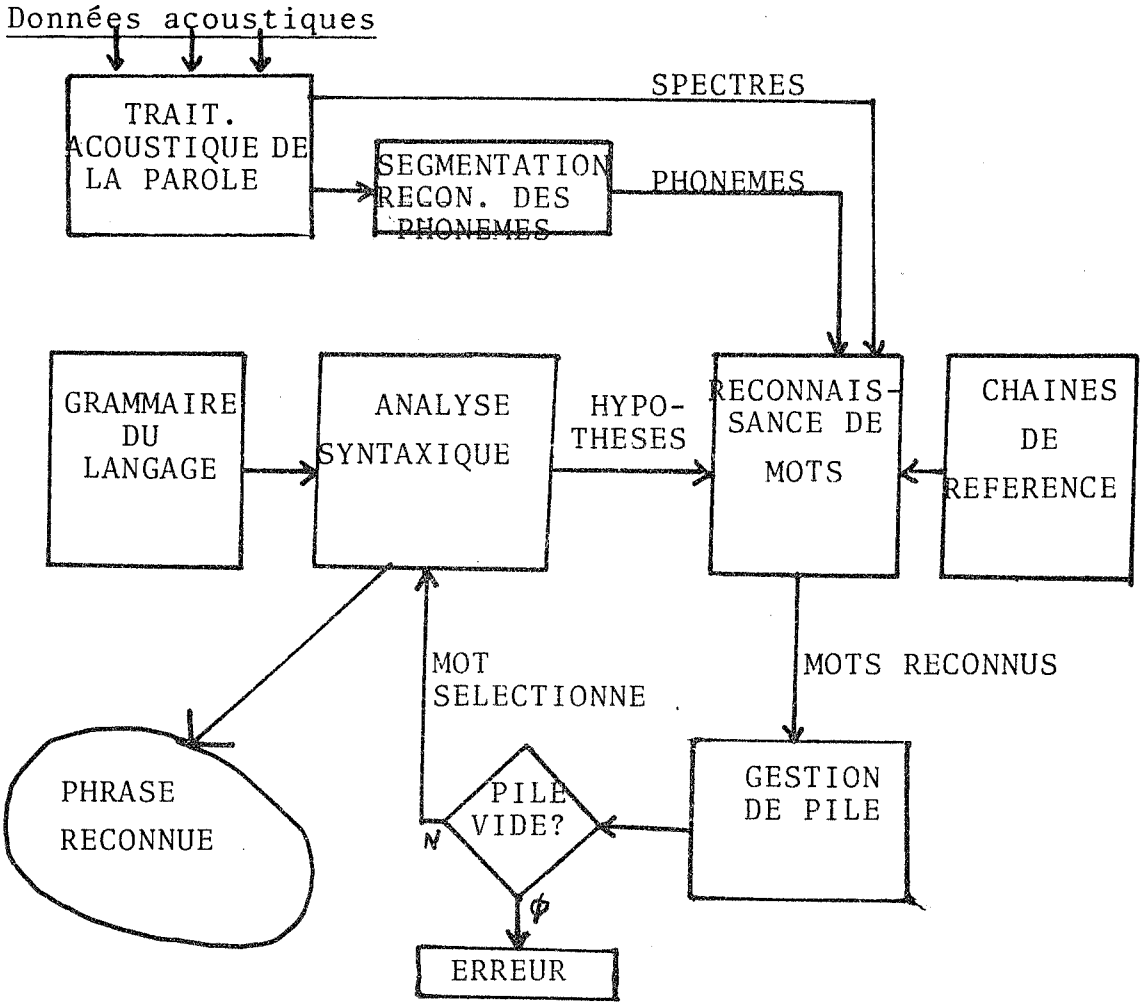
Données en entrée :

a l ɯ ʒ a ʒ ɣ ɯ b i ɛ p ə l e
e ɾ ɔ v ə v ɯ e d ɛ e t a ɾ ɛ
ə o ʒ e ʒ o g e i k e i

Pointeur de départ ↑ A

Pointeur de fin ↑ B

Figure 1



Système de reconnaissance

Figure 2.

Figure 3.

Grammaire du langage

< PHRASE > ::= <APPEL> <DEMANDE>
<DEMANDE>

< APPEL > ::= Allo <SALUT>
allo
< SALUT >

< SALUT > ::= Bonjour <TYPE>
Bonjour
< TYPE >

< DEMANDE > ::= <REQUETE> <POLI>
<REQUETE>
<POLI>

< POLI > ::= s'il vous plaît Merci
s'il vous plaît
Merci

< REQUETE > ::= <INT>
<AFFIR>

< INT > ::= Pourrais-je <DEM>
est-ce que je pourrais <DEM>

< AFFIR > ::= Je voudrais <DEM>
passez-moi <OBJET>
< OBJET >

< DEM > ::= Parler à <PERS>
avoir <OBJET>

< PERS > ::= <TYPE> <NOM>

< OBJET > ::= le <QUALIF> de <PERS>
< PERS >
le <NUM >
le <QUAL 1> <NUM>

< QUALIF > ::= Bureau
<QUAL 1>

< QUAL 1 < ::= Poste
numéro

< TYPE > ::= Monsieur / Madame / Mademoiselle

< NOM > ::= Albert / Dupont / Durand / Pierre / Pierrel /
Bouchet

< NUM > ::= 221 / 222 / 223 / 239 / 240 / 241

en temps réel du système complet permettra de chiffrer effectivement le temps de réponse.

Exemples de phrases testées :

"Allo je voudrais parler à Madame Durand"
"Je voudrais parler à Monsieur Pierre"
"Je voudrais parler à Monsieur Pierrel"
"Je voudrais le poste 239"
"Je voudrais avoir le poste 240"
"Allo bonjour Madame Je voudrais avoir le poste 223 s'il vous plaît, Merci"
"Je voudrais parler à Madame Albert, Merci"
"Je voudrais avoir le poste 223"
"Bonjour je voudrais parler à Madame Bouchet"
"Allo, Mademoiselle, Je voudrais avoir Monsieur Dupont".

5 - *CONCLUSION*.

Nous présentons dans cet article un système de reconnaissance du discours continu utilisant un analyseur syntaxique. Après avoir décrit le fonctionnement de cet analyseur et du système acoustique de reconnaissance, nous donnons quelques exemples de résultats obtenus dans le cas particulier d'une consultation automatique de standard téléphonique. En fait, il ne s'agit là que d'une application possible parmi d'autres, destinée à vérifier la validité de notre approche, le système que nous développons étant capable de traiter d'autres langages. Les performances obtenues sont très encourageantes et nous permettent d'envisager dès maintenant une extension des travaux : généralisation du système aux différents niveaux de traitement (extension de la grammaire, du vocabulaire terminal et du nombre de locuteurs), utilisation d'une analyse sémantique (en particulier en faisant intervenir les indices prosodiques), etc...

BIBLIOGRAPHIE.

[1] A. NEWELL et al. "Speech understanding systems : final report of a study group" C.M.U. Pittsburgh - 1970.

[2] P. VICENS "Aspects of speech recognition by computer" Ph D Thesis, Stanford Univ., 1969.

[3] J.-P. TUBACH "Reconnaissance automatique de la parole" Thèse d'Etat, Université de Grenoble, 1970.

[4] G.-Y. VYSOTSKIY et al. "An experiment in oral control of a computer" Eng. Cybern., n°2, pp. 320-327, 1970.

[5] J.-P. HATON "A practical application of a real-time isolated-word recognition system using syntactic constraints" IEEE Trans., ASSP, 22, n° 6, pp. 416-19, 1974.

[6] R.-B. NEELY, G.-M. WHITE "On the use of syntax in a low cost real-time speech recognition system" IFIP Congress. 74, Stockholm, Aug. 1974.

[7] C. PAIR "Compilation" Ecole d'été d'informatique AFCET, Neufchatel, Juillet 1972.

[8] C. TAPPERT, N. DIXON "Application of sequential decoding for converting phonetic to graphemic representation in automatic recognition of continuous speech" IEEE Trans. AU, 23, n° 3, pp. 225-228, 1973.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

"Compréhension automatique de la parole continue à l'aide de la phonologie, la syntaxe et la sémantique".

G. BATTANI et H. MELONI

Groupe d'Intelligence Artificielle

U. E. R. MARSEILLE-LUMINY

Avec la collaboration de G. MERCIER - C.N.E.T. LANNION

RESUME

Nous présentons un programme de compréhension de la parole continue qui utilise les contraintes phonologiques, syntaxiques et sémantiques.

Notre "entrée" est une chaîne d'unités phonétiques qui représente le résultat de la segmentation et de l'identification de ces unités dans l'émission vocale d'un locuteur prononçant une phrase du français.

Cette chaîne de phonèmes est évidemment remplie d'erreurs par rapport à la transcription phonémique de la phrase prononcée.

Le but est de comprendre la phrase prononcée, d'y répondre et d'en sortir la transcription orthographique.

SUMMARY

This paper presents a program which performs connected speech understanding using phonology syntax and semantics.

The imput of the program is a string of phonemes obtained by sementing and recognizing phonemic units in the utterance of a french speaker.

This string of phonemes is full of errors with respect to the right translation of the uttered sentences.

Our aim is to understand, to answer this sentence, and to carry out its orthographical transcription.

"Compréhension automatique de la parole continue à l'aide de la phonologie, la syntaxe et la sémantique".

G. BATTANI et H. MELONI

A) INTRODUCTION

La reconnaissance automatique de la parole continue ne peut se faire par un décodage du signal acoustique isolé de tout contexte syntaxique et sémantique.

Le fait de savoir que le locuteur s'exprime en français, dans une syntaxe précise, en utilisant un vocabulaire particulier, à propos d'un sujet déterminé, apporte assez d'informations et de contraintes pour contribuer de manière décisive à la reconnaissance de la phrase prononcée.

Le but de notre travail est de construire un programme ("DIALOGUE") capable de mettre en oeuvre une partie des contraintes phonologiques, syntaxiques et sémantiques pour tenter de reconnaître automatiquement des phrases fournies sous la forme d'une chaîne de phonèmes (1).

Des systèmes existent déjà qui fournissent, à partir du signal acoustique émis par un locuteur, une chaîne de phonèmes représentant une tentative de traduction phonologique de ce signal. Cette chaîne contient évidemment de nombreuses erreurs par rapport à la transcription orthoépique de la phrase prononcée.

Nous utilisons dans cette étude, les résultats fournis par de tels systèmes bien qu'un processus de reconnaissance établissant des interactions entre tous les niveaux soit, à terme, certainement plus satisfaisant.

Nous avons utilisé les résultats fournis par le système KEAL (2) du C. N. E. T. - LANNION, et c'est avec la collaboration de G. MERCIER et de l'équipe de Reconnaissance de la Parole que nous avons pu utiliser KEAL pour faire les tests systématiques nécessaires à notre travail.

B) DOMAINE DE LA CONVERSATION

1/ Sujet de la conversation

Il s'agit de répondre aux questions d'un locuteur ayant des connaissances générales sur les systèmes conversationnels de conduite des ordinateurs, qui désire apprendre à se servir du système CP.CMS sur un ordinateur IBM 360-67 (3).

Ce sujet de conversation pouvant d'ailleurs être modifié sans changer fondamentalement la structure du programme.

Il a été choisi pour les raisons suivantes :

- son vocabulaire est limité : nous avons choisi environ 200 mots;
- la syntaxe des questions peut être restreinte, tout en permettant au locuteur de s'exprimer de façon naturelle;
- la sémantique attachée aux questions peut aussi être limitée car le nombre des commandes les plus utilisées est assez faible.

Ce système a également l'avantage d'être clair et concis pour l'utilisateur.

2/ Indications sur la sémantique du domaine

Le locuteur peut poser des questions lui permettant de construire, modifier, effacer des fichiers de programmes ou de données, compiler des programmes, les exécuter, faire écrire les résultats sur différents périphériques ou fichiers, etc...
DIALOGUE doit comprendre les questions et y répondre.

3/ Indication sur la syntaxe des phrases acceptées

Les phrases acceptées ont l'un des 4 types suivants :

P1 : Ce type de phrase est prononcé par le locuteur en réponse à une question de DIALOGUE, ou bien pour arrêter la conversation.

exemples : oui.
non.
au revoir.

P2 : C'est le type de question le plus général du système.

exemples :-Comment est-ce qu'on modifie le contenu du fichier ?
-Quelle est la manière de tester l'existence d'un fichier de programmes ?
-Comment supprimer une instruction ?

P3 : Elle fournit un complément d'information après une question de DIALOGUE.

exemples : - Algol.
- Sur l'imprimante.

P4 : Ce sont des questions au sujet d'un terme sur lequel le locuteur veut obtenir des informations.

exemples : - Qu'est-ce que la désignation d'un fichier ?
- Qu'est-ce qu'une option de compilation ?

4/ Indications sur la prononciation des phrases

Le locuteur prononce les phrases d'une façon naturelle, en effectuant les liaisons qu'il désire.
La phrase est traitée par un système cité dans l'introduction. DIALOGUE prend comme donnée la chaîne résultant de ce traitement.

C) MISE EN OEUVRE DES CONTRAINTES SEMANTIQUES, SYNTAXIQUES ET PHONOLOGIQUES

1/ Précisions sur le langage de programmation utilisé

Le programme est entièrement écrit en PROLOG.
C'est un puissant langage de programmation développé par le Groupe Intelligence Artificielle de MARSEILLE-LUMINY (4). Il peut être comparé à PLANNER, QA4, ND-LISP par exemple. A la base, c'est un démonstrateur automatique de théorèmes pour des problèmes exprimables en clauses de HORN.

Les instructions d'un programme sont :

a) des formules de logiques du 1er ordre (sous la forme de clauses de HORN)

exemple :

```
**CHOIX.

+CHOIX(NIL,NIL,NIL) -/ -IMPASSE.
+CHOIX(*X1*Y,*Z,*S) -CHOIX(*X,*Z,*S).
+CHOIX(*X1*Y,*Z,*S) -/ -CHOIX(*Y,*Z,*S).
+CHOIX(*X.*Y,*Y,*S) -CHOIX(*X,*Z,*S).
+CHOIX(*X.*Y,*Z,*S) -/ -CHOIX(*Y,*Z,*S).
+CHOIX(*X,NIL,*X).
```

b) des règles de reécriture ayant au moins la puissance des règles de type 0 (CHOMSKY) mais dont les terminaux et non terminaux, en nombre illimité sont des termes de logique du 1er ordre (donc pouvant contenir des fonctions et des variables) (5).

Ces règles peuvent également être soumises à des conditions exprimées de la manière (a) et sont entièrement traduisibles sous cette forme.

exemple :

```
** DEBUT DE PHRASE.
:DEBUT(*TP,*CO.*C2)==:COM(*CO.*C1) :ESTCEQUON(*TP,*C1.*C2)-S(1.*CO).
:DEBUT(INF,*CO.*C5)== :QUEL(*GN,*CO.*C1) :EST(*C1.*C2)
    :DET(*GN,NIL,*C2.*C3) : MANIERE(*X,*GN,*C3.*C4)
    :PREP(POUR.DE.NIL,*X,*C4.*C5)-S(1B.*CO).
:DEBUT(INF,*CO.*C2)== =SUJ(*P,*CO.*C1) :VEUX(*P,*C1.*C2)-S(1C.*CO).
```

2/ Traitement syntaxique et phonologique

Il est effectué à l'aide d'un analyseur formé de règles de réécriture (c.1.b)

Il fonctionne suivant une stratégie descendante, avec retour en arrière. Fondamentalement, pour analyser une chaîne il "suppose" une structure puis il vérifie que la chaîne correspond à ce qu'il a supposé, sinon il modifie localement la structure et recommence.
Ces suppositions et vérifications sont bien entendu très nombreuses tout au long de l'analyse.
La meilleure façon d'optimiser ce type d'analyseur, c'est de faire en sorte qu'il soit déterministe, c'est-à-dire que la solution qu'il choisit soit toujours la bonne et qu'il n'ait pas besoin de revenir en arrière.

Mais la composition même de ce que nous analysons nous interdit d'espérer une aussi bonne solution. La chaîne de phonèmes est en effet remplie d'erreurs et dans le cas idéal où il n'y en aurait pas, des risques de mauvaise segmentation en mots ou de segmentations multiples persisteraient. L'indéterminisme de l'analyse pourrait donc se situer à trois niveaux :

- choix de la structure (arbre d'analyse)
- choix des segmentations
- choix dans l'analyse des erreurs.

Nous avons donc choisi une méthode permettant, dans le cas où le locuteur prononce une phrase compréhensible par "DIALOGUE" de faciliter son analyse en réorganisant les choix dans chacun des trois niveaux, de façon à transformer en avantage les inconvénients du non-déterminisme.

a) règles concernant l'analyse structure ← mots

D'une manière systématique, l'analyse ne se fait pas seulement en tenant compte de la syntaxe mais utilise pour son fonctionnement des connaissances qui lui sont fournies soit par la partie sémantique, soit par le début de la phrase en cours d'analyse (avantage de l'analyse descendante).

La partie sémantique indique à l'analyseur quel type de phrase et quel vocabulaire ont la plus forte probabilité d'apparaître dans l'intervention suivante du lo-

cuteur, compte-tenu de l'état de la conversation et des contraintes sémantiques du domaine.
De même, la reconnaissance du début de la phrase et du verbe donne des indications précises sur les compléments possibles.

L'analyseur utilise ces sources de connaissance d'une façon systématique pour choisir la solution la plus probable, ce qui est nécessaire pour des données aussi floues.

b) Règles assurant le passage mots → transcription phonologique

Une partie du programme prenant comme données les mots, leur fonction, leur transcription phonologique génère la plupart des règles de passage des mots à leur prononciation exprimée par des non-terminaux se transformant ultérieurement en phonèmes.

c) Règles de modification de la transcription phonétique (liaisons, pluriel... etc.)

Un ensemble de règles permet de modifier en cours d'analyse la transcription phonétique d'un mot, dans le cas où celui-ci doit être accordé au pluriel ou bien, lorsqu'il est susceptible de prendre une forme acoustique différente lors d'une liaison.

Ces règles sont indépendantes du matériel, du locuteur et des programmes.

d) Règles caractérisant le modèle d'erreurs

Ces règles permettent de modifier la transcription phonétique idéale des mots en contexte afin d'en générer les diverses variantes issues des déformations produites dans les divers composants du système de segmentation.

L'application de ces règles doit être hiérarchisée de manière à obtenir tout d'abord les suites de phonèmes les plus probables.

Ces règles ont la forme suivante :

$$Xa_1 \ldots a_n = Y_1 \ldots Y_m$$

où X est un non-terminal, $a_1, \ldots, a_n$ sont des terminaux $Y_1, \ldots, Y_m$ sont, soit des terminaux, soit des non-terminaux.

Ces règles permettent, entre-autre, de traiter les cas suivants (6) : élision, substitution, insertion, compression, dispersion.

3/ Traitement sémantique

3.1 Organisation du dialogue locuteur-machine

Le dialogue entre le locuteur et la machine se fait sous forme de questions et de réponses alternatives entre les deux "interlocuteurs".

a) interventions du locuteur :

i) questions

Le locuteur désire s'informer d'une part sur la manière dont il peut réaliser une tâche à l'aide du système CP.CMS et, d'autre part, sur le contenu sémantique des divers individus; il dispose pour cela des phrases de type P2 et P4.

ii) réponses

Les réponses permettent d'apporter une information supplémentaire au système, d'affirmer ou d'infirmer une assertion.
Ces phrases viennent en réponse à une question posée par le système; leur syntaxe correspond aux types P1 et P3.

b) interventions du système :

i) questions

Lorsque la phrase prononcée par le locuteur ne désigne pas sans ambiguité une tâche (dans le cas d'une phrase de type P2), le système doit essayer d'obtenir les renseignements qui lui sont utiles pour comprendre la question posée. Il doit donc interroger le locuteur afin d'obtenir ces spécifications.

ii) réponses

Si la question posée par le locuteur permet d'envisager sans ambiguité possible une action, le système doit fournir la réponse qui autorise le locuteur à utiliser les commandes nécessaires à la réalisation de la tâche décrite.

3.2 Traitement des phrases de type P2

a) actions sémantiques :

Les actions sémantiques ont été définies par rapport aux commandes disponibles dans notre sous ensemble du système CP.CMS.
En général, une action sémantique représente une tâche directement exécutable à l'aide d'une instruc-

tion du système de commandes, c'est le cas notamment des actions qui ne demandent pas pour leur exécution l'initialisation d'un contexte particulier.
Certaines actions peuvent être communes à plusieurs requêtes, d'autres exigent un environnement précis. Le système tolère également des actions à caractère informatif qui impliquent tout un ensemble de requêtes.
Les informations qui caractérisent ces actions sont regroupées dans un dictionnaire.

b) accès aux actions sémantiques :

Les phrases syntaxiquement correctes sont transmises par l'analyseur de syntaxe sous la forme d'une structure profonde qui contient l'essentiel des informations utiles pour la description d'une tâche. Cette structure caractérise la classe des phrases qui décrivent une même action.

c) génération des réponses :

La réponse du système est fonction de la coïncidence de la structure profonde de la phrase avec l'une ou plusieurs des actions mémorisées.
Lorsqu'une réponse est donnée au locuteur, elle est construite à partir des informations codées dans les dictionnaires. Sinon DIALOGUE demande des précisions afin d'identifier correctement l'action envisagée.

3.3 Interactions entre l'analyse syntaxique et la sémantique

Lors de la génération des réponses, le système possède des informations sur l'état de la conversation et peut en déduire son évolution future (l'évolution la plus probable). Ces informations sont transmises sous forme de paramètres à l'analyseur afin de limiter ses recherches dans le dictionnaire.

Ces paramètres sont composés, d'une part du type de la phrase attendue (dans une évolution logique du dialogue) et, d'autre part d'une liste de classes de mots représentant les accès au dictionnaire. Ces classes contiennent les mots synonymes (verbes, noms, adjectifs) qui ont une forte probabilité d'apparaître dans la phrase suivante.

4/ Règles phonologiques

Les erreurs continues dans la chaîne de phonèmes à analyser sont liées au matériel utilisé pour la conversion nu-

mérique du signal acoustique, aux programmes de segmentation, au locuteur et aux variations contextuelles des unités phonétiques.

Nous avons établi un modèle des erreurs pour un locuteur déterminé sur lequel a été fait un apprentissage des diverses formes des phonèmes. Ce modèle n'est évidemment pas figé et seuls les tests complets de chacune de ses variantes permettra de choisir celui qui obtient les meilleures performances.

Il a été établi en testant systématiquement toutes les voyelles et les consonnes dans presque tous les contextes de voisinage immédiat possibles. Les altérations de ces phonèmes nous ont conduit à diverses interprétations dont nous avons déduit un ensemble de règles qui rendent compte de la majorité des modifications contextuelles des unités phonétiques.

D) CONCLUSION

Le système est actuellement testé et, bien que l'analyseur syntaxique et la partie sémantique soient dès maintenant terminés, les règles qui caractérisent le modèle d'erreur sont sujettes à de nombreuses améliorations en fonction de l'efficacité de la reconnaissance.

Un modèle trop large des erreurs autorise la production d'un certain nombre d'ambiguités, alors qu'un modèle trop restreint ne permet pas la reconnaissance de toutes les phrases prononcées. Nous devons donc établir un système de règles qui limite les inconvénients de chacun des extrèmes.

BIBLIOGRAPHIE

(1) G. BATTANI - H. MELONI : Thèse de 3ème cycle à soutenir.
(2) MERCIER et Col. : Reconnaissance de grands dictionnaires prononcés par plusieurs locuteurs, 5èmes journées d'étude du groupe communication parlée, mai 1974, ORSAM.
(3) IBM CP 67/CMS : Users Guide (Oct. 70)
(4) G. BATTANI - H. MELONI : INterprèteur du langage de programmation PROLOG.
(5) A. COLMERAUER : Article à paraître sur les grammaires de métamorphose.
(6) BUISSON et coll. Phonétic decoding for automatic recognition of words. P. 189-196, speech com. seminar - Stockholm Avril 1974.

Nous donnons à titre d'exemple une conversation tenue avec le système. La suite des phonèmes est entrée sous la forme d'un arbre dont les noeuds sont marqués par la fonction "." (point) à deux arguments (elle est notée sous forme d'opérateur). Cette chaîne ne présente pas d'erreurs mais seulement les altérations produites par les liaisons.

```
-ECOUTE!

K.O.M.AN.K.ON.F.I.L.EI.UN.F.R.O.G.R.A.M.E.NIL,        ← chaîne d'entrée
COMPILER FICHIER                                      ← partie de la structure profonde

COMMENT COMPILER UN PROGRAMME                         ← Phrase reconnue
VOUS DESIREZ:                                         ← Réponse
COMPILER UN FICHIER ALGOL, OU COMPILER UN FICHIER FORTRAN?

A.L.G.O.L.NIL,
ALGOL

ALGOL

VOUS DEVEZ:
CREER UN FICHIER.
VOUS TAPEZ:
ALGOL (NOM DU FICHIER)
LE TYPE DU FICHIER DEVANT ETRE 'ALGOL'
VOUS TROUVEREZ LES INDICATIONS PAGE 324.

K.O.M.AN.T.AI.S.K.ON.K.R.EI.S.E.F.I.CH.Y.EI.NIL,
CREER FICHIER

COMMENT EST CE QU' ON CREE CE FICHIER

VOUS DEVEZ:
INITIALISER EDIT.
VOUS TAPEZ:
INPUT
POUR SORTIR DE L'ENVIRONNEMENT INPUT, FAIRE DEUX (FC)
VOUS TROUVEREZ LES INDICATIONS PAGE 61.

J.E.V.EU.M.O.D.I.F.Y.EI.MON.P.R.O.G.R.A.M.E.NIL,
MODIFIER FICHIER

JE VEUX MODIFIER MON PROGRAMME

VOUS DESIREZ:
MODIFIER LE CONTENU D'UN FICHIER, OU MODIFIER LA DESIGNATION D'UN FIC
HIER?

L.E.K.ON.T.E.N.U.NIL,
CONTENU

LE CONTENU

VOUS AVEZ LA POSSIBILITE DE SUPPRIMER, D'INSERER, OU DE REMPLACER
DES LIGNES DANS UN FICHIER.
VOUS TROUVEREZ LES INDICATIONS PAGE 59.
```

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

RECHERCHES SUR LA STRUCTURATION PROSODIQUE DE LA PHRASE FRANCAISE (ESSAI D'ANALYSE PHONOSYNTAXIQUE)

Albert DI CRISTO (Institut de Phonétique d'Aix)

RESUME

Ce travail, fondé d'une part sur l'étude acoustique et perceptuelle des structures prosodiques d'un important corpus de français parlé et, d'autre part, sur l'analyse de certaines structures syntaxiques de ce même corpus, permet de proposer un premier modèle phonosyntaxique, susceptible d'être applicable à la reconnaissance de la parole.

SUMMARY

This work was founded on one hand on a detailed acoustical and perceptual study of theprosodic features of a large corpus of spoken french,and on the other hand on the analysis of certain syntactical structures. As a result, a first phonosyntactical model, which could be applied to speech recognition, has been proposed.

RECHERCHES SUR LA STRUCTURATION PROSODIQUE DE LA PHRASE FRANCAISE (ESSAI D'ANALYSE PHONOSYNTAXIQUE)

Albert DI CRISTO

(Institut de Phonétique d'Aix)

Si certains linguistiques affirment, à la suite de M.A.K. HALLIDAY (1961,1963,1967), que toutes les fonctions assumées par l'intonation sont grammaticales, d'autres partagent les réticences de A. CRUTTENDEN (1970) et considèrent que ces fonctions n'ont aucun rapport direct avec la syntaxe.

Entre ces attitudes radicalement opposées, il est possible de relever un certain nombre de conceptions plus nuancées, que nous nous contenterons de résumer brièvement.

a) De nombreux auteurs (DANES 1960, SIERTSEMA 1962, FREI 1968) ont souligné l'importance des facteurs prosodiques pour la signalisation du thème et du prédicat de l'énoncé. Dans cette perspective, M. ROSSI (ROSSI 1973) a dégagé récemment les indices acoustiques et perceptuels caractéristiques de l'intonation prédicative dans les phrases françaises transformées par permutation.

b) Selon G. FAURE (1962, 1969, 1970), le découpage en groupes mélodiques assure l'actualisation des structures syntaxiques de l'énoncé et permet de mettre en évidence la hiérarchie des unités successives qui le constituent. D'autre part, en se fondant sur l'analyse d'oppositions comme : "Mais oui mon cher, réellement" / "Mais oui mon cher Rey, elle ment" et "Elle est maladroite" / "Elle est mal à droite", G. FAURE a montré qu'un simple déplacement de la frontière mélodique pouvait faire apparaître des unités lexicales et syntaxiques entièrement différentes. Ces observations paraissent confirmer l'hypothèse de l'Ecole Structuraliste (cf. TRAGER and SMITH, 1951), suivant laquelle la structure constituante de la phrase serait décodée par l'auditeur à partir des indices acoustiques présents dans le signal de parole.

c) D'autres spécialistes pensent en revanche que ces indices prosodiques ne sont pas nécessaires à l'identification des structures syntaxiques, ni par conséquent, à l'interprétation des énoncés (LIEBERMAN 1967, NASH 1970). D'après Ph. LIEBERMAN, l'intonation ne jouerait un rôle structural effectif que dans les cas d'ambiguïté : "It is only when the speaker is trying to disambiguate the sentence that he will consistenly segment smaller constituents by means of

intonation" (LIEBERMAN 1965, p. 124).

d) Après les publications de N. CHOMSKY et de ses disciples, de plus en plus nombreux sont les chercheurs qui, reprenant le voie précocément tracée par R.P. STOCKWELL (1960), s'attachent à définir la place de l'intonation dans la grammaire générative et transformationnelle (BIERWISCH 1966, RIVARA 1973, YORIO 1973, HIRST 1974, HIRST and GINESY 1974, GROSS 1974). Les travaux les plus récents poursuivis dans cette direction (DOWNING 1970, POPE 1971, BRESNAN 1972, STOCKWELL 1972, BERMAN and SZAMOSI 1972, LAKOFF 1972) semblent aboutir à la conclusion que les structures intonatives ne sont pas exclusivement déterminées par les constituants des structures superficielles et que l'on doit pouvoir prédire la localisation et la forme de certains contours intonatifs à partir de l'analyse des structures profondes et mi-profondes (deep and shallow structures) sous-jacentes.

Nous nous proposons d'évaluer, dans cette communication, la valeur opérationnelle des indices prosodiques pour la structuration prosodique de la phrase française. Nous limiterons volontairement notre recherche aux structures constituantes dérivées (structures superficielles) et nous tenterons de répondre en partie aux questions suivantes : Peut-on décaler dans le signal de parole des indices acoustiques et perceptuels révélateurs de la structure constituante ? Ces indices sont-ils toujours présents ? Quelle est leur nature ? Sont-ils uniquement liés à la structuration syntaxique, ou bien sont-ils également tributaires d'autres facteurs (variantes individuelles et stylistiques, contraintes physiologiques, etc...) ?

PRINCIPES D'ANALYSE

Comme l'a justement souligné MAEDA (1974), dans son étude sur les contours intonatifs de l'anglo-américain, la plupart des recherches qui traitent des relations entre la prosodie et la syntaxe sont fondées sur des impressions auditives. Or, ce type d'analyse a été vivement critiqué par Ph. LIEBERMAN. Celui-ci a en effet montré qu'il existait des divergences manifestes entre la réalité physique du signal prosodique et la transcription auditive de ce même signal, effectuée par des linguistiques (LIEBERMAN 1965). Sans doute est-ce là une des raisons qui ont conduit de nombreux chercheurs à n'étudier les phénomènes prosodiques que du seul point de vue objectif (MAEDA 1974, VAISSIERE 1974). Il est toutefois regrettable que ces auteurs aient choisi de fonder leur étude sur l'analyse d'un seul paramètre : les variations de la fréquence fondamentale (Fo). Bien que l'importance de ce dernier soit couramment admise, il reste encore à prouver que les autres paramètres : durée, intensité, tempo, timbre ne jouent qu'un rôle négligeable dans la structuration

de l'énoncé. C'est pourquoi nous avons choisi d'adopter une approche de type paramétrique (cf. CRYSTAL 1969), qui consiste à analyser systématiquement tous les paramètres constituant le signal prosodique.

Le corpus qui a servi de base à notre recherche est constitué de 250 phrases (1) enregistrées par cinq locuteurs masculins représentatifs du français général (2).

Au cours d'une première étape, nous avons présenté à un groupe d'informateurs ces enregistrements, afin d'en exclure tous les exemples jugés artificiels.

Nous avons procédé ensuite à une analyse acoustique exhaustive (Fo, intensité, durée) des phrases retenues, en faisant abstraction des phénomènes micro-prosodiques (COHEN and't HART 1967, WENDAHL 1967, KIM 1968, LEON et MARTIN 1970, VAISSIERE 1971, WANG 1972, LARREUR et BOE 1973).

La troisième partie de notre travail a porté sur la transformation des valeurs objectives obtenues en valeurs perceptuelles. Cette conversion n'a pu être réalisée qu'au terme d'une double procédure qui a consisté, d'une part, à pondérer les données objectives en fonction des caractéristiques intrinsèques (HOUSE and FAIRBANKS 1953, LEHISTE and PETERSON 1961, MOHR 1971, ROSSI 1971b, BOE 1972) et, d'autre part, à corriger ces données en tenant compte des valeurs de seuil connues (ROSSI 1971a, ROSSI 1972, ROSSI et CHAFCOULOFF 1972b) (3).

---

(1) Ce corpus comprend notamment les 50 phrases proposées par P. DELATTRE dans son article sur "L'intonation par les oppositions" (DELATTRE 1969). Les 250 phrases comportent de 1 à 6 groupes prosodiques. La moyenne étant d'environ 3 groupes par phrase, notre étude porte sur 250 x 5 x 3 = 3 750 groupes prosodiques.

(2) Les enregistrements ont été effectués dans une chambre anéchoïde, à l'aide d'un magnétophone Philips EL.3503. LA calibration de chaque enregistrement a été réalisée en utilisant un signal pur de 1000 hz et de 80 db, diffusé par une source située à 30 cm du microphone. Pour l'analyse instrumentale, nous avons utilisé le mélodimètre de l'Institut de Phonétique d'Aix et l'intensimètre des Laboratoires d'Electronique et de Radio Electricité de l'Université de Grenoble.

(3) En ce qui concerne le seuil différentiel d'intensité et le seuil de variation d'intensité, les valeurs nous ont été obligeamment communiquées par M. ROSSI qui poursuit actuellement d'importants travaux dans ce domaine.

Ces nouveaux résultats nous ont permis de déterminer les niveaux intonatifs perceptuels respectifs de nos cinq locuteurs, en appliquant la méthode définie par (ROSSI et CHAFCOULOFF 1972b).

Nous avons, enfin, au cours d'une dernière étape, procédé à la segmentation des différentes phrases en groupes prosodiques (G. P.), ce dernier étant défini comme une unité suprasegmentale, délimitée par une variation perceptuelle significative d'un ou de plusieurs paramètres prosodiques.

Il convient de préciser, dès à présent, que durant cette phase de segmentation, chaque G. P. a été l'objet d'une double analyse : qualitative et quantitative. La figure 1 représente, à titre d'exemple, les variations perceptuelles schématisées du Fo d'un G. P. continuatif. Nous appellerons respectivement : attaque, la syllabe initiale du G. P., tonique, la dernière syllabe du G. P., et prétonique, l'avant-dernière syllabe du G. P.. Nous considérerons l'attaque, la prétonique et la tonique comme les points-clés de notre analyse.

La tonique constitue le contour du G. P. (centre prosodique où s'effectuent généralement en français les variations paramétriques significatives). L'ensemble (prétonique + contour) correspond à la cadence du G. P. Enfin, la portion comprise entre l'attaque et la prétonique délimite le précontour du G. P.

L'analyse qualitative consiste à décrire les configurations caractéristiques du G. P., du contour et du précontour.

L'étude quantitative porte sur les point suivants :

a) Calcul des écarts types des valeurs perceptuelles de l'attaque, de la prétonique, du début et de la fin de la tonique

b) Evaluation des rapports : R1 (attaque/fondamental usuel)(4), R2 (fin tonique/prétonique) et R3 (fin tonique/début tonique), ainsi que de leurs écarts types.

Une analyse similaire des points-clés mentionnés plus haut a également été effectuée pour les durées et les intensités.

---

(4) Le fondamental usuel est déterminé, pour chaque locuteur, par le calcul de la valeur moyenne des syllabes inaccentuées (cf. LEON et MARTIN 1970, ROSSI et CHAFCOULOFF 1972a).

RESULTATS

1) Définition des règles phonosyntaxiques

On admet généralement que la phrase française minimale (sans enchassement ni transformation), définie par la modalité de base : assertion positive, est formée de deux constituants immédiats : le groupe nominal (G. N.) et le groupe verbal (G. V.), qui a le statut de prédicat (DUBOIS 1969).

(I) P ⟶ G. N. + G. V.

La structure constituante P est délimitée par une frontière terminale, tandis que les constituants immédiats (G. N. et G. V.) sont séparés par une frontière syntaxique non terminale majeure (F. S. n. T. M.) (5).

Pour de nombreux générativistes, la règle (I) peut être réécrite sous la forme :

(II) P ⟶ G. N. + G. Préd. + (Circonst.)

Si nous adoptons cette notation, nous obtiendrons la série suivante :

(III) G. N. ⟶ [Nom propre]
[Dét. + Nom]
[Proclitique (Je, tu, il, ça, etc...)]

(IV) G. Préd ⟶ Auxil. + G. V.

(V) G. V. ⟶ Base verb. + G. N. + G. Prép.

(VI) G. Prép. ⟶ Prép. + G. N.

(VII) Circonst. ⟶ [Adv.]
[prép. + G. N.]

Les résultats de l'étude acoustique et perceptuelle, effectuée au cours de la première partie de notre étude, nous permettent de formuler la règle phonosyntaxique obligatoire suivante :

(A) - G. N. ⟶ G.N./,,/ / $\underset{P}{((\text{———})}$ + (G. Préd.)$\underset{P}{)}$

(5) Il s'agit en fait d'une F. S. n. T. M. dérivée puisque nous nous préoccupons exclusivement des structures superficielles.

Cette première règle implique que dans une structure constituante : P → G. N. + G. Préd., les constituants immédiats sont séparés par une frontière prosodique non terminale Majeure (F. P. n. T. M.), que nous symboliserons ainsi : /,,/ (6)

exemples : (1) Paul /,,/ mange /./ (7)
(2) Les enfants/,,/ s'amusent/./

Toutefois, la règle (A) n'est plus applicable si G.N. est un proclitique (Je, tu, il, elle, nous, vous, ils, elles, ça), c'est-à-dire après une transformation par pronominalisation.

Plusieurs possibilités sont offertes dans ce cas.

a) P → G.N. + G.Préd.

Si le G.Préd. n'est constitué, outre l'auxiliaire, que d'une base verbale, la phrase ne comporte qu'un seul groupe prosodique, délimité par une frontière prosodique terminale (F. P. T.).

exemples : (3) IL part /./
(4) Elle travaille/./

Quand le G.Préd. comprend, outre l'auxiliaire, une base verbale plus un G.N. la frontière prosodique non terminale majeure (F. P. n. T. M.) se situe devant le G.N. du G.V. (c'est-à dire devant G.N. 2)

(5) ((elle) ((invite) /,,/ (le président)))
G.N.1 base G.N.2
G.Préd. G.Préd.
P. P.

Nous avons relevé une structuration prosodique analogue dans la phrase dérivée :

(6) ((il) ((a vendu) (son château)))
G.N.1 base G.N.2
G.Préd. G.Préd.
P. P.

---

(6) En ce qui concerne les indices prosodiques perceptuels de cette F.P.n.T.M., voir plus loin.

(7) Le symbole /./ indique une frontière prosodique terminale (F.P.T.).

ainsi que dans les phrases du type :

(7) Elle demande /,,/ qui va rentrer/./
(8) Elle a dit/,,/ quel scandale/./

qui sont représentées par le même indicateur syntagmatique (voir les diagrammes 1a et 1b).

Nous pouvons ainsi formuler une seconde règle phonosyntaxique :

(B) G.N.2 $\rightarrow$ /,,/G.N.2/ ((proclit.) + ((base) + (——)))
G.N.1 G.Préd. G.Préd.
P. P.

Il convient de noter que si l'on substitue au proclitique du G.N.1. un nom propre, ou la suite : (dét. + nom) :

(9) Jean-Jacques a vendu son château/./
(10) Marie demande qui va rentrer /./,

la règle (A) fonctionne de nouveau, mais les résultats des analyses acoustiques et perceptuelles permettent, dans ce cas, de déceler la présence d'une frontière prosodique non terminale mineure (F. P. n. T. m.), représentée par /,/ devant le G.N.2. D'où la règle phonosyntaxique (C) :

(C)/G.N.1x G.N.2y /$\rightarrow$G.N.1/,,/x /,/ G.N.2y / ((—)+((base) + (——)))
P. P.

Cette F.P.n.T.m. n'est d'ailleurs pas toujours présente, ainsi que le révèle l'analyse instrumentale. Nous dirons que la règle qui la postule est facultative :

G.N.2 $\rightarrow$ (/,/) G.N.2

b) P $\rightarrow$ G.N. + G.Préd. + Circonst.

- Lorsque le G.Préd. ne comprend, outre le constituant auxiliaire, que la base verbale; et le circonstanciel, un constituant ultime monosyllabique :

(11) Elle travaille bien/./,

la phrase ne comporte qu'un groupe prosodique, délimité par une frontière prosodique terminale (F.P.T.), comme c'était le cas pour (3) et (4). Nous pensons que les réalisations (3), (4), (11) sont conditionnées par des contraintes de nature physiologique.

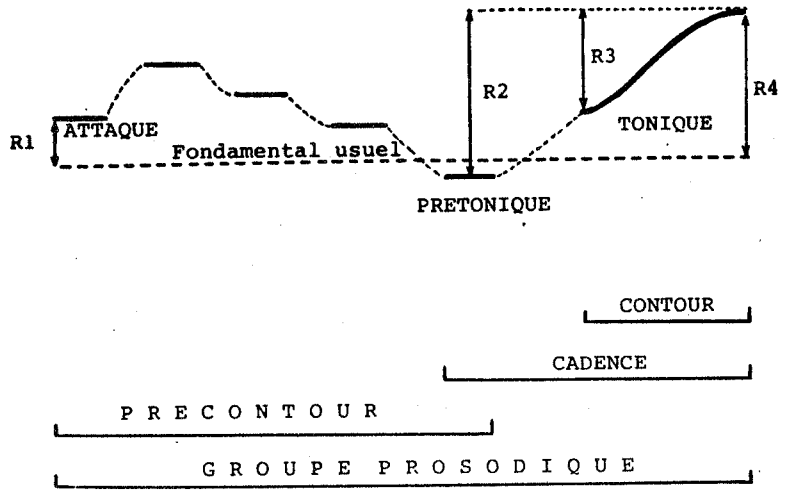


Figure 1. Représentation schématique des variations perceptuelles du Fo d'un Groupe Prosodique(G.P.) continuatif. R1 exprime le rapport entre la valeur de l'attaque du GP et le fondamental usuel. R2, le rapport entre la valeur finale de la tonique et celle de la prétonique.R3, le rapport entre les valeurs finales et initiales de la tonique. R4, le rapport entre la valeur finale de la tonique et le fondamental usuel.

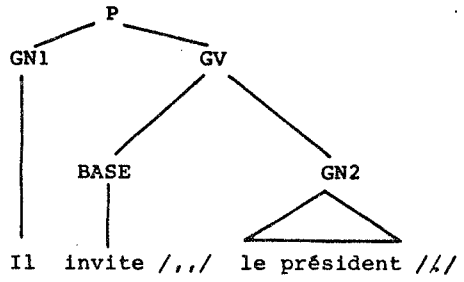


Diagramme 1.a.

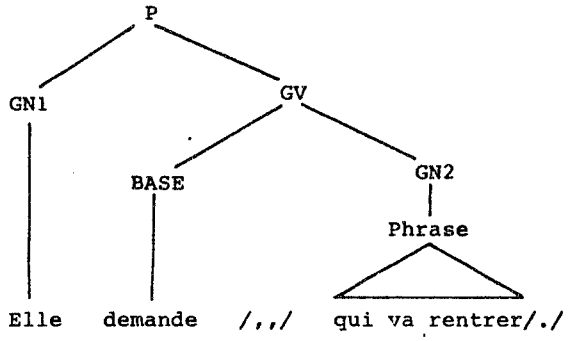


Diagramme 1.b.

- Lorsque le G.Préd. est formé, outre l'auxiliaire, d'une seule base verbale, et le circonst., d'un constituant ultime plurisyllabique, une frontière prosodique non terminale mineure (F.P.n.T.m.) apparaît devant le circonst.

(12) Il travaille /,,/ considérablement/./

La substitution d'un nom propre ou d'une séquence (dét.+nom) au pronom personnel a pour conséquence l'affectation d'une F.P.n.T.M. après le G.N.1. et d'une F.P.n.T.m. devant le circonstanciel :

(13) Marie /,,/ travaille /,/ considérablement/./

Ce qui nous conduit à la règle (D) :

(D) /G.N.1.x Circonst.y/ $\rightarrow$ G.N.1/,,/x /,/circonst.y/$_{P.}$((—)+((base) + (——)))$_{P.}$

- Lorsque le G.Préd. comprend : (Auxil. + base + G.N.)

exemple :

(14) Il a vendu son château en Espagne

(= C'est en Espagne qu'il a vendu son château),

nous obtenons une F.P.n.T.M. devant le circonstanciel "en Espagne" et une F.P.n. T.m. (facultative) devant le G.N.2. "son château" (cf. diagramme 2a).

(15) Il a vendu/,/ son château/,,/ en Espagne/./

Ce qui nous donne la règle phonosyntaxique (E) :

(E) /G.N.2.x Circonst.y/ $\rightarrow$ /,/G.N.2.x /,,/circonst.y/$_{P.}$((——) + (((base) + (G.N.2.)) + (——)))$_{P.}$

L'introduction d'un nom propre ou d'une séquence (dét.+nom) dans le G.N.1., entraîne la réalisation de deux F.P.n.T.M., l'une après le (G.N.1.) et l'autre devant le (circonst.), ainsi que d'une F.P.n.T.m. facultative devant le (G.N.2.).

(16) Les Dupont/,,/ont vendu/,/leur château/,,/ en Espagne/./

On notera que la suite terminale proposée en (14) est ambiguë et peut signifier aussi : "Il a vendu son château qui se trouve en Espagne" (cf.DELATTRE 1969). L'indicateur syntag-

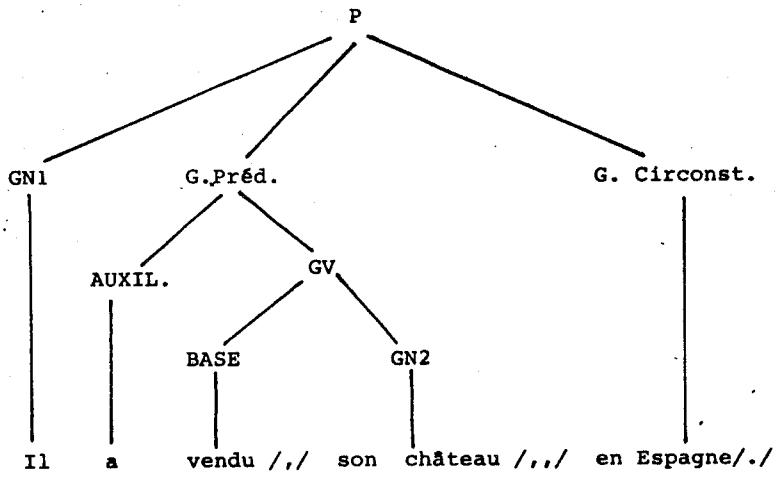


Diagramme 2.a.

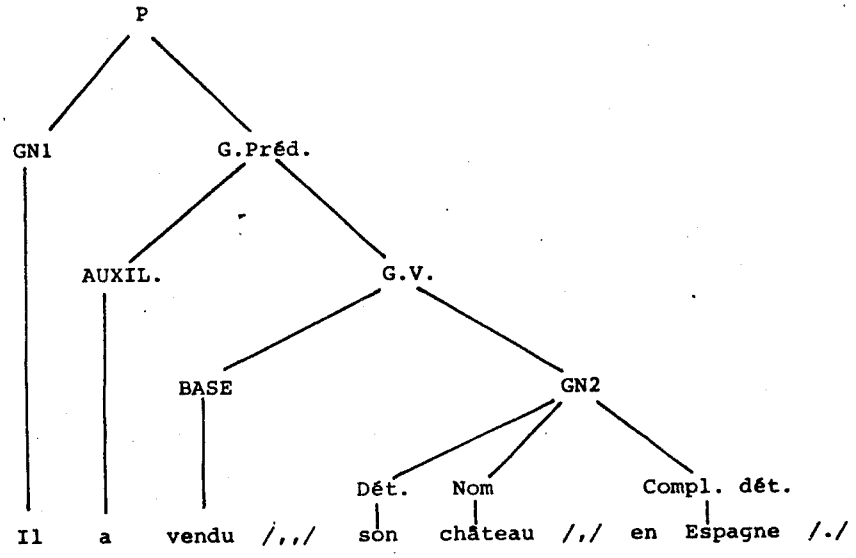


Diagramme 2.b.

matique dérivé de cette structure constituante a été représenté dans le diagramme 2b. L'analyse acoustique et perceptuelle d'exemples de ce type a montré que les sujets réalisaient et percevaient une F.P.n.T.M. devant le G.N.2. (son château) et une F.P.n.T.m. devant le complément déterminatif (en Espagne), cette dernière pouvant être facultative. D'où la règle phonosyntaxique (F) :

(F) /N2x Compl.dét.y→/,,/N2x/,/Compl.dét.y./((GN1) (Aux.+base+
P. G.Préd.

—— + ——))
G.Préd.
P.

N. B. N2→dét. + Nom (son château).

Nous interrompons ici, faute de place, la liste des règles phonosyntaxiques que nous avons réussi à dégager au cours de nos études sur le français. Nous nous intéressons tout particulièrement, à l'heure actuelle, au problème des enchassements (enchassements des adjectifs dans le G.N. et enchassement des relatives attributives et déterminatives). Ces recherches nous permettent de conclure que les indices acoustiques et perceptuels de la structure constituante sont toujours présents et ne se manifestent pas, contrairement à ce qu'affirmait LIEBERMAN, que dans les cas d'ambiguïté. C'est à la définition de ces indices que nous consacrerons la dernière partie de cette étude.

2° Les indices prosodiques perceptuels de la structure constituante.

Nous nous attacherons surtout à décrire les indices perceptuels des F.P.n.T.M. et F.P.n.T.m.

A- La frontière prosodique non terminale majeure

La F.P.n.T.M., dont nous avons montré l'importance pour la structuration syntaxique de la phrase française, est perçue en fonction de plusieurs indices perceptuels réalisés simultanément.

a) Une variation tonale atteignant le niveau 3 (infra-aigu), ou une rupture dans ce même niveau, à la fin du G.P. L'importance de la variation tonale (rapport entre la fin de la tonique et son début) semble être conditionnée par différents facteurs.

- caractéristiques individuelles : le contour est presque toujours réalisé par une variation mélodique chez le locuteur 1, alors que c'est rarement le cas pour les locuteurs 2, 3, et 4.

- contraintes contextuelles : la variation du contour est plus importante lorsque la consonne qui précède la tonique est sonore. Dans le cas d'une sourde, le début du contour est plus élevé et le rapport : fin tonique/début tonique diminue.

- contraintes temporelles : les variations du contour sont moins grandes et vont jusqu'à disparaître si le débit est très rapide.

- facteurs stylistiques : on décèle des variations nettement plus marquées dans la lecture et dans le style des conférenciers ou des journalistes que dans la parole spontanée.

La configuration du contour, contrairement à ce que pensait DELATTRE (DELATTRE 1966), ne semble pas être significative (cf. Figure 2), pas plus d'ailleurs, que la configuration du précontour dont les variations sont liées à des facteurs individuels et stylistiques.

Le trait constant, pour tous les locuteurs, est la rupture dans le niveau 3, comme le démontrent les résultats présentés dans la figure 3. On peut y observer que les écarts types des valeurs de la tonique n'excèdent jamais les limites de ce niveau. Il est nécessaire pour celà que l'écart type du R 4 (fin tonique/Fo usuel) soit compris entre le rapport : limite supérieure du niveau 3/Fo usuel et limite inférieure du niveau 3/Fo usuel. Tel est le cas, pour tous les locuteurs (cf. Tableau 1).

Il convient de préciser que la rupture de F.P.n.T.M. est bilatérale (entre la prétonique et la tonique ; entre la tonique et la voyelle initiale du G.P. subséquent).

Les valeurs de σ sont moins fines en ce qui concerne l'attaque, principalement pour les locuteurs 1 et 4.

Les valeurs de σ de la prétonique montrent que celle-ci à tendance à se placer dans le bas du médium et dans le grave, pour le locuteur 1, alors qu'elle occupe plutôt le haut du médium chez les locuteurs 3 et 4 et qu'elle couvre pratiquement tout le médium pour le locuteur 2.

Nous ajouterons que ces résultats confirment pleinement la validité de la méthode élaborée par ROSSI et CHAFCOULOFF pour la définition des niveaux intonatifs perceptuels.

b) La F.P.n.T.M. se caractérise également par une rupture bilatérale d'intensité, toujours supérieure au seuil différentiel. Les travaux que nous poursuivons actuellement montrent que ce paramètre joue un rôle décisif dans la structuration de l'énoncé, ainsi que dans celle de certains types de phrases.

c) Le troisième paramètre, qui contribue à la reconnaissance de la F.P.n.T.M. est le ralentissement du tempo, qui semble affecter toute la cadence. A titre d'exemple, nous avons pu relever les valeurs moyennes suivantes : syllabe précédant la cadence (60 msec.), syllabe prétonique (100 msec.), syllabe tonique (200 msec.), syllabe postonique (60 msec.).

d) Le quatrième paramètre est constitué par une pause de durée variable (40 à 200 msec.). Mais celle-ci, qui est rarement présente, ne doit pas être considérée comme un indice significatif.

B - La frontière prosodique non terminale mineure

La F.P.n.T.m. se réalise également par un faisceau d'indices perceptuels simultanés.

a) Une rupture tonale dans le niveau 3, mais qui n'est pas nécessairement bilatérale (cf. Loc. 1, figure 4). Etant donné que les toniques des F.P.n.T.M. et F. P.n.T.m. se situent dans le même niveau, le creusement de la prétonique est nettement marqué dans le cas de la F.P.n.T.M., et la configurations du G.P. a, de ce fait, tendance à prendre une forme concave.

Si nos travaux justifient en partie le bien fondé de la distinction établie par DELATTRE (DELATTRE 1966) entre continuité mineure et continuité majeure (Cm ≠ CM), ils infirment en revanche la valeur des traits que ce dernier leur attribue (rappelons que dans l'analyse de DELATTRE, la Cm est définie par une configuration convexe 2-3 et la CM, par une configuration convexe 2-4).

b) Une rupture d'intensité, qui peut être, soit monolatéralement négative (tableau 4, loc. 4), soit bilatéralement négative (tableau 4, locs. 1 et 3).

c) Un ralentissement du tempo de la cadence (mais ce dernier est moins important que pour la F.P.n.T.M.).

d) L'absence de pause à la fin du G.P.

CONCLUSION

Le travail que nous venons de présenter ne constitue en fait que la première étape d'une vaste recherche qui vise, d'une part, à dégager de nouvelles règles phonosyntaxiques et, d'autre part, à étendre l'étude de ces règles aux transformations interrogatives (T. Interr.) et contrastive (T. Contr.).

Les résultats que nous avons déjà obtenus devraient nous conduire à définir prochainement un algorithme permettant, soit de générer les structures prosodiques à partir des structures constituantes, soit de déduire de l'analyse des structures prosodiques les constituants syntaxiques qu'elles actualisent.

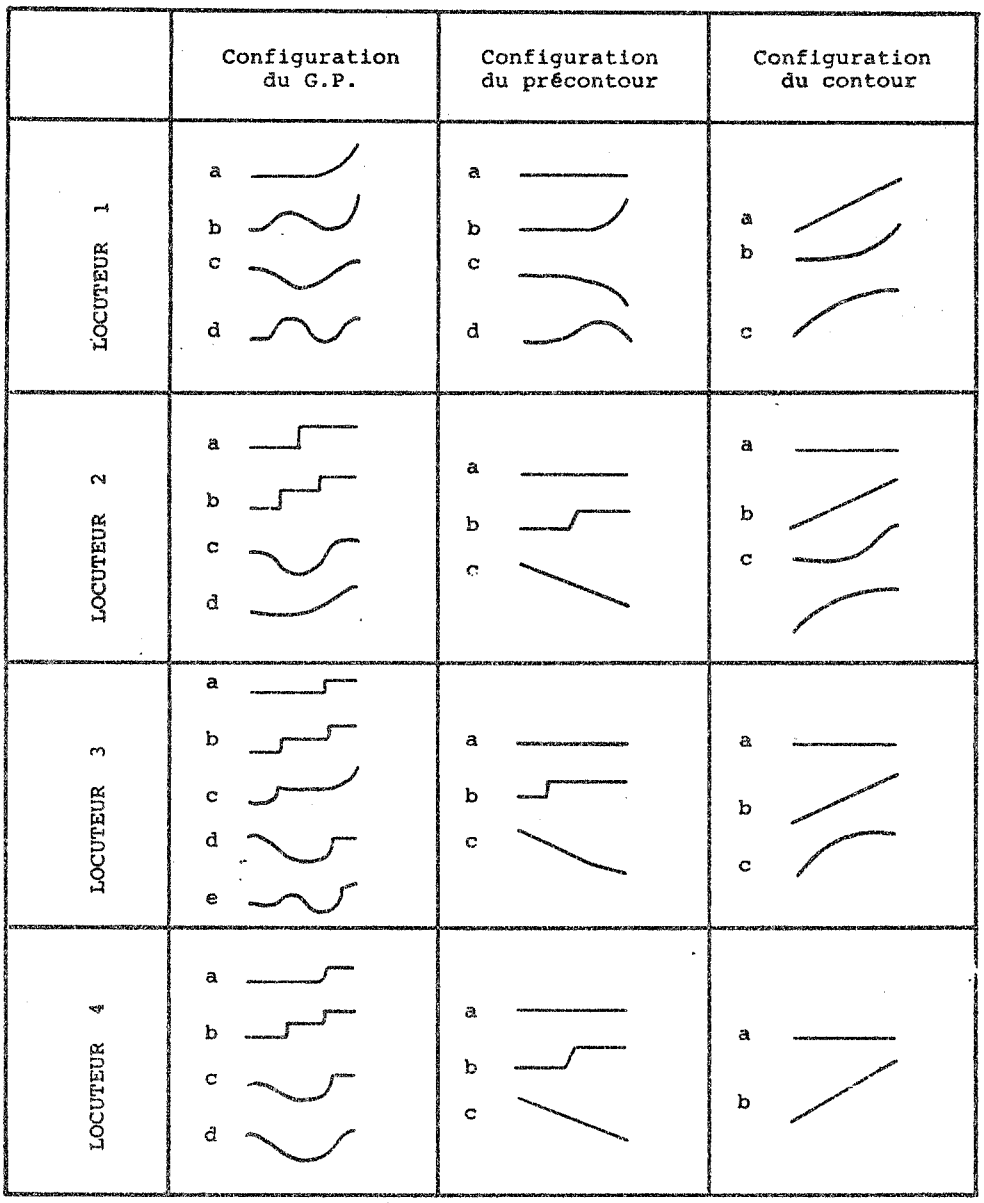


Figure 2. Configurations caractéristiques du Groupe Prosodique, du précontour et du contour des Groupes Prosodiques non terminaux.

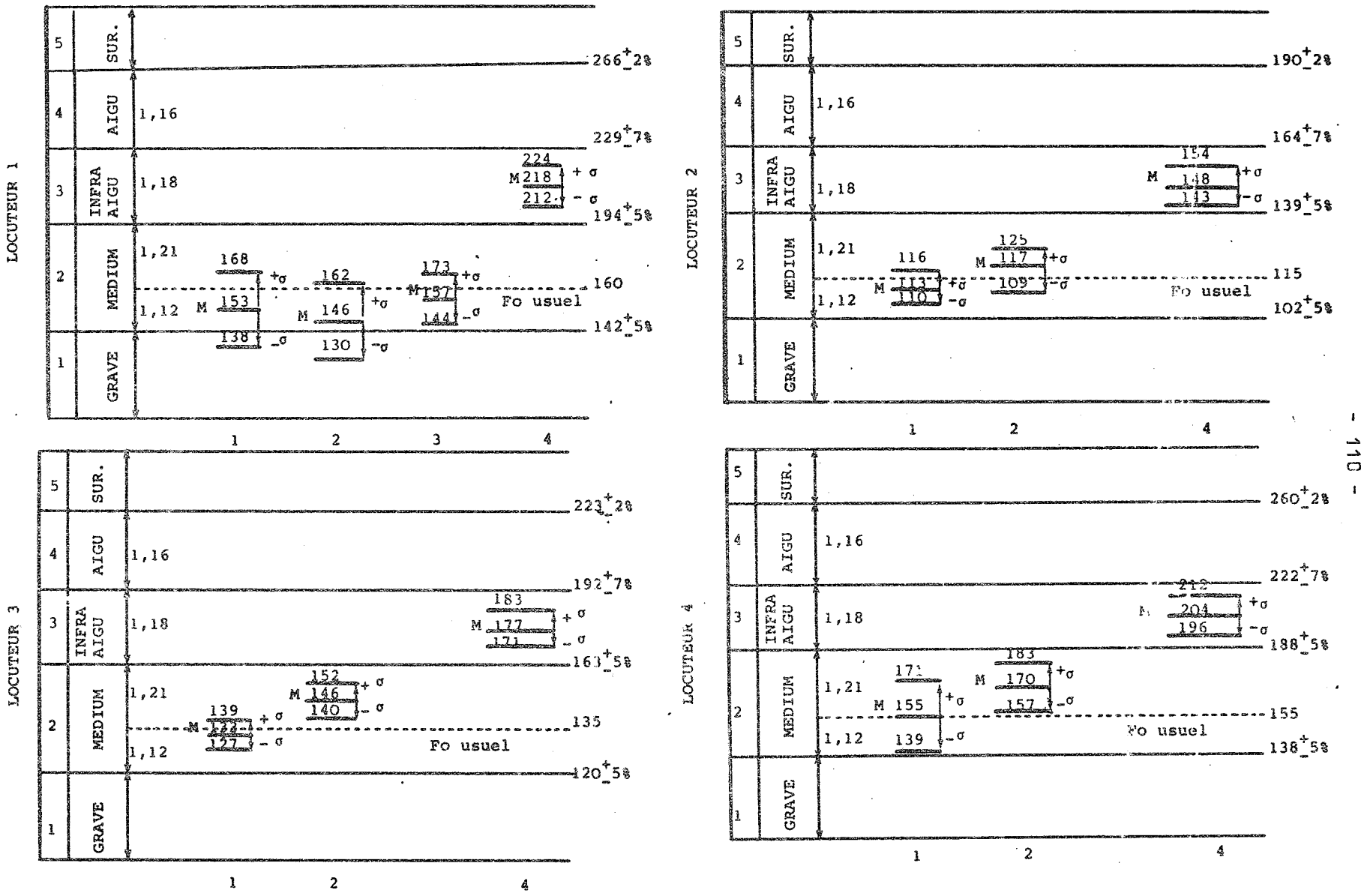


Figure 3. Valeurs perceptuelles moyennes et écarts types des points : attaque(1), prétonique(2), début tonique(3), fin tonique(4). L'absence de (3) indique que la syllabe tonique est réalisée statique.

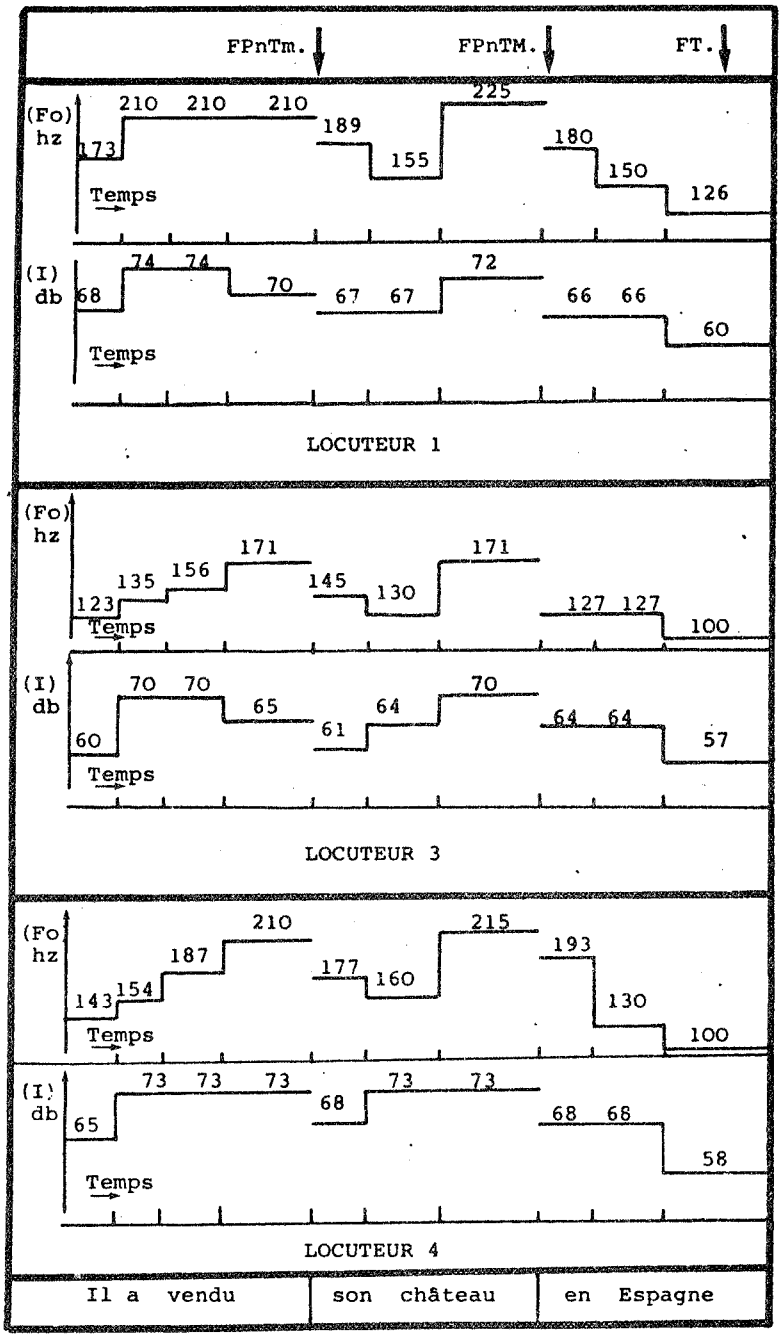


Figure 4. Schématisation des paramètres perceptuels des FPnTm.,FPntM. et F.T. dans la phrase:"Il a vendu/,/ son château/,,/ en Espagne/./ (Locuteurs 1,3,4).

A

	R 2 Fin Tonique Prétonique	R 3 Fin Tonique Début Tonique	R 4 Fin Tonique Fond. Usuel.
LOCUTEUR1	1 46±0,06 1,52 / 1,40 ~ 1 quinte	1,41±0,12 1,53 / 1,29 ~1 quinte diminuée	1,37±0,04 1,41 / 1,33 ~ 1 quarte
LOCUTEUR2	1,22±0,06 1,28 / 1,16 ~1 tierce mineure	1	1,28±0,04 1,32 / 1,24 ~1 tierce majeure
LOCUTEUR3	1,23±0,04 1,27 / 1,19 ~1 tierce majeure	1	1,33±0,04 1,37 / 1,29 1 quarte
LOCUTEUR 4	1,21±0,09 1,30 / 1,12 ~1 tierce mineure	1	1,31±0,05 1,36 / 1,26 ~ 1 quarte

Tableau I - A) Valeurs Moyennes et écarts types des rapports R2, R3 et R4 des G. P. non terminaux (continuités).

B

	ATTAQUE	PRETONIQUE	TONIQUE	R 5 Prétonique Fin Tonique	R 6 Fo Usuel Fin Tonique
LOCUTEUR1	182±24 206 / 158	145 ± 5 150 / 140	126 ± 3 129 / 123	1,15±0,03 1,18 / 1,12 ~1 seconde augmentée	1,26±0,03 1,29 / 1,23 ~1 tierce majeure
LOCUTEUR2	136 ± 12 148 / 124	112 ± 7 119 / 105	90 ± 3 93 / 87	1,23±0,05 1,28 / 1,18 ~1 tierce majeure	1,28±0,04 1,32 / 1,24 ~1 tierce majeure
LOCUTEUR3	135 ± 15 150 / 120	118 ± 6 124 / 112	98 ± 3 101 / 95	1,21±0,04 1,25 / 1,17 1 tierce mineure	1,39±0,03 1,42 / 1,36 ~ 1 quarte augmentée
LOCUTEUR 4	169 ± 18 187 / 151	131 ± 12 143 / 119	100 ± 10 110 / 90	1,33±0,07 1,40 / 1,26 1 quarte	1,57±0,14 1,71 / 1,43 1 quinte augmentée

Tableau I - B) Valeurs Moyennes et écarts types de l'attaque, de la prétonique et de la tonique ainsi que des rapports R 5 et R 6 des G. P. Terminaux (Finalités)

Références Bibliographiques

BERMAN A. and SZAMOSI M. (1972) Observations on sentential stress, Language, 48,304-25.

BIERWISCH M. (1966) Regeln für Intonation Deutscher Sätze, Studia Grammatica, 7, 99-101.

BOE L.J. (1972) Etude de l'intéraction source laryngienne-conduit vocal dans la détermination des caractéristiques intrinsèques des voyelles orales du français (fréquence laryngienne), Bulletin de l'Institut de Phonétique de Grenoble, 1, 25-43.

BRESNAN J.W. (1972) Stress and Syntax : a reply, Langage, 48,326-42.

COHEN A. and't HART J. (1967) On the anatomy of intonation, Lingua, 19, 177-92.

CRUTTENDEN A. (1970) On the so-called grammatical function of intonation, Phonetica, 21, (3), 182-92.

CRYSTAL D. (1969) Prosodic Systems and Intonation in English, Cambridge University Press.

DANES F. (1960) Sentence intonation from a functional point of view, Word, 16 (1), 34-54.

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

DELATTRE P. (1969) L'intonation par les oppositions, Le Français dans le Monde, 64, 6-12.

DOWNING B.T. (1970) Syntactic Structure and Phonological Phrasing in English, Dissertation, University of Texas.

DUBOIS J. (1969) Grammaire Structurale du Français : La Phrase et les Transformations, Larousse.

FAURE G. (1961) L'intonation et l'identification des mots dans la chaîne parlée, Proceedings of the 4th Intern. Congress Phon. Mouton, 598-609.

FAURE G. (1969) Contribution à l'étude des apports du système prosodique à la structuration de l'énoncé en français moderne, Proceedings of the Xth Intern. Congress of Linguists, Ed. de l'Académie de la République Socialiste de Roumanie, II, 1079-90.

FAURE G. (1970) Contribution à l'étude du statut phonologique des structures prosodématiques, Prosodic Feature Analysis, Studia Phonética 3, Didier,93-107.

FREI M. (1968) Signes intonationnels de mise en relief, Festschrift Walther von Wartburg zum 80 Geburtstag, I, 611-16.

GROSS M. (1974) Sur la place de l'intonation dans une grammaire transformationnelle, 5es Journées d'Etude du Groupe "Communication Parlée" II, 1-9.

HALLIDAY M.A.K. (1961) Categories of the Theory of Grammar, Word, 17, 241-92.

HALLIDAY M.A.K. (1963) Intonation in english grammar, Transactions of the Philological Society, 143-69.

HALLIDAY M.A.K. (1967) Intonation and Grammar in Britsh English, Mouton, The Hague.

HIRST D. (1974) La Levée de l'Ambiguïté Syntaxique par les Traits Intonatifs, Thèse de 3e Cycle, Université d'Aix.

HIRST D. and GINESY M. (1974) An approach to the integration of intonation in the syntactic description of english, Linguistics, 121, 45-55.

HOUSE A.S. and FAIRBANKS G. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels, J. A. S. A. 25, 105-13.

KIM K. (1968) Fo variations according to consonantal environments, Monthly Internal Memorandum, Phonoly Laboratory, Univers. Calf., Berkeley, september, 33-43.

LAKOFF G. (1972) The global nature of the nuclear stress rule, Language, 48, 285-303.

LARREUR D. et BOE L.J. (1973) Les caractéristiques intrinsèques des consonnes voisées du français dans la parole continue, Bulletin de Phonétique de Grenoble, II, 25-29.

LEHISTE I. and PETERSON G.E. (1961) Some basic considerations in the analysis of intonation, J. A. S. A., 33 (4), 419-25.

LEON P. et MARIN Ph. (1970) Prolégomènes à l'Etude des Structures Intonatives, Studia Phonetica, Didier.

LIEBERMAN PH. (1965) On the acoustic basis of the perception of intonation by Linguists, Word, 21,40-54.

LIEBERMAN Ph. (1967) Intonation, Perception and Language, Research Monograph n° 38, M. I. T. Press, Cambridge, Mass.

MAEDA S. (1974) A characterization of fundamantal frequency contours of speech,Quarterly Progress Report, 114, July, M. I. T.

MOHR B. (1971) Intrinsic variations in the speech signal, Phonetica, 23, 65-93.

NASH R. (1970) John likes Mary more than Bill. An experiment in disambiguation using synthezided intonation contours, Phonetica, 22 (3), 170-88.

POPE E. (1971) Answers to yes-no questions, Linguistic Inquiry, 2, 69-82.

RIVARA R. (1973) Pour une description intégrée de l'intonation, Linguistics, 117, 59-76.

ROSSI M. (1971a) Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole, Phonetica, 29, 1-33.

ROSSI M. (1971b) L'intensité spécifique des voyelles, Phonetica, 24, 129-61.

ROSSI M. (1972) La perception de la durée et ses implications phonétiques, Travaux de l'Institut de Phonétique d'Aix, 1, 151-64.

ROSSI M. et CHAFCOULOFF M. (1972a) Les niveaux intonatifs, Travaux de l'Institut de Phonétique d'Aix, 1, 167-76.

ROSSI M. et CHAFCOULOFF M. (1972b) Recherche sur le seuil différentiel de fréquence fondamentale dans la parole, Travaux de l'Institut de Phonétique d'Aix, 1, 179-85.

ROSSI M. (1973) L'intonation prédicative dans les phrases transformées par permutation, Linguistics, 103, 64-94.

SIERTSEMA B. (1962) Timbre, pitch and intonation, Lingua, 11, 388-98.

STOCKWELL R.P. (1972) The role of intonation : reconsiderations and others considerations, in Intonation, BOLINGER D.L. ed., Penguin Books, 87-109.

TRAGER G.L. and SMITH H.L. (1951) An Outline of English Structure, Studies in Linguistics n° 3, Battenburg, Norman, Okla.

VAISSIERE J. (1971) Contribution à la Synthèse par Règles du Français, Thèse de 3e Cycle, Université de Grenoble.

VAISSIERE J. (1974) On french prosody, Quarterly Progress Report, M. I. T., n° 114, July, 212-23.

WANG S.Y. (1972) The many uses of Fo, Papers in Linguistics and Phonetics to the Memory of Pierre DELATTRE, Mouton, 487-503.

WENDAHL R.W. (1967) Glottal wawe periods in V. V. C. environments, J. A. S. A., 42, p. 1208.

YORIO C.A. (1973) The generative process of intonation, Linguistics, 97, 111-25.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UTILISATIONS DE L'INFORMATION PROSODIQUE EN SEGMENTATION DE LA PAROLE CONTINUE.*

L. BUISSON, G. MERCIER, Centre National d'Etudes des Télécommunications, Lannion

RESUME En vue d'applications à la segmentation de la parole continue, on recherche des paramètres permettant d'accéder de façon simple à l'information prosodique. On choisit les maxima de la fréquence fondamentale et l'énergie totale des voyelles pour une étude préliminaire d'un corpus d'énoncés utiles dans le dialogue avec une machine.

SUMMARY

The question is raised whether it is possible to use prosodic features to segment continuous speech. Two parameters : namely, fundamental frequency maxima and vowels energy are used to study a set of utterances useful in man-machine communication.

* Recherche effectuée dans le cadre du contrat SESORI n° 74-80.

Utilisation de l'information prosodique en segmentation de la parole continue

L. BUISSON, G. MERCIER, Centre National d'Etudes des Télécommunications, Lannion

Dans un système de reconnaissance automatique de la parole, la reconnaissance phonétique traduit généralement l'énoncé à reconnaitre en une chaine de phonèmes avec ou non possibilité de réponses multiples. Ceci constitue évidemment une approximation car, dans la parole continue, tous les phonèmes ne sont pas sur le même plan. Intensité, durée et fréquence fondamentale varient au cours de l'énoncé. Ces trois paramètres traduisent une information de niveau supérieur à celle des phonèmes : syntaxe et sémantique, même s'ils sont aussi relatifs à la nature des phonèmes eux-mêmes.

Pour passer de la reconnaissance phonétique au sens exprimé par l'énoncé, se pose donc le problème d'utiliser cette information prosodique. Les travaux de W.A.LEA [1,2] montrent qu'on peut utiliser cette information - en anglais - avant même d'avoir fait la reconnaissance phonétique. Nous nous situerons cependant dans l'optique où les phonèmes ont été segmentés et identifiés au préalable.

I - Les Paramètres

a) Le Fondamental

Il semble bien qu'en français, la mélodie soit utilisée par le locuteur pour rendre plus aisée la segmentation de la phrase en syntagmes (groupes de sens) et même en mots lexicaux [3,4]

C'est donc le paramètre qu'il convient d'étudier en premier. Le vocodeur à canaux muni d'un détecteur de pitch [5] donne des résultats satisfaisants qui ne sont pas cependant totalement exempts d'erreurs. L'application envisagée permet d'effectuer des corrections ne laissant subsister que l'enveloppe des variations de la fréquence fondamentale :

- élimination de quelques valeurs grossièrement erronées (doublement, division par 2, transitions entre sons voisés et non voisés)

- lissage éliminant des variations non significatives pour la segmentation.

Le problème le plus difficile est la restitution automatique d'un contour global de mélodie pour l'énoncé, malgré les sons

non voisés. Pour l'instant ce problème n'a pas été abordé et on a étudié visuellement les courbes de mélodie.

b) Intensité et durée

Après segmentation automatique de la parole continue, on peut attribuer à chaque voyelle son énergie totale. Ce facteur réunit à la fois intensité et durée du segment vocalique. Il donne, à la reconnaissance, le caractère de proéminence à certaines syllabes. Contrairement au premier paramètre nous l'avons pris à la sortie du programme de segmentation et identification des phonèmes. Il possède un défaut génant c'est celui d'être sensible aux erreurs de segmentation. Dans notre étude, nous avons corrigé ces erreurs - d'ailleurs peu nombreuses - à la main mais dans un système complet cela ne pourrait se faire que par un retour d'information venant du niveau lexical.

II - Etude d'un ensemble d'énoncés

L'étude préliminaire a porté sur 124 "énoncés" servant soit à décrire des circuits logiques en conception assistée par ordinateur soit à dialoguer avec un système de documentation automatique. Cet ensemble est hétérogène et comprend :

- des énoncés qui ne sont pas des phrases naturelles en français mais qui ressemblent à des lectures de fragments d'instructions de programmation. Ex "au pas de un".

- des énoncés laconiques. Ex "fiche suivante".

- des phrases affirmatives. Ex "je voudrais analyser la sélection".

En ce qui concerne la mélodie, deux aspects ont été étudiés :

a) La chute de la fréquence fondamentale précédant une pause indique la fin d'une phrase. Or on sait que la détermination de la fin d'un énoncé n'est pas un problème trivial dans une salle bruyante (cf. par ex. 6). Sur notre corpus, (pas de phrases interrogatives), les valeurs atteintes par la fréquence en fin d'énoncé sont très inférieures à toutes les valeurs en cours d'énoncé et ceci donne donc une première segmentation qui devrait se faire pratiquement avec certitude par application d'un critère sur la fréquence.

b) A l'intérieur des énoncés l'utilisation du contour global se heurte à deux difficultés. La première - déjà signalée- est celle des intervalles non voisés et la seconde est relative aux problèmes généraux de la reconnaissances des formes. On a donc essayé dans un premier temps de se baser uniquement sur les propriétés locales de la courbe. A cet effet on a recherché les maxima de la fréquence. Il ne semble pas y avoir de problème de détection car tous les cas

pouvant entrainer des erreurs sont supprimés par le prétraitement dont nous avons parlé.

Pour chacun des énoncés du corpus on a également marqué les syllabes accentuées (*) par le deuxième facteur : intensité-durée.

Sans pouvoir pour l'instant donner de statistiques pour un corpus trop peu étendu on peut dire que dans le cas général les pics de fréquence marquent soit un début de mot soit une fin de groupe de sens (Fig. (a) - (d)). On remarque d'ailleurs que la notion de groupe de sens est liée avec les caractéristiques phonatoires de la phrase : par exemple le nombre de syllabes (Fig. (b) et (c)). Dans le cas général également, l'accent d'intensité-durée est bien corrélé avec les fins de mots (Fig. (a) - (d)).

Il faut cependant noter la présence de plusieurs catégories d'exception.

a) L'accent est parfois absent là où on pourrait le prédire d'après les groupes de sens. Cette catégorie d'exceptions n'est pas gênante si on envisage une stratégie guidée par la prosodie puisqu'elle entraine seulement qu'on ne peut tirer profit d'un niveau inférieur de l'arbre syntaxico-sémantique.

b) Plus gênante est la deuxième catégorie : détection d'accents en surnombre par rapport à ce que l'on peut prédire. On peut essayer d'expliquer ces différents cas d'exception. D'une part, certains semblent liés au caractère non naturel de quelques énoncés. Le locuteur semble essayer de recréer une structure par exemple en accentuant chaque syllabe d'un mot qui ne pourrait avoir le même type d'occurence dans une phrase française (Fig. (e) et (f)). D'autre part la connaissance du travail à accomplir et de la signification d'un énoncé dans l'étape de travail peut conduire à détacher exagérément certaines parties (Fig. (g) - (h)). (Voir même remarque pour l'anglais dans [7]). Enfin certains cas sont difficilement explicables (Fig (i)).

* Pour faire court, nous parlons d'accent au sujet des syllabes marquées par nos deux paramètres. Cela n'implique pas, de notre part, une prise de position sur la nature des liens avec l'accent au sens strict en français.

III - Limites d'étude et extensions possibles

Pour l'étude de la fréquence fondamentale, on n'a pas étudié l'aspect des niveaux. Ils nous ont paru moins significatifs mais cela demanderait à être vérifié. Même du point de vue de la forme de la courbe, nous avons déjà dit que l'étude devrait être étendue à son caractère global et qu'elle devrait être automatisée.

Les premiers résultats permettent cependant de poser certaines questions sur l'utilisation future de la prosodie en reconnaissance automatique. La détection de la fin de l'énoncé semble le seul point où on peut s'appuyer sans crainte sur la mélodie. Par contre on peut se demander quelle serait l'efficacité de systèmes guidés par la prosodie. Les exceptions semblent suffisamment nombreuses pour entrainer un grand nombre d'analyses inutiles si tout était fondé sur le caractère prédictif de la prosodie. Il semble au contraire que le rôle de la prosodie soit plus dans une vérification d'hypothèses construites avec des informations de niveau supérieur, y compris le niveau "pragmatique de l'application".

non, carte répétition (a)

je voudrais poser une question (b)

je voudrais voir un dictionnaire (c)

je désire donner le format d'édition (d)

entrée soixante dix huit (e)

poser une question (f)

non, fin de la description des modules (g)

je veux changer de procédure (h)

je voudrais explorer le fichier (i)

Fig. Exemples d'accentuation d'énoncés :

/ détectée par les maxima de la fréquence

/ détectée par le paramètre intensité-durée

REFERENCES :

(1) W.A.LEA : An approach to Syntactic Recognition without Phonemics, IEEE Tr. on Audio, vol.AU-21, Nr 3, June 1973.

(2) W.A.LEA et al. : A Prosodically-guided Speech Understanding Strategy, IEEE Tr. on Acoustics, vol. ASSP-23 Nr 1, February 1975.

(3) J.VAISSIERE : On French Prosody, M.I.T. Quaterly Progress Report Nr. 114, XVI Speech Communication, pp.212-223.

(4) J.VAISSIERE : Fréquence Fondamentale des phrases déclaratives en Français, 5ème Journées d'Etudes du Groupe "Communication Parlée", Orsay, Mai 1974.

(5) J.ZURCHER : Dispositif de Détection et de Mesure du Fondamental de la Parole Humaine, Note Technique TMA/ETA/18.

(6) L.R.RABINER et al. : An Algorithm for determining the end points of isolated utterances, Bell System Technical Journal, vol.54,n°2, Feb. 1975.

(7) T.C. DILLER : Prosodies appearing in the SDC Vocal Data Management Dialogues, July 1973.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

A PROPOS DE MARQUEURS LEXICOSYNTAXIQUES : Quelques exemples commentés de phrases issues de l'analyseur du projet A.P.I.A

D. DOURS* - R. FACCA* - Y. LAURENTIE** - G. MAURAND*** - G. PERENNOU****

RESUME L'analyse d'un phrase, basée uniquement sur les éléments phonétiques et le lexique, est extrêmement lourde. Lorsque certains mots ou groupes de mots peuvent être isolés, cette analyse s'allège considérablement. Elle peut encore s'alléger lorsqu'il est possible d'affecter un sens ou une fonction aux éléments isolés. A partir de données fournies par l'analyseur du projet A.R.I.A. |1|, on examine trois types de marqueurs (ou délimiteurs) fondés sur :

- Elévation du niveau énergétique des voyelles
- Discontinuité de la mélodie
- Apport des configurations consonantiques

Les durées jouent également un rôle, leur influence n'est pas examinée ici.

SUMMARY The analysis of a sentence based only on phonetical and lexical elements is extremely awkard. When certain words or groups of words can be isolated this analysis is easier. It becomes even easier when it is possible to assign a meaning or a function to the isolated elements. From data given by the A.R.I.A. project analyser we consider three types of markers :

- Elevation of the energetic level of the vowels
- Discontinuity of the melody
- Contribution of the consonantic configurations.

Durations are also significant, but their influence is not examined here.

* Assistant - Université Paul Sabatier Toulouse
** Chercheur - Laboratoire C.E.R.F.I.A. Toulouse
*** Professeur - Université du Mirail Toulouse
**** Professeur - Université Paul Sabatier Toulouse

A PROPOS DE MARQUEURS LEXICOSYNTAXIQUES : Quelques exemples commentés de phrases issues de l'analyseur du projet A.R.I.A.

D. DOURS - R. FACCA - Y. LAURENTIE - G. MAURAND - G. PERENNOU

1. INTRODUCTION

Dans beaucoup de systèmes de reconnaissance de la parole, la segmentation en mots est supposée résolue par les contraintes de prononciation. Dans ce cas, l'approche analytique consiste à découper chaque mot en une suite de phonèmes ou autres constituants phonétiques, puis d'identifier cette chaîne à un terme du lexique.

Lorsqu'on s'intéresse à la reconnaissance de la parole continue, deux difficultés apparaissent : d'une part on ne peut s'appuyer sur un prédécoupage en mots ; d'autre part, il est bien connu que lorsqu'ils sont insérés dans la chaîne phrastique les mots s'influencent les uns les autres.

Il est néanmoins envisageable de reconstituer le message vocal, en partant d'une séquence phonétique et en procédant comme suit :

1°) Dans un fichier $\mathcal{P}$ se trouvent répertoriées les représentations phonétiques de chaque mot du lexique, y compris celles qu'introduisent les liaisons et les différentes prononciations.

2°) Le signal est décomposé en une suite de segments contigus, identifiés chacun à un symbole phonétique (comme c'est le cas dans le module de reconnaissance du projet A.R.I.A.)

3°) A la lecture du 1er symbole phonétique, une consultation du fichier $\mathcal{P}$ permettra de déterminer s'il correspond à un mot, un début de mot ou encore les deux à la fois. Un quatrième cas doit être envisagé, ce n'est ni un mot ni un début de mot.

L'analyse se poursuit en lisant les symboles phonétiques les uns après les autres. A chaque lecture, il faut distinguer les quatre possibilités suivantes : un nouveau mot vient de se compléter, un nou-

veau mot se complète tout en étant le début d'un autre mot, un début de mot apparaît seul, enfin apparaît une chaîne lexicalement impossible. On conçoit qu'en procédant ainsi, après la lecture de plusieurs symboles phonétiques, il faut conserver présente la possibilité de plusieurs découpages de la phrase en une succession de mots.

Plus précisément, on peut formaliser cette analyse de la manière suivante :

A un stade donné de la lecture, se trouvent mémorisés les couples (ℓ_i, d_i) $i = 1,2,...$ où ℓ_i est une suite de mots du lexique et d_i est un début de mot.

Soit x le nouveau symbole phonétique lu. Chacun des couples est alors modifié comme suit :

- Si $d_i x$ est un début de mot mais non un mot, alors :

$$(\ell_i, d_i) \rightarrow (\ell_i, d_i x)$$

- Si $d_i x$ est un mot mais non un début de mot, alors :

$$(\ell_i, d_i) \rightarrow (\ell_i', e)$$

où ℓ_i' est la liste ℓ_i augmentée du mot $d_i x$ et "e" est le "début vide".

- Si $d_i x$ n'est ni un mot ni un début de mot, le couple (ou hypothèse) (ℓ_i, d_i) est annulé.

- Si $d_i x$ est à la fois un début et un mot, on dédouble (ℓ_i, d_i) en :

$$(\ell_i, d_i x) \text{ et } (\ell_i', e)$$

Pour alléger l'analyse, il faut mettre en place des procédures permettant d'abandonner des couples (ℓ_i, d_i) qui sont peu crédibles du point de vue prosodique. Il faut également diminuer autant que possible les accès au fichier lexical.

Dans cette optique, nous allons examiner successivement quelques possibilités de simplification.

2. MARQUEURS LEXICOSYNTAXIQUES

Partant d'une chaîne de phonèmes, notre but est de reconstituer des combinaisons de phonèmes de diverses longueurs (syllabe, groupe de syllabes, mots ou syntagmes). Ceci peut être fait à l'aide de marqueurs prosodiques ou bien à partir des configurations consonantiques.

2.1. Marqueurs prosodiques

Nous n'envisagerons ici que les manifestations purement physiques des prosodèmes (force, hauteur, durée).

- La force, encore appelée accent dynamique, est fonction de l'énergie articulatoire, elle est due à l'amplitude des vibrations des cordes vocales et au degré de fermeture.

- La hauteur ou accent musical, dépend des variations de fréquences du ton fondamental.

- La durée ou accent quantitatif, est une extension relative dans le temps de chaque unité, en général la syllabe.

2.1.1. Accent dynamique

A l'issue de l'analyse nous disposons d'une chaîne phonétique dans laquelle chaque phonème est identifié soit à une voyelle soit à une consonne à l'aide d'un algorithme basé sur un critère énergétique. En effet, les consonnes sont produites avec un retrécissement du conduit vocal, voire même avec une fermeture partielle ou totale de ce dernier. De ce fait, le signal correspondant à la réalisation d'une consonne est toujours plus faible que dans le cas d'une voyelle dans un contexte à court terme.

Dans le cas des sons voisés (consonnes ou voyelles), nous calculons l'énergie sur une période j du fondamental de durée T échantillons par la formule :

$$E_j = \sum_{i=1}^{T} x_i^2$$

En envisageant divers critères énergétiques :

$$E_{TOT} = \sum_{j=1}^{N} E_j$$

$$E_{MOY} = E_{TOT}/N$$

$$E_{MAX} = \max_{j} E_j \qquad j = 1,N$$

N est le nombre de périodes du fondamental correspondant à la réalisation du phonème

nous avons pu mettre en évidence certains phénomènes intéressants. Nous avons constaté en particulier une élévation du niveau énergétique des voyelles en début de mot ou de groupe de mots. La figure 1 donne les différentes segmentations obtenues à l'aide des divers critères énergétiques envisagés.

Il semble que le critère de l'énergie moyenne E_{MOY} soit le plus efficace car c'est celui qui donne la meilleure segmentation dans les exemples envisagés.

L'intérêt d'un tel critère est évident. En effet, le calcul est très rapide et la segmentation obtenue est très proche de la segmentation idéale. Seuls quelques cas particuliers ne sont pas segmentés correctement (dernier mot d'une phrase interrogative chez certains locuteurs...)

2.1.2. Accent musical

Partant de la même chaîne phonétique, nous avons également envisagé la segmentation en mots ou groupes de mots à partir des discontinuités du fondamental. La figure 2 donne les différentes segmentations obtenues. Il semble que dans ce cas, la méthode ne soit pas aussi efficace que la précédente. Ceci est dû en particulier à la difficulté de trouver un seuil de discontinuité optimal et de plus l'accent musical ne semble pas posséder les propriétés de segmentation que nous recherchons. Il est par contre intéressant d'envisager l'accent musical dans la recherche de structures syntaxiques.

Partant des études faites par G. Maurand |2|, il est possible de mettre en oeuvre un algorithme utilisant les résultats de l'analyseur du projet A.R.I.A. La méthode ainsi envisagée permet de mettre en évidence des structures syntaxiques relativement simples. Il semble

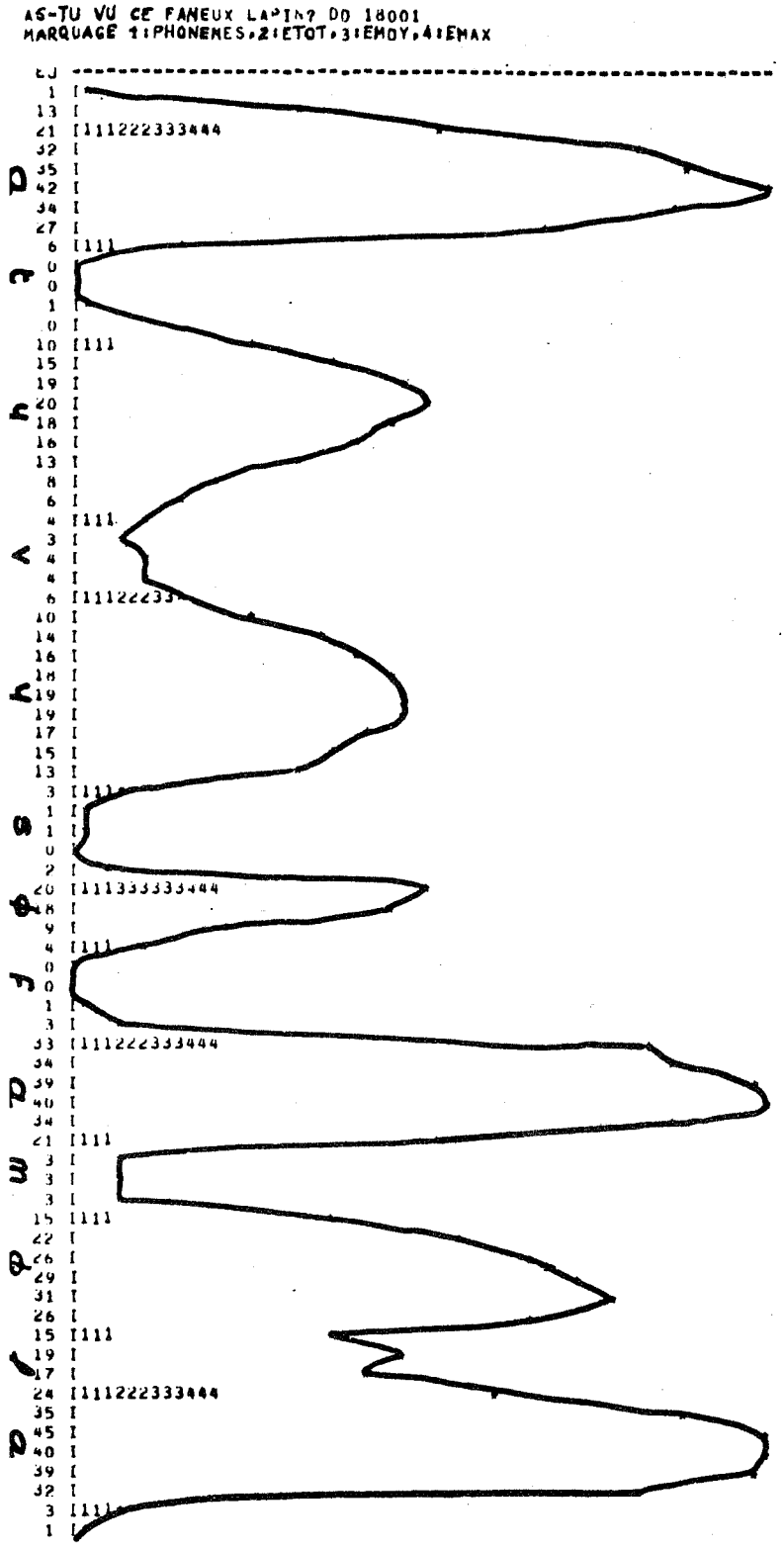


Figure 1-a

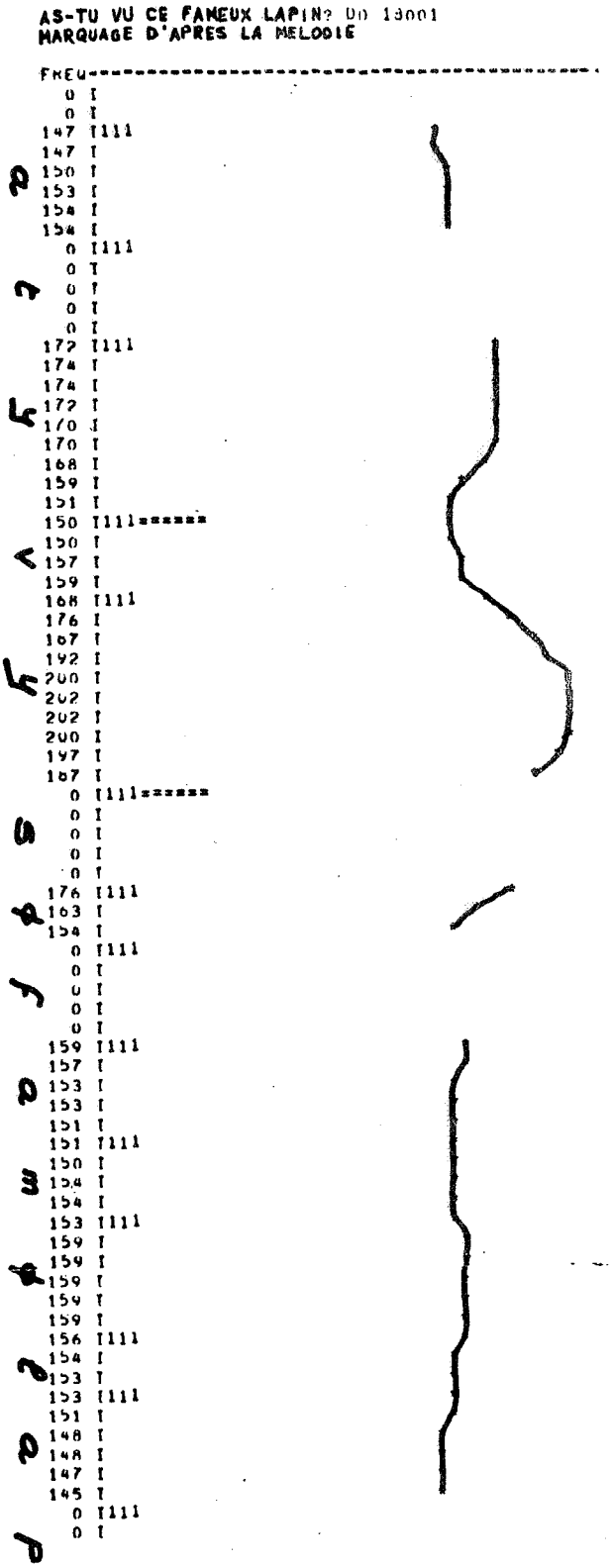


Figure 2-a

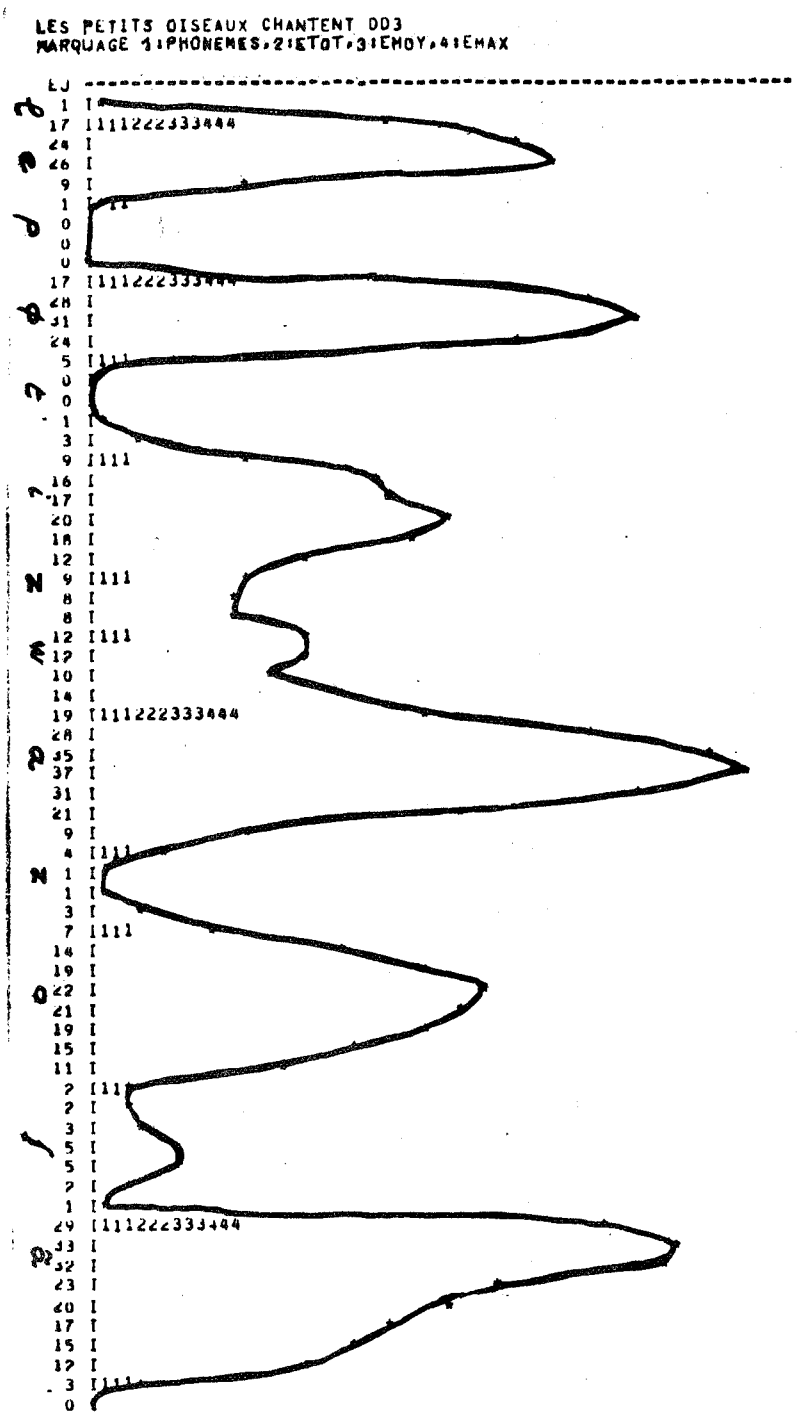


Figure 1-b

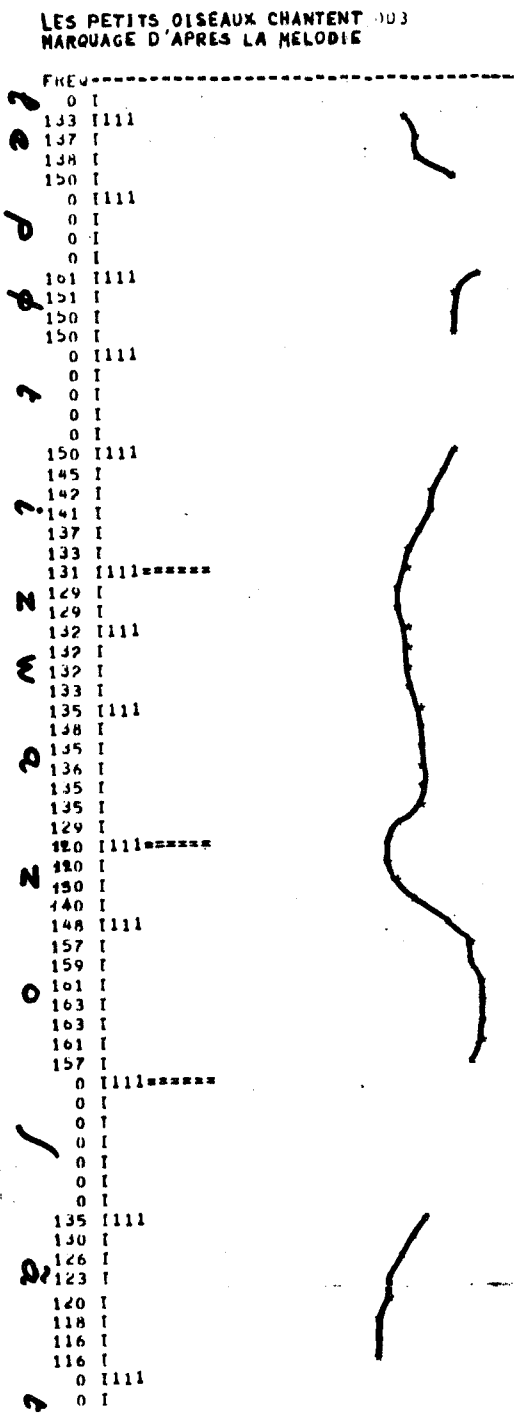


Figure 2-b

toutefois que l'apport de l'accent quantitatif rende la méthode encore plus efficace. En effet, on peut modifier le contenu d'une phrase en mettant l'accent sur tel ou tel mot. Mais l'accent dont on parle ici est ce que l'on appelle un "accent de mise en valeur" ou "d'insistance" |3|. L'accent d'insistance,est,bien sûr,fonction de l'accent musical, mais aussi de l'accent quantitatif et de l'accent dynamique.

Le critère énergétique E_{TOT}, envisagé précédemment, semble donc bien indiqué puisqu'il tient compte à la fois de l'énergie et du temps. Des études plus approfondies apporteront certainement beaucoup dans de domaine.

2.2. Apport des configurations consonantiques

Nous nous contentons d'envisager dans cette étude les groupes consonantiques à deux éléments.

2.2.1. Place de ce démarcateur dans l'ensemble des signes démarcatifs

Les éléments phoniques qui permettent de déterminer les limites du mot assurent une fonction dite démarcative. Ils peuvent être étudiés selon le classement établi par N.S. Troubetzkoy |4| : "D'abord d'après leur rapport avec la fonction distinctive, deuxièmement d'après leur caractère homogène ou complexe, troisièmement selon qu'ils indiquent l'existence ou l'absence d'une limite. Dans le cadre ainsi proposé les groupes consonantiques à deux éléments fonctionnent comme des signes phonématiques complexes, positifs ou négatifs.

2.2.2. Mécanisme de la démarcation

L'opération de commutation prouve que les dites "semi-voyelles" jouent dans la chaîne parlée un rôle consonantique ; aussi les incluons-nous dans la classe des consonnes, à la suite de la plupart des phonéticiens modernes, notamment Pierre Delattre. Toutefois la langue française n'admet pas les semi-voyelles |w,ɣ| en position implosive (finale de syllabe) ; nous ne les ferons donc pas figurer dans l'inventaire des consonnes occupant la première place dans les groupes consonantiques binaires.

Les groupes consonantiques envisagés sont susceptibles d'apparaître dans les quatre positions suivantes : 1) à l'initiale de mot,

2) en position médicale, 3) à la finale de mot, 4) à la suture de deux mots. Dans ce dernier cas, la consonne finale de mot se combine avec la consonne initiale du mot suivant.

Théoriquement toutes les combinaisons sont réalisables à la suture des mots puisque chacune des consonnes du français, exception faite des limitations particulières aux semi-voyelles, apparaît à la finale comme à l'initiale de mot ; mais elles ne le sont pas dans chacune des trois autres positions. A partir de la prononciation du français standard, nous avons dressé sur un même tableau l'inventaire des combinaisons consonantiques apparaissant dans les diverses positions du mot. Les huit possibilités théoriques de distribution, selon le nombre des positions représentées, sont attestées dans la langue. Le principe de la démarcation repose sur le fait que le nombre des combinaisons réalisées est plus grand à la suture des mots que dans les autres positions. Il reste à passer en revue, selon le classement proposé plus haut, les différents types de procédés démarcatifs.

2.2.3. Types de procédés démarcatifs

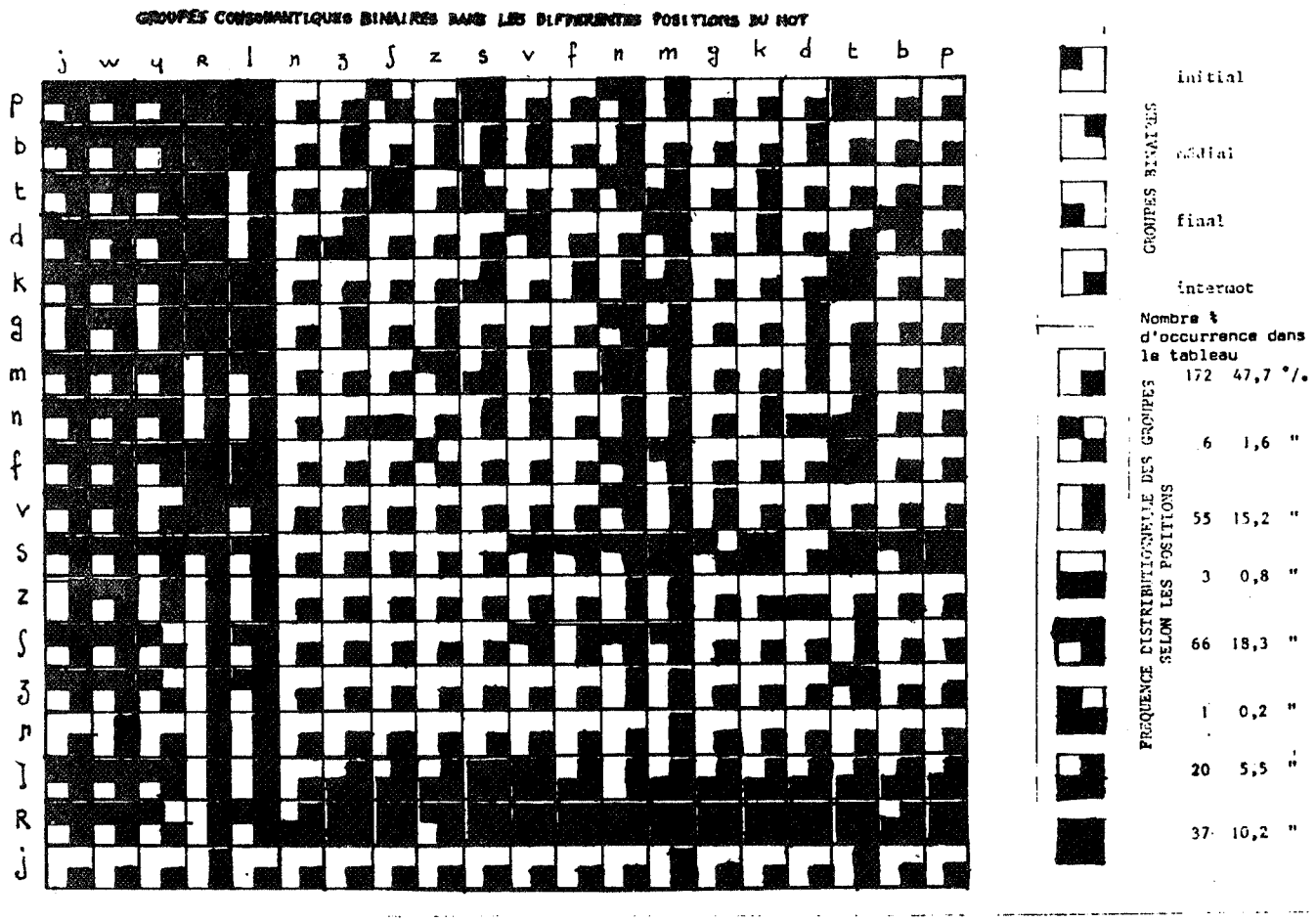
1) Les combinaisons attestées seulement en position 4 fonctionnent comme signes démarcatifs positifs en indiquant une frontière de mots entre les deux consonnes ; ainsi /pp/, dans /kap profõ/ "cap profond".
2) Les combinaisons attestées dans les positions 1,4 fonctionnent comme signes démarcatifs négatifs par rapport à la fin du mot. Il ne peut y avoir de limite de mot derrière le groupe, la frontière de mots se trouvant soit devant le groupe, soit entre les deux éléments ; ainsi /rɣ/, dans /le rɣiso/ "les ruisseaux", /par ɣit/ "par huit".
3) Les combinaisons attestées dans les positions 2 et 4 fonctionnent comme signes démarcatifs négatifs par rapport aux deux limites du mot, la frontière de mot ne pouvant se trouver ni juste avant ni juste après le groupe ; ainsi /pm/, dans /ekipmã/ "équipement", /ekip modit/ "équipe maudite".
4) Les combinaisons attestées dans les positions 3, 4 fonctionnent comme signes démarcatifs négatifs par rapport au début du mot. Il ne peut y avoir de limite de mot devant le groupe consonantique ; ainsi /nʃ/ dans pœ̃nʃ/ "punch", /bɔn ʃɛR/ "bonne chère".

5) Les combinaisons attestées dans les positions 1, 2, 4 fonctionnent comme procédés démarcatifs négatifs par rapport à la fin du mot ; ainsi /dm/ dans /dm̃ade/ "demander", /admɛtr/ "admettre", /Ryd m̃ote/ "rude montée".

6) La combinaison /ts/ attestée dans les positions 1, 3, 4 ne peut être un procédé démarcatifs négatif précis par rapport aux limites du mot : /tsigan/ "tsigane", /ɛRzats/ "ersatz", /pat sal/ "patte sale".

7) Les combinaisons attestées dans les positions 2, 3, 4 fonctionnent comme procédés démarcatifs négatifs par rapport au début du mot ; ainsi /lb/ dans /kylbyt/ "culbute".

8) Les combinaisons attestées dans les quatre positions ne fonctionnent pas comme procédés démarcatifs phonématiques ; cela n'implique pas que les réalisations soient identiques dans les différentes positions. Le problème des démarcateurs aphonématiques sera repris dans une étude ultérieure.



3. CONCLUSION

L'étude des marqueurs en tant qu'aide à la reconnaissance de la parole en continu montre déjà l'existence d'indices possédant une grande efficacité pour la segmentation en mots ou en syntagmes. Ce type d'étude mérite d'être poursuivi afin de préciser, soit d'autres indices, soit les limites d'utilisation des indices existants, soit encore les procédures prenant en compte simultanément le lexique et ces indices.

REFERENCES

|1| - "Méthode de segmentation et d'analyse par traitement direct du signal vocal
Application à la classification et à la reconnaissance des voyelles et des consonnes".
D. DOURS, R. FACCA - Thèses 3ème cycle - U.P.S. Toulouse 1974

|2| - "Contribution à l'étude du rôle syntaxique de l'intonation".
G. MAURAND - Annales de l'Université de Toulouse le Mirail. 1974

|3| - "Les domaines de la phonétique"
B. MALMBERG - PUF

|4| - "Principes de phonologie"
N.S. TROUBETZKOY - Klincksieck - p. 291

THEME 1C

ROLE DES CONTRAINTES SEMANTIQUES, PHONOLOGIQUES, CONTEXTUELLES

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UTILISATION DE CONTRAINTES PHONOLOGIQUES DANS UNE METHODE DE COMPARAISON DYNAMIQUE

C.BERGER-VACHON - G.MESNARD

Laboratoire de Physique Electronique
Université de LYON I

RESUME

On détermine par une analyse mathématique des sons élémentaires du langage, les "structures différentielles" entre les phonèmes consécutifs d'un mot. Ceci permet la constitution d'archétypes compacts qui sont comparés, par une méthode de "comparaison dynamique", avec des enregistrements de mots prononcés par plusieurs locuteurs. Les résultats numériques obtenus montrent l'intérêt de la méthode.

SUMMARY

A "differential structure" between consecutive phonemes in a word is obtained from a mathematical analysis of elementary sounds ; this leads to compact archetypes. Through a "dynamic programming method" the archetypes are compared with different utterances (several words and several speakers). The method is further analyzed by numerical computation, leading to worthwhile results.

UTILISATION DE CONTRAINTES PHONOLOGIQUES DANS UNE METHODE DE COMPARAISON DYNAMIQUE

C.BERGER-VACHON et G.MESNARD

I - INTRODUCTION

Pour améliorer les performances des premières machines ayant pour but de reconnaître la parole, on a largement utilisé les techniques générales de reconnaissance des formes. Mais la parole a ses règles propres [1] et l'emploi systématique de techniques générales [2], non adaptées au problème spécifique à résoudre, laisse de côté l'aspect physique du problème ; l'optimisation des solutions n'est pas quelque chose de trivial, même dans le cadre d'un cahier des charges très précis ; l'effort d'incorporation dans la conception d'un système de reconnaissance, des composantes venant de plusieurs domaines, doit être considéré actuellement comme une exigence minimale. C'est ce que nous avons essayé de faire en utilisant des techniques de programmation dynamique et des contraintes phonémiques dans un mot pour "marquer" les échantillons qui représentent l'évolution temps-fréquence du signal vocal.

La programmation dynamique, telle qu'elle est utilisée pour la reconnaissance de la parole donne en général des résultats meilleurs que les méthodes analytiques (cf.par exemple [3]) ; il faut définir un archétype et comparer une forme inconnue avec cette structure de référence, en imposant au paramètre temps la souplesse lui permettant de s'accommoder aux variations inhérentes au phénomène de la phonation [4] . Dans ce qui suit nous nous proposons d'analyser des sonogrammes à l'aide d'opérateurs d'assez mauvaise qualité, mais qui nous permettent quand même de repérer les phonèmes qui constituent un mot [5] .

Pour cela, nous utiliserons les contraintes de la parole. On sait que les successions de sons ne sont pas libres dans le langage naturel [6] ; en se limitant à des mots connus, on peut aller encore plus loin car les successions sont imposées. Par exemple, l'examen du chiffre SEPT montre qu'une occlusion suit obligatoirement le phonème /ɛ/. Nous tirerons parti de cette remarque dans la suite.

Les opérateurs utilisés sur le sonogramme ne retiennent que les paramètres qui semblent utiles pour identifier une suite de phonèmes (en pratique, le mot phonème a une signification plus large que celle de la théorie acoustique ; il peut par exemple regrouper un ensemble de sons élémentaires). On ne cherchera pas, avec ces opérateurs, à localiser les formants car les techniques actuelles, même les plus sophistiquées [7] sont souvent mises en échec ; on ne retiendra pas non plus le fondamental qui n'est pas essentiel dans la discrimination des sons.

Les méthodes que nous utiliserons seront d'autant plus efficaces que les suites de phonèmes sont formées d'éléments distincts. L'emploi de vocabulaires spéciaux [8] peut, dans certains cas délicats, augmenter leur efficacité.

II - LA METHODE UTILISEE

2-1 Le pré-apprentissage - Il s'agit de repérer les principales caractéristiques des phonèmes du langage.

Sur le vocodeur du CNET à Lannion, nous avons enregistré un ensemble de sons prononcés de façon aussi isolée que possible ; en effet, même dans ce cas qui semble très favorable, il n'est pas possible de localiser avec précision les phonèmes ; c'est ainsi (cf.figure 1) que toute méthode de séparation de l'énergie et du bruit peut être discutée, or la position des limites conditionne les paramètres du son.

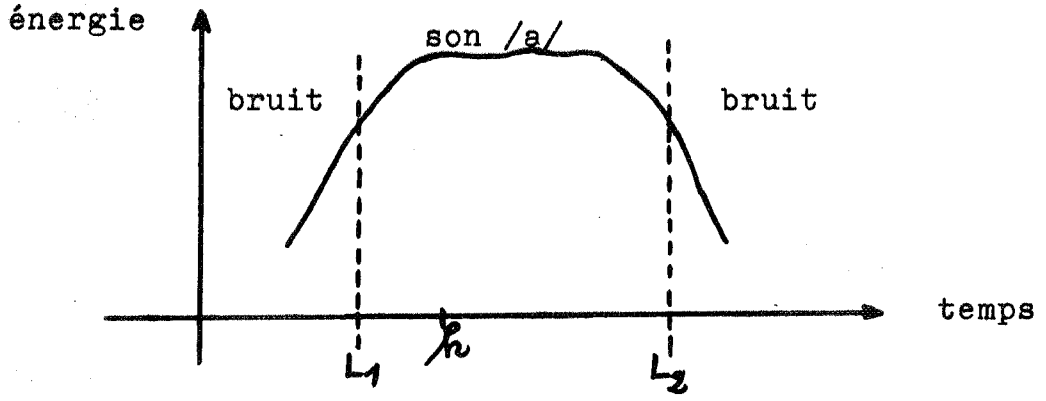


Figure 1 - Evolution de l'énergie en fonction du temps lorsque le phonème /a/ est prononcé devant le vocodeur. L_1 et L_2 représentent les limites entre le son et le bruit.

Le vocodeur fournit des échantillons toutes les 20 millisecondes ; chaque échantillon est formé de 14 valeurs qui représentent la répartition fréquentielle de l'énergie au cours des 20 millisecondes qui précèdent la lecture des registres électroniques.

Nous appellerons $\overrightarrow{X_{ik}}$ le k^e vecteur obtenu à la suite de la prononciation du son ω_i devant la machine. Les composantes de $\overrightarrow{X_{ik}}$ sont formées par les 14 répartitions fréquentielles qui correspondent au k^e échantillon : $\overrightarrow{X_{ik}} \in \mathbb{R}^{14}$.

Le son ω_i sera caractérisé par les vecteurs $\overrightarrow{V_i}$ et $\overrightarrow{W_i}$ représentant respectivement la moyenne et la variance des $\overrightarrow{X_{ik}}$ lorsque $L_1 < k < L_2$. Précisons ces vecteurs.
Soient x_{ikp}, v_{ip}, w_{ip} les p^e composantes des vecteurs $\overrightarrow{X_{ik}}$, $\overrightarrow{V_i}$ et $\overrightarrow{W_i}$. On peut écrire :

$$v_{ip} = \frac{1}{m} \sum_{k \in [L_1, L_2]} x_{ikp} \qquad (1)$$

$$w^2_{ip} = \frac{1}{m-1} \sum_{k \in [L_1, L_2]} (x_{ikp} - v_{ip})^2 \qquad (2)$$

où m représente le nombre d'échantillons correspondant au son.

Comme on dispose de plusieurs prononciations du son ω_i, on peut encore moyenner $\overrightarrow{V_i}$ et $\overrightarrow{W_i}$ sur l'ensemble de ces prononciations.

Soient maintenant deux sons ω_i et ω_j. Ils sont représentés par des couples de vecteurs $\overrightarrow{V_i}$ et $\overrightarrow{W_i}$ d'une part et $\overrightarrow{V_j}$ et $\overrightarrow{W_j}$ d'autre part.
Si on considère chacune des dimensions de $\mathbb{R}^{14}$ on peut rechercher celles qui permettent de discriminer au mieux ω_i et ω_j. On suppose

que les quantités x_{ikp} et x_{jkp} sont distribuées normalement ; on peut poser en première approximation :

$$x_{ikp} \longrightarrow N(v_{ip}, w_{ip}) \qquad (3)$$

$$x_{jkp} \longrightarrow N(v_{ip}, w_{jp}) \qquad (4)$$

N désignant la distribution normale. Représentons sur la figure 2 les distributions (3) et (4) ;

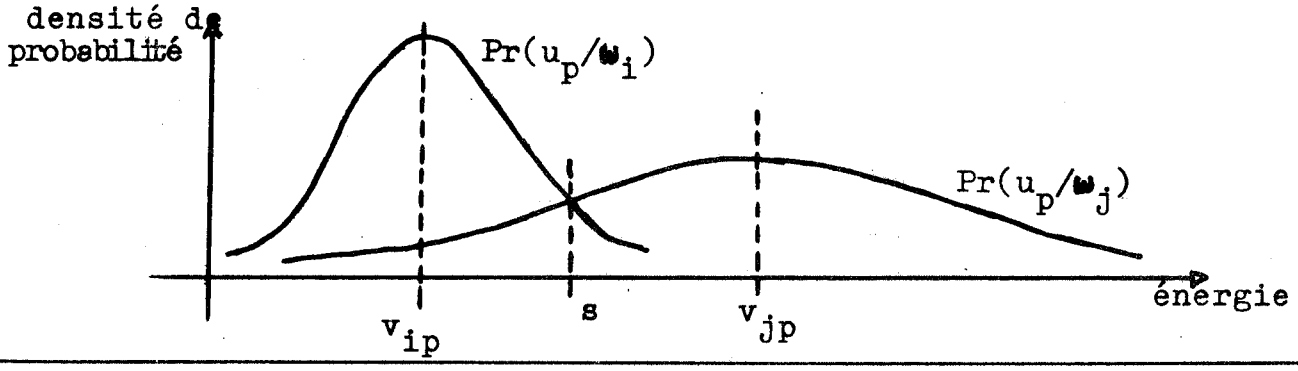


Figure 2 - Distribution de la répartition énergétique des sons ω_i et ω_j dans le p^e canal fréquentiel (auquel on associe la p^e dimension de $\mathbb{R}^{14}$). u_p représente l'énergie dans le k^e canal et s le seuil de décision théorique.

Les théories classiques du maximum de vraisemblance [9] appliquées à ce type de problème permettent d'écrire la probabilité de mauvaise classification R_{ijp} sous la forme :

$$R_{ijp} = \frac{1}{2}\left[\int_{-\infty}^{s} Pr(u_p/\omega_j)\, d\,u_p + \int_{s}^{\infty} Pr(u_p/\omega_i)\, d\,u_p\right] \qquad (5)$$

La quantité $Q_{ijp} = 1 - R_{ijp}$ (6) représente la qualité de la p^e dimension pour séparer les classes w_i et w_j. L'équation (5) suppose l'égalité des probabilités a priori des classes ω_i et ω_j, ce qui est compatible avec les analyses que nous ferons par la suite. On voit donc qu'il est possible de déterminer mathématiquement le pouvoir discriminant de chacune des dimensions de $\mathbb{R}^{14}$ pour les sons ω_i et ω_j sans recourir à des méthodes lourdes et plus difficilement exploitables [10]. Nous retiendrons donc les meilleurs canaux séparant ω_i et ω_j.

2-2 Le marquage des échantillons

2-2-1 Description d'un sonogramme - Un sonogramme est organisé selon le schéma indiqué sur la figure 3.

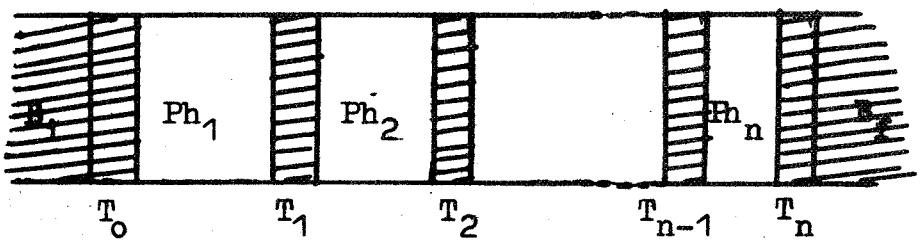


Figure 3 - Décomposition d'un mot en phonèmes. Ph_1, Ph_2, ..., Ph_n représentent les phonèmes. T_o, T_1, T_2, ..., T_n représentent les transitions. B_i et B_f sont les bruits, initial et final (non analysés).

Un mot est représenté par une suite de phonèmes séparés par des transitions. Sur ce modèle, on introduit les notions suivantes :

. le point "d'ancrage". C'est un instant de référence, facile à obtenir et fiable. On a choisi pour le déterminer la montée en énergie ; le point d'ancrage correspond au premier dépassement d'un seuil énergétique. Le seuil dépend du mot car une donnée trop rigide ne peut pas s'adapter à l'ensemble des situations recontrées sur la parole. Il est situé en général dans la transition T_1.

. le premier phonème, Ph_1, est de faible énergie : on suppose que tous les mots commencent par une zone de faible énergie, ce qui permet d'utiliser le modèle de la figure 3 dans toutes les situations. En effet, considérons le mot ZERO qui débute par un /z/. Si le /z/ est trop long, son début pourra ne pas être sur le sonogramme ; ceci est lié aux conditions de démarrage de l'enregistreur associé au vocodeur de Lannion, qui prend systématiquement les 8 échantillons qui précèdent un seuil énergétique (dans le cas de ZERO le déclenchement est dans la transition entre le /z/ et le /e/). Lorsque le mot débute par un élément de forte énergie (exemple:DEUX), le bruit est incorporé au mot et il forme le premier phonème. Par exemple, DEUX s'écrit :

$$\underbrace{\text{bruit}}_{Ph_1}\qquad\underbrace{\text{d Ø}}_{Ph_2}$$

. les phonèmes ne sont pas strictement ceux de la théorie acoustique Lorsque des phonèmes sont trop difficiles à distinguer, on les englobe en une seule unité, appelée phonème par extension. C'est par exemple le cas de /dØ/ ci-dessus.

2-2-2 Le marquage proprement dit - Le principe du marquage des échantillons est alors le suivant : on considère un enregistrement connu (par exemple ZERO). L'analyseur logique recherche le point d'ancrage (qui est situé dans la montée énergétique vers le /e/). Ensuite il examine les échantillons en fonction de la transition à venir. Prenons par exemple la transition entre Ph_2=/e/ et Ph_3=/R/ ; on utilise les canaux les meilleurs pour séparer les deux phonèmes Ph_2 et Ph_3, d'après les valeurs des coefficients Q_{23p} (équation 6)[2] ; la transition est représentée schématiquement sur la figure 4 :

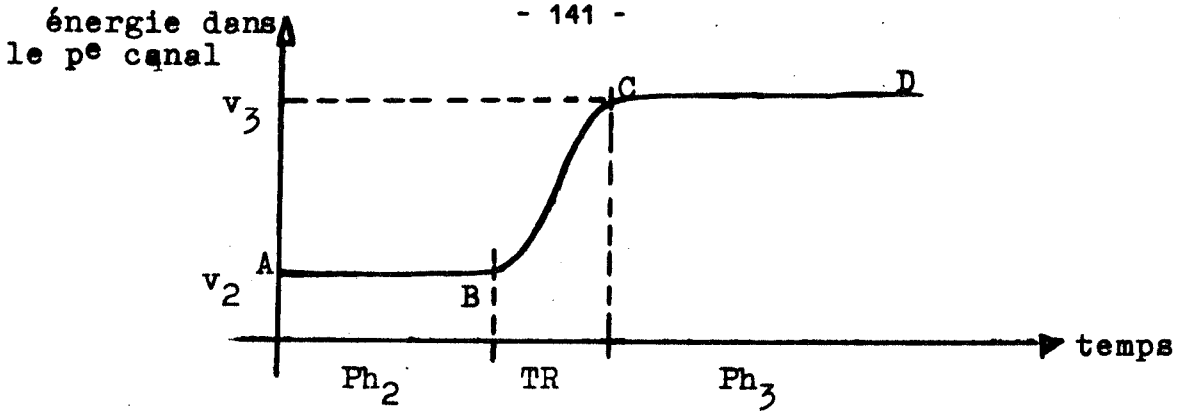


Figure 4 - Transition entre les phonèmes Ph_2 et Ph_3. En ordonnée, on représente l'énergie dans un canal.

En pratique, les segments AB et CD ne sont pas aussi rectilignes, mais l'utilisation de l'équation (6) permet de choisir les canaux qui se rapprochent le plus de ce schéma.

On introduit une distance de Hamming entre les structures "types" définies en 2-1 et les structures réelles observées au cours de la prononciation du mot :

$$D_{ix(t)} = \sum_{s=1}^{S} | x_s(t) - v_{is} | \qquad i=1,2 \qquad (7)$$

- v_{is} est la valeur type de Ph_i sur le s^e canal
- $x_s(t)$ est la valeur observée sur le s^e canal pour le signal au temps t considéré
- S est le nombre de canaux "sensibles" (ayant un coefficient Q_{ijp} élevé)
- $D_{ix(t)}$ est la distance de Hamming du signal au i^e phonème à l'instant t considéré.

On a représenté les distances D_2 et D_3 sur la figure 5.

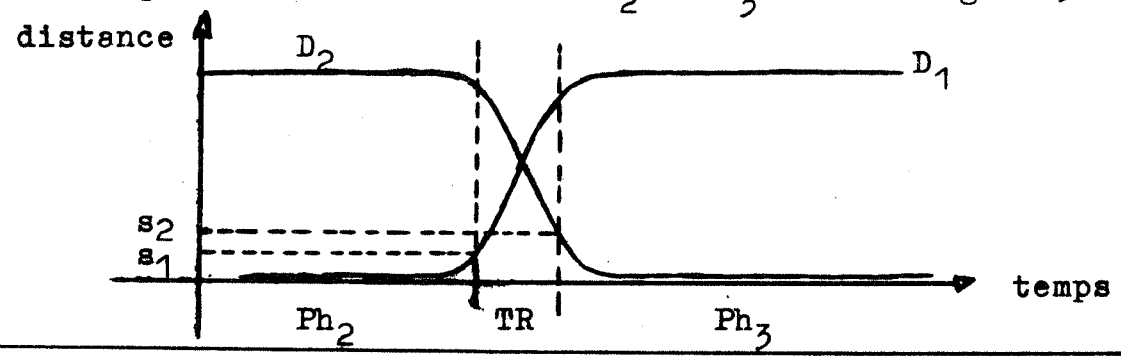


Figure 5 - Evolution des distances de Hamming en fonction du temps.

On voit que la distance D_2 est particulièrement bien adaptée pour indiquer le début de la transition et que D_3 annonce la fin de cette transition. s_1 et s_2 sont deux seuils ; lorsque D_2 est supérieur à s_1 on décide que la transition est commencée et on calcule alors D_3. Dès que D_3 devient inférieur à s_2, on conclut que Ph_3 a commencé. Il faut alors attendre le passage $Ph_3 \longrightarrow Ph_4$. La logique de cette séquence est indiquée sur la figure 6.

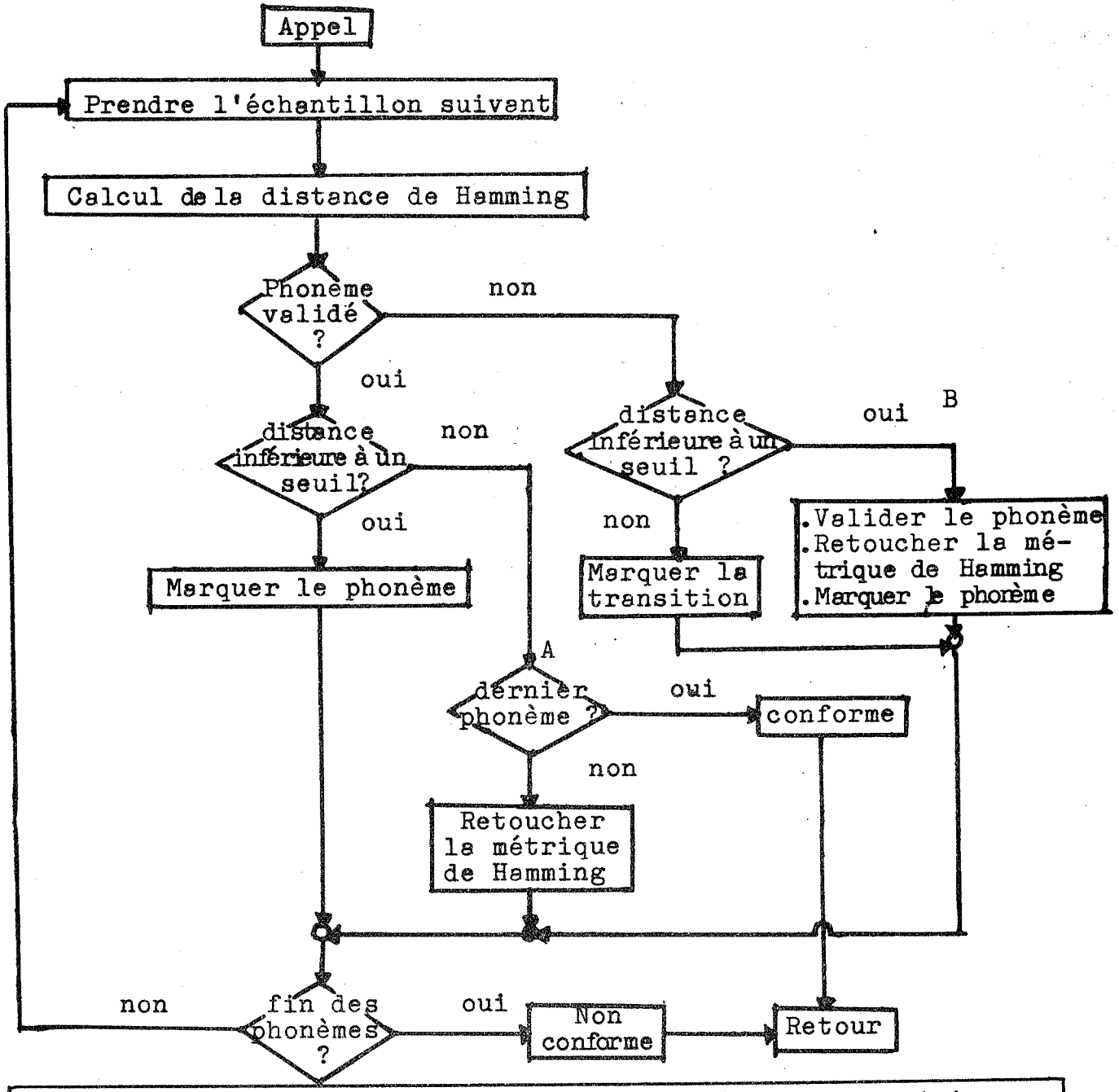


Figure 6 - Logique des changements de structures de référence lors de l'analyse des échantillons d'un enregistrement. Le programme effectuant le travail s'appelle MARQUECH.

En pratique, la machine ne calcule qu'une seule distance de Hamming par échantillon : D_2 avant la transition et D_3 pendant la transition. Après la transition, on s'intéresse au phonème Ph_3 en tenant compte du prochain passage de Ph_3 à Ph_4 ; de nouveaux canaux et de nouvelles valeurs sont concernées.

Revenons à l'organigramme de la figure 6 ; un phonème est dit "validé" lorsque l'analyseur logique n'étudie pas une transition. Le point A indique le début d'une transition ; si ce début correspond à la fin du dernier phonème, l'analyse est terminée. Cette voie de sortie est la seule normale. Si le mot n'est pas terminé, on change les valeurs de référence de v_2 à v_3 (pour les mêmes canaux) et la distance de Hamming D_3 est calculée.

Le point B correspond à la fin d'une transition. Il faut donc indiquer quels sont les canaux qui séparent Ph_i de Ph_{i+1} et les charger avec les S valeurs v_{i1}, ..., v_{is} de Ph_i (ou ω_i).

Si l'analyse se termine sans que tous les échantillons du mot de référence aient été reconnus, il y a une erreur et on l'indique.

En pratique, l'analyse du phonème 1 se fait en revenant en arrière à partir du point d'ancrage, tandis que les phonèmes Ph_2, Ph_3, ..., Ph_n sont analysés en progression normale selon un organigramme qui ressemble à celui de la figure 6. Chaque échantillon est étiqueté selon la structure représentée figure 3.

2-3 La structure des archétypes

Un mot est donc défini par ses phonèmes, eux-mêmes repérés par les canaux différentiels ou sensibles. La structure de /zéro/ est indiquée sur le tableau 7.

T_o /bruit/ → /Z/	1	2	6	7	10	14	
	5.3	4.5	3.2	3.5	3.4	3.8	14
T_1 /Z/ → /e/	1	2	4	5	6	7	
	5.3	4.5	1.5	1.5	3.2	3.5	15
	14.5	14.	7.2	5.8	6.1	7.8	10
T_2 /e/ → /R/	1	2	7	8			
	14.5	14.	7.8	10.			10
	7.	7.8	2.5	2.5			15
T_3 /R/ → /O/	1	2					
	7.	7.8					11
	14.7	14.1					2
$T_4=T_n$ /O/ → /bruit/	1	2	4	5	11		
	14.7	14.1	10.0	6.5	5.0		10

Tableau 7 - Description du mot /ZERO/. Chaque transition est décrite en indiquant d'abord les numéros des canaux sensibles. Considérons la k^e transition ; elle est entourée par les phonèmes Ph_{k-1} et Ph_k représentés par les vecteurs $\vec{V}_{k-1}$ et $\vec{V}_k$. On a indiqué sur le tableau les valeurs de V_{k-1} et V_k qui correspondent aux canaux sensibles.

Les transitions T_o et T_n sont plus réduites car l'entrée dans le bruit signale la fin de l'analyse. A droite, on a noté les seuils correspondant aux distances de Hamming. L'archétype, tel qu'il est décrit sur le tableau 7 (pour ZERO) est comparé selon le protocole indiqué figure 6 avec des enregistrements acoustiques ; l'élasticité en temps de MARQUECH est très grande.

MARQUECH a deux réponses possibles : conforme et non conforme. On pourra donc juger de la similitude entre l'archétype et les enregistrements.

2-4 Les résultats numériques

2-4-1 Les sons de même espèce - L'apprentissage a été effectué sur le locuteur numéro 1. Les résultats suivants ont été obtenus, pour des mots prononcés par ce locuteur (tableau 8) :

ZERO/ZERO	12/14
DEUX/DEUX	15/15
SEPT/SEPT	12/12
VIRGULE/VIRGULE	7/7
CONTINUE/CONTINUE	8/8
PUISSANCE/PUISSANCE	6/7
EXPONENTIEL/EXPONENTIEL	6/8

Tableau 8 - Résultats du programme MARQUECH avec le locuteur numéro 1. Le mot de gauche indique la structure et celui de droite l'enregistrement sur lequel elle a été appliquée. A droite, la fraction indique le taux de "conforme".

Le taux de reconnaissance est de 93%. En modifiant les valeurs des seuils (par exemple, pour ZERO, on adopte les seuils : 14 15 10 10 15 11 3 18), le résultat a été amené à 100%. Le 100% n'est pas systématique : si on augmente trop les seuils un phonème sera toujours reconnu et si on les abaisse trop, il ne sera jamais reconnu. Les archétypes ont été appliqués (avec les seuils qui ont donné 100% sur le locuteur numéro 1) sur des prononciations issues d'autres locuteurs. Les résultats sont résumés sur le tableau 9.

ZERO (I)	ZERO (I)	3/10
	ZERO (II)	6/10
DEUX (I)	DEUX (II)	6/10
	DEUX (III)	10/10
SEPT (I)	SEPT (II)	6/10
	SEPT (III)	6/8

Tableau 9 - Résultats obtenus, en comparant des archétypes calculés sur le locuteur numéro 1, avec des enregistrements fournis par les locuteurs II et III.

Les résultats sont assez mauvais. Nous les avons améliorés par une nouvelle modification de la ligne des seuils (14 15 15 13 12 7 8 18 pour ZERO) qui a donné :

ZERO (I)	ZERO (II)	9/10
	ZERO (III) ...	10/10

2-4-2 Les résultats différentiels - Un archétype qui permet de bien reconnaître le mot qui lui correspond peut aussi reconnaître des mots totalement différents. On a analysé les chiffres, prononcés par le locuteur numéro 1, à l'aide de la structure de ZERO (obtenue sur le locuteur numéro 1) et ne donnant du 100% que pour ce locuteur, on n'obtient qu'une seule reconnaissance .... celle du ZERO.

En analysant les chiffres, prononcés par le locuteur numéro 1, à l'aide d'une structure de SEPT, donnant du 100% sur les SEPT des locuteurs II et III, on a obtenu sept reconnaissances (0, 2, 4, 5, 7, 8, 9) et trois rejets (1, 3, 6). Mais souvent, la longueur des transitions est anormalement grande, les phonèmes étant détectés sur des transitoires. Il semble donc que l'élasticité en

temps soit trop grande pour qu'une extension de MARQUECH à de nouveaux locuteurs permette des reconnaissances avec un taux convenable.

III - GENERALISATIONS ET CONCLUSIONS

3-1 Correction automatique des structures phonémiques

Si on veut utiliser les résultats d'un premier apprentissage sur de nouveaux locuteurs, il est indispensable de recalculer les vecteurs $\vec{V}_i$ et $\vec{W}_i$ qui caractérisent un phonème. On a vu qu'un élargissement des seuils permet d'étiqueter correctement un mot connu ; il suffit donc de repérer, dans un ensemble de mots, les structures qui représentent un phonème pour obtenir de nouvelles représentations pour les sons et refaire l'étude différentielle. Les seuils seront alors fixés à des valeurs assez basses et MARQUECH pourra fonctionner de façon fiable pour le nouveau locuteur.

3-2 Segmentation automatique

MARQUECH attribue à chacun des échantillons d'un enregistrement une des trois étiquettes :

- Numéro de phonème
- Transition
- Echantillon non analysé.

Cet étiquetage correspond au schéma de la figure 3 (un échantillon n'est pas analysé s'il est avant Ph_1 ou après Ph_n).

Dans chaque transition, on peut déterminer quel est l'échantillon dont la configuration est la plus proche d'une configuration moyenne entre les structures des phonèmes qui entourent cette transition. La détermination d'une telle configuration moyenne doit faire appel à des méthodes statistiques pour tenir compte de la variabilité sur chacun des canaux du vocodeur.

3-3 Génération d'archétype

La segmentation automatique de différents enregistrements correspondants à plusieurs prononciations d'un même mot, ouvre la voie à la normalisation en longueur des différents phonèmes. On peut alors, à partir de chacun des enregistrements, obtenir une structure normalisée. La moyenne des structures normalisées conduit à un archétype standard. La reconnaissance des enregistrements, à l'aide d'une méthode de programmation dynamique un peu moins souple que la précédente et utilisant l'archétype standard peut être envisagée.

3-4 Conclusions

La méthode de reconnaissance des successions de phonèmes à l'aide des caractères différentiels a conduit à des résultats valables pour étiqueter les échantillons d'une structure connue. Mais, bien entendu, la souplesse trop grande de la méthode ne permet pas de l'utiliser pour la reconnaissance. Elle doit être considérée comme une étape dans la construction de structures standard qui seront utilisées ensuite avec des techniques plus rigides en vue de la reconnaissance.

Les auteurs remercient les personnes qui les ont aidés dans la constitution de la banque de données qui a été utilisée : MM. J.Y.GRESSER, BUISSON, REJAUD et MERCIER du CNET à Lannion.

BIBLIOGRAPHIE

[1] JL FLANAGAN "Speech Analysis, Synthesis and Perception". Springer Verlag Berlin (1972)

[2] G.SEBESTYEN "Decision making processes in pattern recognition" Mac Graw New York (1962)

[3] R.VIVES, L.BUISSON, J.Y.GRESSER, G.MERCIER, M.QUERRE "Reconnaissance de grands dictionnaires prononcés par plusieurs locuteurs" Journées d'études sur la parole - Orsay (15-17 mai 1974)

[4] C.BERGER-VACHON et G.MESNARD "Aspect statistique des variations acoustiques liées aux mécanismes de la phonation. Applications à la reconnaissance de la parole". A paraître sur la Revue d'Acoustique (1975)

[5] C.ROCHE, F.CHATEAUNEUF "Reconnaissance de la parole par algorithme d'apprentissage d'opérateurs". Journées d'études sur la parole - Orsay (15-17 mai 1974)

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

[7] R.W.SCHAFFER, L.R.RABINER "System for Automatic Formant Analysis of Voiced Speech" - The journal of the Acoustical Society of America [47] pp.634-648 (Feb.1970)

[8] M.CARTIER "Reconnaissances de vocabulaires spéciaux" Analyse et synthèse de la parole - Volume I (1972/73) pp.39-46 - Département Etudes et Techniques Acoustique CNET 22301 Lannion

[9] C.BERGER-VACHON et G.MESNARD "Evaluation de l'efficacité d'un système de paramètres dans un problème de reconnaissance des formes" Onde Electrique vol.50 fas.11 pp.920-933 (Dec.1970)

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

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

INFERENCE AUTOMATIQUE DE GRAMMAIRES FORMELLES

L. MICLET

RESUME

En Reconnaissance syntaxique des formes, le problème de l'apprentissage peut se formuler ainsi : est-il possible d'inférer une grammaire formelle à partir d'échantillons du langage qu'elle accepte ?

On présente ici quelques réponses proposées à ce problème, peu abordé malgré son grand intérêt théorique et pratique.

SUMMARY

In Syntactic Pattern Recognition, the problem of learning can be formulated as follows : is it possible to infer a formal grammar, from samples of its generated language ?

We present here partial answers to this problem, badly known in spite of its great interest.

6ème JOURNEES D'ETUDE SUR LA PAROLE
TOULOUSE 28 AU 30 MAI 1975

INFERENCE AUTOMATIQUE DE GRAMMAIRES FORMELLES

L. MICLET

* * * * *

INTRODUCTION

C'est en reconnaissance des caractères manuscrits que les premières méthodes que l'on peut qualifier de "syntaxiques" ont fait leur apparition. Il semblait en effet fructueux, plutôt que d'utiliser les procédés classiques, de décomposer les caractères en sous-formes primitives, par exemple des segments horizontaux, obliques, etc..., et de chercher pour un caractère donné, les relations entre ces primitives. On parvient de la sorte, en explicitant les primitives et leur règles d'assemblage, à caractériser les classes de caractères, et à les reconnaître.

Depuis les premières approches dans ce sens, les méthodes syntactiques, qui font porter l'effort d'analyse sur la structure de la forme à reconnaître plutôt que sur l'étude statistique de paramètres qu'on en extrait, ont été surtout appliquées à la reconnaissance des images et à l'analyse de scène. On trouve aussi des applications en reconnaissance de la parole, où par exemple la connaissance de la structure de la phrase prononcée aide considérablement à l'examen de cette phrase.

D'une façon générale, les méthodes syntactiques, qui proposent une exploration structurelle des formes, semblent bien adaptées à toute une classe de problèmes, ceux pour lesquels la recherche de primitives et de leur règles d'assemblage est nécessaire pour la reconnaissance. On dispose alors d'un certain nombre d'outils, qui ont été développés à partir de la Théorie des Langages Formels, sous la forme de Grammaires et d'Automates. Par exemple, l'expérience de la Compilation des langages de programmation a été très utile pour les méthodes syntaxiques, en fournissant des algorithmes pour examiner les formes. On possède un cadre théorique cohérent, et des outils opératoires dans ce cadre, pour formaliser de nombreux problèmes de reconnaissance syntaxique.

Un des problèmes abordés en reconnaissance des formes classique est celui de l'apprentissage. Il s'agit de définir les fonctions de choix sur les formes à partir de formes d'apprentissage connues. Un grand nombre de travaux, à base d'outils statistiques, ont été proposés sur ce point, qui est naturellement d'un très grand intérêt.

Si on se pose le même problème pour les méthodes syntaxiques, dans le cadre des Grammaires Formelles, on aborde une partie

../..

peu explorée de cette théorie. On l'appelle "inférence grammaticale". Est-il possible d'"inférer" une grammaire qui représente la structure d'une classe de formes, à partir d'échantillons de cette classe ?

Avant de décrire quelques solutions proposées au problème de l'inférence, nous allons formaliser le cadre des grammaires, et préciser leur utilisation pour la reconnaissance des formes.

I - Les Grammaires Formelles en Reconnaissance des Formes

I-1. Définition

Une "grammaire formelle" G est par définition un quadruplet $G = (V_T, V_N, S, P)$ où

V_T est le "vocabulaire terminal" de G

V_N le "vocabulaire auxiliaire" de G

S l'"axiome" de G

P l'ensemble des "règles de productions" de G

Les éléments de V_T (resp. : V_N) sont appelés "éléments (ou symboles) terminaux" (resp. : auxilaires, ou non-terminaux). On les notera par des lettres minuscules (resp. : majuscules). On notera V_T^* l'ensemble de toutes les "chaînes" que l'on peut former en juxtaposant ("concaténant") les éléments terminaux, et en y adjoignant un élément, neutre pour cette opération associative.

Par exemple, si $V_T = \{a, b\}$

$V_T^* = \{ \varepsilon, a, b, aa, ab, ba, bb, aaa, ... \}$

Les éléments de V_T^* sont appelés "phrases terminales", ou simplement "phrases". Un élément de P, ensemble des règles de productions de G, s'écrit sous la forme :

$$\alpha \rightarrow \beta$$

$$\text{où } \alpha, \beta \in (V_T \cup V_N)^*$$

On désigne aussi les éléments de P sous le nom de "règles de réécriture", pour une raison qui va être explicitée ; on dira aussi simplement "règles".

Nous allons nous limiter ici à l'étude de deux types particuliers de grammaires, selon la forme de leurs règles :

1-Grammaires régulières

Leurs règles sont de la forme : $M \rightarrow a\,N$

ou $M \rightarrow a$

avec $M, N \in V_N$

$a \in V_T$

../..

Autres noms de ces grammaires :

rationnelles, de type 3, K-grammaires.

2-Grammaires algébriques

Leurs règles sont de la forme : $M \rightarrow \alpha$

avec $M \in V_N$

$\alpha \in (V_T \cup V_N)^*$

Autres noms : context-free, de type 2, C-grammaires.

I-2. Utilisation : exemple régulier

Le mécanisme d'une grammaire formelle va être décrit dans un exemple de Reconnaissance des Formes. Dans ce cadre

- une forme est représentée par une chaîne de terminaux. Les éléments primitifs que l'on a extrait de la forme sont les terminaux de la grammaire.
- une classe de formes est représentée par une grammaire $G = (V_T, V_N, S, P)$.
- pour décider si une forme f appartient à la classe de G, il faut constater si on peut engendrer, à partir de S, à l'aide de règles de P, la chaîne f sur V_T^*.

Pour décrire ce processus génératif, prenons la grammaire régulière suivante :

$G = (V_T, V_N, S, P)$

avec $V_T = \{a, b, c\}$

$V_N = \{A, B, C, D, E\}$

P =

1	S	→ a A	8	C	→ b D
2	S	→ b E	9	C	→ b
3	S	→ c C	10	D	→ a A
4	A	→ b B	11	D	→ a
5	A	→ b	12	E	→ a A
6	B	→ c C	13	E	→ c C
7	B	→ c	14	E	→ a
			15	E	→ c

Soit la forme f = c b a b c b a b $\in V_T^*$

En partant de l'axiome S, et en appliquant des règles de P, cherchons à générer f, On a : S → c C (règle 2) mais C → b D (règle 8).

Donc, on a la suite : ../..

$S \rightarrow c\ C \Longrightarrow c\ b\ D$

(le signe $\Longrightarrow$ indique l'application d'une règle à un non-terminal)

En appliquant successivement les règles 10, 4, 6, 8, 10, on arrive à $S \rightarrow \ldots \Rightarrow c\ b\ a\ b\ c\ b\ a\ A$

La règle finale à appliquer est 5, et on trouve :

$S \rightarrow \ldots \Longrightarrow f_{\pi}$

On note : $s \Longrightarrow f$

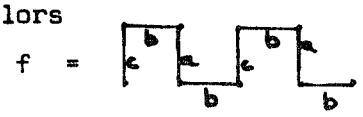
On a donc réussi à engendrer la forme f par la grammaire G. On dit que f est un élément du langage engendré par G, noté L(G); un langage est donc simplement ici une partie de V_T^*.

$$f \in L(G) \Longleftrightarrow S \overset{*}{\Rightarrow} f$$

Il est facile dans cet exemple d'avoir une idée de cette classe de formes qu'est L (G).

Si on suppose que les terminaux sont issus du pré traitement d'images de courbes, et que l'on ait symbolisé par a, b et c les segments orientés

On aura alors

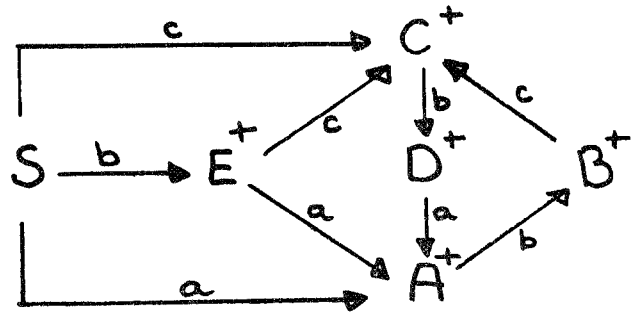


L (G) est en fait l'ensemble de toutes les suites de créneaux c b a b, sans contrainte sur le segment de début ou de fin. On peut obtenir une représentation plus visuelle, sous forme d'"automate fini", c'est-à-dire un graphe où les sommets seront les éléments de V_N, et les arcs seront étiquettés par ceux de V_T. Il y aura un arc

$$A \xrightarrow{a} B \quad \text{dans le graphe}$$

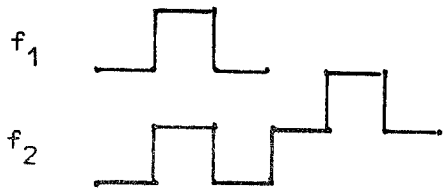
Ssi $A \rightarrow aB \quad \in P$.

En introduisant le sommet de départ S et les sommets finaux (marqués par "+") on arrive à :



../..

Pour générer une phrase f de V_T^*, il faut trouver un chemin entre S et un sommet de sortie tel que la suite des étiquettes des arcs de ce chemin soit égale à f. A l'aide du prétraitement, qui doit transformer les courbes en phrases sur V_T^*, on peut donc séparer les deux formes



On a en effet : $f_1 \in L(G) \subset V_T^*$

$f_2 \in V_T^* - L(G)$

Aucun chemin dans l'automate ne décrit f_2.

I-3. Exemple context-free

Dans le cas des grammaires context-free, on ne peut pas avoir de représentation de la grammaire sous forme de graphe. Par contre, on peut visualiser la génération d'une phrase par la grammaire selon un arbre, appelé "arbre de dérivation".

Soit par exemple la grammaire suivante, qui génère des expressions parenthesées, ce qui est très caractéristique de la "puissance" des grammaires context-free

$G = (V_T, V_N, S, P)$

VT = { a, b, o, 1 .... 9, +, *,), (}

V_N = { A, B, C, D }

P = 1 S → A
2 S → S + A
3 A → B
4 A → A * B
5 B → C
6 B → D
7 B → (S)
8 C → a
9 C → b
10 D → o
:
19 D → 9

$f = (a + b) * 9 \in V_T^*$

Montrons que $S \overset{*}{\Rightarrow} f$, c'est-à-dire : $f \in L(G)$

$S \underset{1}{\rightarrow} A \underset{4}{\rightarrow} A * B \underset{3}{\Rightarrow} B * B \Rightarrow (S) * B$

$\underset{2}{\Rightarrow} (S + A) * B \underset{1}{\Rightarrow} (A + A) * B \underset{3}{\Rightarrow} (B + A) * B$

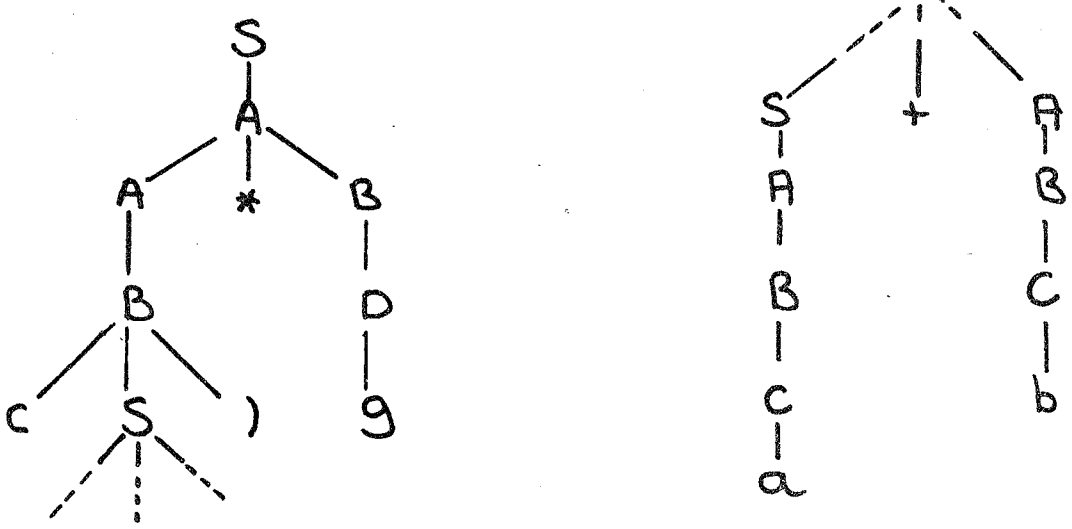
../..

$\Rightarrow_{7}$ (C + A) * B $\Rightarrow_{8}$ (a + A) * B $\Rightarrow_{3}$ (a + B) * B

$\Rightarrow_{5}$ (a + C) B $\Rightarrow_{9}$ (a + b) * B $\Rightarrow_{6}$ (a + b) * D

$\Rightarrow_{9}$ (a + b) * 9

L'arbre de dérivation de f est le suivant :



Les éléments de f sont les sommets pendants de l'arbre de dérivation, lus de gauche à droite, puisque la génération de f a été produite en dérivant systématiquement le terminal le plus à gauche.

I-4. Applications

On a utilisé pour de nombreux problèmes de reconnaissance des formes le cadre théorique des grammaires formelles. Dans le domaine des images, on trouvera dans les références (11), (13), (14), quelques exemples de cette utilisation. Pour la parole, les méthodes syntaxiques sont utilisées à tous les niveaux du processus de reconnaissance. Quelques exemples sont cités dans les références (8), (9), (10), (12).

Dans le cadre de ces applications, on utilise souvent des grammaires dont la définition n'est plus exactement celle d'une grammaire context-free, mais en dérive par des modifications. Il existe par exemple toute une classe de grammaires, dites "stochastiques", pour lesquelles on attribue à chaque règle de production une probabilité ; une chaîne sur V_T^* sera acceptée avec une probabilité égale au produit des probabilités des dérivations qui l'engendrent (s'il n'y a qu'une seule façon de l'engendrer par la grammaire).
On peut signaler aussi les grammaires dites "de précédence" et les grammaires LR (k), pour lesquelles il existe des algorithmes très performants d'analyse syntaxique (c'est-à-dire qui répondent au problème :

../..

étant donnés f ε V_T^* et G, décider si f ε L (G)
ou si f ∉ L (G)).

Un grand nombre de modifications ont été ainsi apportées au formalisme classique des grammaires, dans le but, soit d'en étendre le champ d'application, soit de simplifier certains algorithmes. C'est ainsi que certains procédés d'inférence, très liés à la forme des règles de grammaire à trouver, s'appliquent seulement à des sous-ensembles de la classe des grammaires context-free.

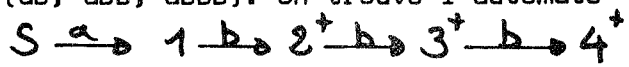
II - Le problème de l'inférence

II-1. Introduction

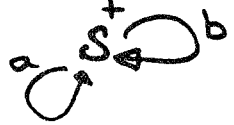
Soit un espace de formes E. Vu sous l'angle de la reconnaissance syntaxique, E est une partie de V_T^*, où V_T est un vocabulaire terminal de grammaire. On dispose d'un certain nombre de grammaires sur V_T^*, G_1 —— G_n, qui modélisent les classes de formes sur E. On dira que la forme f appartient à la classe i si f ε L (Gi), c'est à dire si la forme f ε V_T^* est reconnue par la grammaire Gi. Si les grammaires ont été bien construites, une forme quelconque de E ne sera acceptée que par une seule grammaire. On pourra donc classer une forme inconnue, en cherchant parmi les grammaires G_1 —— Gn celle qui peut la générer.

Cependant, il est particulièrement intéressant d'essayer de construire automatiquement les grammaires G_1 —— Gn à partir d'échantillons connus des classes qu'elles représentent, c'est-à-dire de résoudre le problème de l'apprentissage pour la reconnaissance des formes selon les méthodes syntaxiques.

En formulant le problème en terme de grammaires formelles, ceci revient à poser la question :
est-il possible de trouver une grammaire G = (V_T, V_N, P, S) à partir d'un ensemble fini I, avec I c L (G) ?.
Les inconnues sont évidemment V_N et P, car il est naturel de prendre pour V_T les éléments distincts apparaissant dans les chaînes sur V_T^* de I.

Il y a une infinité de solutions à ce problème. Les deux premières sont triviales.
1-Il est très facile de trouver la grammaire qui engendre I et seulement I. Ce sera une grammaire de type 3, que nous noterons G_I, par exemple, si V_T = {a, b} I = {ab, abb, abbb}. On trouve l'automate fini qui n'accepte que I : 

cette solution n'a pas d'intérêt, car la grammaire trouvée n'acceptera pas de forme inconnue, et ne saura réconnaître que les formes de son ensemble d'apprentissage.
2-L'autre grammaire triviale est celle qui accepte V_T^* entier. C'est aussi une grammaire de type 3. Pour V_T = {a, b}, la grammaire se représente par



../..

Elle n'a pas non plus d'intérêt puisqu'elle accepte tout élément de V_T^* et ne peut servir à définir un langage. C'est la grammaire universelle de V_T.

3-On peut toujours, étant donné une grammaire G1 trouver une autre grammaire G2 telle que L (G1) = L (G2).
A partir d'une solution, on peut en engendrer ainsi une infinité.
Il faut donc circonscrire le problème, et se définir des critères, plus ou moins bien formulés, que l'on cherchera à optimiser.

Dans un problème d'inférence, les solutions cherchées doivent tendre à s'équilibrer entre deux pôles : d'une part la "complexité" de la grammaire solution, d'autre part la bonne correspondance entre le langage qu'elle engendre et l'ensemble des phrases d'apprentissage, c'est-à-dire en fait entre les deux solutions triviales. On a proposé de classer les méthodes proposées pour l'inférence grammaticale en deux types de procédés, selon que l'on optimise en contrôlant l'un ou l'autre critère.

Le premier procédé consiste à énumérer un grand nombre de grammaires d'un type donné, sur le vocabulaire terminal de l'échantillon, et de mesurer leur adéquation à l'échantillon.

Le second consiste au contraire à essayer de remonter à partir de l'échantillon de phrases I, d'y découvrir des règles, et d'induire ainsi la grammaire à partir de ses productions connues.

Les méthodes pratiques sont cependant difficiles à classer de façon aussi tranchée, car l'énumération est bien souvent contrôlée par l'échantillon, et l'induction par des modifications systématiques sur les symboles de la grammaire.

Un des problèmes importants de l'inférence grammaticale est de décider à quel moment on doit introduire des structures récursives dans la grammaire générée. Par exemple, à V_t = {a, b} et I = {ba, baa, baaa}. Doit-on généraliser I = $\{ba^n\}$ n = 1, 2, 3 à L (v) = (ba^n) n = 1, ... et ainsi inférer la grammaire

G S → b A
   A → a A
   A → a

Naturellement, en l'absence d'informations supplémentaires, le choix de cette généralisation est tout à fait arbitraire. C'est pour cette raison que les méthodes d'inférence font souvent appel soit à un informateur, soit à un paramètre arbitraire.

Nous allons donner quelques indications sommaires sur les principales méthodes avancées, en signalant qu'il ne s'agit ici que d'une introduction aux idées proposées en inférence grammaticale, et non pas d'une description exhaustive des algorithmes. On pourra trouver dans les références, le détail des méthodes et parfois des exemples élaborés d'inférence.

../..

III - Méthodes d'inférence

III-1. Horning (1) utilise les probabilités pour mesurer d'une part la complexité d'une grammaire, d'autre part l'adéquation d'une grammaire donnée à un échantillon de langage donné. Il s'intéresse aux grammaires stochastiques, c'est-à-dire qu'à chaque règle de la grammaire est affectée une valeur, que l'on peut interpréter comme la probabilité d'application de cette règle.

Si C est la classe des grammaires dans laquelle on cherche à décrire l'échantillon I, composé des éléments (I_1 ... I_n), on suppose que l'on peut mesurer la probabilité P (Gi) de la grammaire Gi dans le contexte C. D'autre part, il faut affecter une probabilité P (Ij) à chaque élément de I, et mesurer la valeur $P(I_j|G_i)$ probabilité avec laquelle la chaîne I_j peut être engendrée par la grammaire G_i.

Le théorème de Bayes permet alors de calculer la probabilité de la grammaire Gi conditionnellement à I :

Les échantillons étant supposés indépendants, on a :

$$P(I|G_i) = \prod_{j=1,n} P(I_j|G_i)$$

Et la formule :

$$P(Gi|I) = \frac{P(Gi).P(I|Gi)}{\prod_j P(Gj).P(I|Gj)}$$

On peut alors construire un algorithme, dont Horning montre qu'il fournit la grammaire la plus probable pour l'échantillon.

Malgré l'organisation performante de son énumération, cette méthode est très avide de temps de calcul, et elle n'a permis de trouver que des grammaires extrêmement simples. Cependant elle possède l'intérêt de définir un certain type d'optimalité, et de se placer dans un cadre rigoureux.

III-2. Cook (2) se place dans la classe des grammaires context-free stochastiques. Il définit la complexité d'une grammaire de cette classe en calculant l'information qu'elle possède, selon la longueur et le nombre de ses règles, et le nombre d'occurences des symboles dans une règle. Il définit encore une distance entre deux langages stochastiques : ceci servira à chaque étape de l'algorithme à mesurer l'adéquation entre une grammaire trouvée (c'est-à-dire le langage qu'elle engendre) et l'échantillon, supposé encore ici affecté de probabilités.

L'algorithme part de la grammaire triviale GI telle que L (GI) = I et autorise à chaque étape des manipulations sur les symboles de la grammaire trouvée tels que : substitutions, disjonctions (création des règles A → B et A → C et remplacement de B et C par A dans les autres règles), etc... A chaque étape, il s'agit de trouver les manipulations qui diminuent le plus la complexité de la grammaire, sans augmenter la distance de cette grammaire à l'échantillon. L'algorithme d'optimisation de ces deux paramètres qui est proposé n'assure pas la convergence vers la "meilleure" grammaire

../..

mais fournit de bons résultats pratiques.

A partir d'un échantillon d'une vingtaine d'éléments, la grammaire d'addition parenthésée suivante a été inférée :

$S \rightarrow a \mid b \mid (S) \mid S + S$

(la notation "|" signifie "ou" par exemple : $S \rightarrow a \mid b \Leftrightarrow S \rightarrow a$; $S \rightarrow b$)

III-3. Pao (6) a proposé un algorithme analogue pour les grammaires régulières. En partant de la grammaire triviale GI, on essaye toutes les fusions possibles de non-terminaux. L'ensemble des grammaires dérivées par cette opération de GI forme un treillis pour la relation

$$Gi > Gj \Leftrightarrow L(Gi) \supseteq L(Gj)$$

Il faut parcourir cet ensemble de grammaires à l'aide d'un contrôle de l'opérateur : le programme lui proposera des phrases caractéristiques de V_T^*, qui ne sont pas dans I, et selon la réponse "cette phrase est dans L (G)", ou "cette phrase n'est pas dans L (G)", le programme choisira son parcours dans l'ensemble des grammaires.

La grammaire trouvée correspondra à la confusion d'un certain nombre de non-terminaux de la grammaire triviale, donc à une de ses généralisations possibles.

III-4. Crespi Reghizzi (3) suppose deux choses sur l'échantillon d'apprentissage. D'abord qu'il appartient au langage généré par une grammaire de précédence, c'est-à-dire qu'il existe des relations binaires entre les symboles de la grammaire, permettant de faire une analyse syntaxique directe, à la manière de celle pratiquée dans les automates finis. D'autre part, l'échantillon doit être structuré: chacun de ses éléments doit comporter une information supplémentaire (un parenthésage) qui indique la profondeur de ses éléments dans son arbre de dérivation. L'algorithme d'inférence est alors directement dérivé d'une analyse syntaxique classique. Il conduit à inférer des grammaires assez complexes, à l'intérieur du cadre où il s'est placé.

III-5. Biermann et Feldman (5), pour les grammaires régulières, définissent à partir de l'échantillon I des sous-langages, notés Sw, tels que : $Sw = \{ n \mid w\,n \in I \}$. Deux sous-langages sont équivalents s'ils sont identiques pour toutes les chaînes de longueur inférieure ou égale à un paramètre k. A chaque sous-langage on associera un non-terminal de la grammaire à inférer, par la règle : $A \rightarrow a\,B$, s'il existe deux sous-langages Sw et Sw' et une chaîne $w \in I$ telle que $w' = w\,a$.

On démontre que pour k égal à la longueur de la chaîne la plus longue dans I, on infère ainsi la grammaire triviale GI, et pour $k = o$, la grammaire universelle. On dispose donc d'un paramètre réglable, en relation directe avec la puissance des grammaires inférées.

III-6. Evans (4) cherche à réduire la grammaire GI issue de l'échantillon à l'aide d'une série de règles de manipulation.

../..

- Chercher les couples de non-terminaux tels que, si on les identifie, on obtient des règles multiples. Choisir le couple qui produit par confusion le plus de règles multiples. Les supprimer toutes sauf une.
- Chercher une paire (A, a) ε $V_N \times V_T$ telle que l'addition de la règle $N \rightarrow n$ et le remplacement de n par N conduisent à des règles multiples. Agir de même que précédemment.

Evans a développé cette méthode pour des généralisations des grammaires formelles appliquées aux images : les relations entre "primitives" (terminaux) peuvent être nombreuses, au lieu d'être réduites à la concaténation. C'est une ouverture intéressante à l'inférence dans un formalisme moins strict.

CONCLUSION

Il y a pour le moment peu de résultats probants en inférence grammaticale, qui est pourtant un problème très important, non seulement dans le cadre formel de la Théorie des langages, mais bien aussi pour son intérêt essentiel en Reconnaissance Syntaxique des Formes. L'une des raisons à cela est la grande difficulté de faire un lien entre les deux représentations d'un langage, l'une sous forme générative : sa grammaire G, l'autre sous forme exhaustive : la suite de ses phrases L (G). On ne sait pas évaluer systématiquement l'influence sur L (G) de l'addition d'une règle à G, ou de la confusion de deux de ses non-terminaux par exemple.

D'autre part, l'intérêt d'avoir un modèle génératif est que celui-ci représente une infinité potentielle de phrases sous une forme particulièrement condensée. Puisque l'échantillon d'apprentissage est fini, il faut choisir avec soin dans quelle "direction" on doit l'agrandir à l'infini, en inférant une grammaire.

Enfin, il est difficile de juger jusqu'à quel point une grammaire inférée est optimale vis à vis de l'échantillon. Est-elle trop générale, ou "colle"-t'elle par contre trop près aux phrases qui ont servi à la créer ? La réponse ne peut venir que de l'utilisation postérieure à l'apprentissage, et pour le moment on dispose de peu d'exemples complets d'apprentissage, puis de reconnaissance en méthodes syntaxiques. Cependant, la voie ouverte semble très riche, et la nécessité de l'explorer apparaît de plus en plus au fur et à mesure que se développent les procédés syntaxiques de reconnaissance des formes.

REFERENCES :

1 - Horning : A procedure for grammatical inference.
Inf. Proc. 71 - North-Holland, Pub. Co. 1972.

2 - Cook -Rosenfeld : Some experiments in grammatical inference.
Nato advanced study institute : "Procédures informatiques d'apprentissage". Bonas 1974.

3 - Crespi Reghizzi : An effective model for Grammar Inference.
Inf. Proc 71, North-Holland Pub. Co. 1972

../..

4 - Evans : Grammatical inference Techniques in Pattern analysis. Software Engineering vol 2, Ed. by Tou. Academic Press 1971.

5 - Biermann-Feldman : On the synthesis of Finite-State Machines from samples of their Behaviour. IEEE on Computers Juin 1972.

6 - Biermann-Feldman :A survery of results in grammatical inference. Frontiers of Pattern Recognition. Ed. by Watanabe. Academic Press 1972.

7 - Fu : Syntactics Methods in Pattern Recognition. Academic Press 1974.

8 - Levinson : An artificial intelligence approach to automatic speech recognition. PhD Thesis/Univ. of Rhode Island 1974.

9 - De Mori - Laface - Piccolo.
Research for a grammar of Speech.

10 - Thon : Error correction in automatic speech recognition systems by use of formal language syntax.
9 et 10 : Speech Communication Seminar. Stockholm 1974.

11 - Narasimhan. The role of syntactic models in picture processing.

12 - Neely-White : On the use of syntaxe in a low cost real time speech recognition system.
11 et 12 : IFIP Congress 74 - Stockholm - Vol 4.

13 - Thomason-Gonzalez : Classification of imperfect syntactic pattern structures.

14 - Fung,Fu : Stochastic Syntactic classification of noisy patterns.
13 et 14 : 2° I J C P R Copenhague 1974.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

LES CONTRAINTES PHONOLOGIQUES DANS UN SYSTEME DE RECONNAISSANCE DE LA PAROLE

Mario ROSSI

(Institut de Phonétique d'Aix-En-Provence)

-=-=-=-=-=-

RESUME Les contraintes phonologiques traduisent les liens de dépendance entre les éléments de la chaîne de parole. On propose une segmentation de la phrase à partir d'indices prosodiques, la reconnaissance prioritaire du noyau accentuel dans des algorithmes où sont introduites les règles de contraintes, enfin la reconnaissance de la chaîne inaccentuée des groupes intonatifs à l'aide d'un modèle de transitions d'état, image des contraintes sur la structure des consonnes inaccentuées.

SUMMARY Phonological constraints display the dependence relationships between the elements of the speech chain.

A segmentation of the sentence is proposed, based on prosodic cues, the priority recognition of the stress nucleus in algorithms in which constraint rules are introduced, and finally the recognition of the unstressed chain of intonation groups thanks to a model of state transitions, an image of the constraints on the structure of unstressed consonants.

LES CONTRAINTES PHONOLOGIQUES DANS UN SYSTEME DE RECONNAISSANCE DE LA PAROLE

Mario ROSSI
(Institut de Phonétique d'Aix-en-Provence)

-=-=-=-=-=-

La procédure d'analyse des contraintes phonologiques exposée ici ne se substitue pas aux techniques de reconnaissance acoustique ; elle peut et doit s'insérer à un certain niveau dans les algorithmes de reconnaissance afin d'en améliorer le rendement.

Cette procédure suppose un système de reconnaissance par poursuite de formants, mais elle peut également être appliquée, après les adaptations nécessaires, à toute autre technique (phonatomes(1), vecteurs Δ (2), etc...).

Toutefois, elle est prévue pour un système (3) susceptible de fournir :

1° L'évolution de l'amplitude,
2° L'évolution du fondamental,
3° L'évolution des trois premiers formants sur les voyelles accentuées,
4° La valeur moyenne des trois premiers formants sur les voyelles atones.

La procédure que nous proposons est prévue pour la reconnaissance de phrases.

La contrainte est définie comme un lien de dépendance entre deux ou plusieurs éléments, lien de dépendance qui élimine certaines combinaisons et augmente la redondance du système.

Les contraintes seront d'abord identifiées au sein du noyau accentuel, situé en français à la fin du groupe intonatif. La reconnaissance du noyau accentuel s'impose en effet de façon prioritaire car il constitue,dans la perception, un point d'ancrage essentiel à partir duquel l'auditeur émet un certain nombre d'hypothèses sur la phrase.

Nos expériences prouvent que la suppression du noyau accentuel rend la phrase incompréhensible. Les noyaux accentuels en effet sont mieux définis que les syllabes atones : les voyelles et les consonnes sont environ deux fois plus longues à la fin du groupe intonatif qu'à l'intérieur du

groupe ; la voyelle accentuée atteint aisément son régime permanent : de ce fait, les zones de dispersion et de recouvrement formantiques sont très réduites dans ce contexte. Les travaux de Haton et de Vicens (4) montrent que seules les voyelles accentuées sont correctement segmentées et reconnues. D'autre part, les transitions de formants aux bornes de ces éléments stationnaires sont nettement marquées. Les bornes de la voyelle et les transitions formantiques sont accessibles à partir d'un certain seuil de variation de l'amplitude globale qui est fonction du rapport d'amplitude consonne/voyelle (5).

L'identification du noyau accentuel implique une première segmentation de l'énoncé en groupes intonatifs. Di Cristo (6), Faure (7) et Rossi (8) définissent le lien qui existe, dans certains types de phrases, entre le nombre et la forme des schémas intonatifs et la structuration syntaxique de l'énoncé et dégagent les indices prosodiques qui autorisent la segmentation en groupes intonatifs.

A - Les contraintes et la situation.

Les combinaisons possibles dans le noyau accentuel sont théoriquement très élevées. Mais une première série de contraintes est imposée par la situation.

Le rôle que joue la situation dans la perception n'est plus à démontrer. On peut affirmer qu'il est impossible de procéder à une reconnaissance correcte si on ne tient pas compte des contraintes imposées par la situation. Mais il peut sembler chimérique -en l'état actuel de la recherche- d'identifier les éléments qui définissent une situation et de les formaliser.

Toutefois, si on envisage la situation comme un univers où le lexique et la combinatoire phonologique sont limités, il est possible d'appréhender certains des éléments qui la caractérisent.

On peut penser en effet que les règles de redondance et de neutralisation utilisées par un auditeur sont propres à une situation. Une façon, par conséquent, de faire intervenir la situation dans un système de reconnaissance consisterait à définir des contraintes non pas sur l'ensemble du lexique du français mais sur l'univers limité qui correspond à une situation donnée.

Afin de ne pas limiter arbitrairement la situation, nous avons choisi comme univers de référence les 5 000 mots du Français Fondamental présenté par Juilland (9). Ce lexique a l'avantage de fournir les mots les plus fréquents et d'éliminer les mots rares. Ce choix permet d'éviter le re-

TABLEAU 1 a

Lexique (F. F.) : classé en fonction du noyau accentuel.

	U	MDG	MDF
1 - TUT	600	1	1
1 - toute,a.			
2 - TIK	419	342	25
2 - politique,n.	63.84		
3 - politique,ad.	55.16		
4 - critique,n.f.	20.13		
5 - pratique,n.	18.34		
6 - pratique,ad.	17.65		
7 - domestique,n.	12.47		
8 - dramatique,ad.	12		
9 - caractéristique,a.	10.59		
10 - boutique,n.	8.87		
11 - poétique,a.	8.48		
12 - antique,a.	8.29		
13 - identique,a.	8.34		
14 - authentique,a.	7.61		
15 - critique,n.m.	7.45		
16 - mystique,ad.	7.36		
17 - sympathique,a.	7.27		
18 - rustique,a.	7.12		
19 - critique,a.	5.25		
20 - ecclésiastique,a.	5		
21 - mathématique,a.	5		
22 - diplomatique,a.	4.60		
23 - domestique,a.	4.45		
24 - plastique,a.	4.41		
25 - mystique,n.	3.64		
26 - gymnastique,n.	3.41		
3 - TIT	270	1	1
27 -petite,a.	258		
28 - petite,n.	12		
4 - TYD	258	37	14
29 - étude,n.	72.66		
30 - habitude	58.71		
31 - attitude	44.84		
32 - inquiétude	19.29		
33 - incertitude	12.59		
34 - certitude	9.72		
35 - solitude	8.29		
36 - altitude	7.17		
37 - exactitude	6.15		
38 - gratitude	5		
39 - lassitude	4.12		
40 - servitude	3.41		
41 - aptitude	3.08		
42 - ingratitude	3.08		
5 - KÕT	162.18	12	4
43 - compte,n.	120		
44 - compte,v.	20		
45 - comte,n.	15.80		
46 - conte,n.	6.38		
6 - TÈT :	152	10	1
47 - tête			
7 - TÃT	105	121	6
48 - tante,n.	12.29		
49 - entente,n.	7.21		
50 - attente,n.	6.95		
51 - éclatante,a.	6		
52 - détente,n.	3.08		
53 - tente,n.	1		
8 - KÃT :	53.95	38	3
54 - cinquante,a.	46.36		
55 - fréquente,a.	4.59		
56 - éloquente,a.	3		
9 - TÕD :	43.89	6	1
57 - méthode,n.			
10 - TÕB :	31.82	3	1
58 - tombe,v.	22		
59 - tombe,n.	9.82		
10 - PID :	31.28	11	2
60 - rapide,a.	25.96		
61 - stupide,a.	5.32		

TABLEAU 1 b

10 - KUT :	31	2 - 1
62 - écoute,v.	22	
63 - coûte,v.	9	
11 - TIP :	25.79	9 - 1
64 - type,n.		
11 - KÈT :	25.24	34 - 3
65 - conquête,n.	11.80	
66 - enquête,n.	7.36	
67 - casquette,n.	6.08	
12 - KÔK :	24.71	3 - 2
68 - quelconque,a.	21.35	
69 - quiconque,pr.	3.36	
12 - KYP :	24	1 - 1
70 - occupe,s.pl.v.		
13 - KÒT :	20.91	3 - 2
71 - côte,n.(partie du corps)	10.91	
72 - côte,n.	10	
14 - PAT :	18.11	6 - 1
73 - patte,n.		
15 - KAT :	16	6 - 1
74 - délicate,n		
16 - TAK :	15.59	3 - 1
75 - attaque,n.		
16 - TIG :	15.42	1 - 1
76 - fatigue,n.		
17 - KÒK :	12.93	8 - 2
77 - coq,n.	9.29	
78 - coque,n.	3.64	
17 - TAP :	12.80	2 - 2
79 - étape,n.	12.03	
80 - tape,v.	0.77	
18 -KÛP :	11.41	4 - 2
81 - coupe,n.	7.41	
82 - coupe,v.	4	
19 - PÈT :	10.97	20 - 2
83 - tempête,n.	6.85	
84 - trompette,n.	4.12	
19 - PÂT :	10.93	20 - 3
85 - pente,n.	7.93	
86 - frappante,a.	2	
87 - grimpante,a.	1	
19 - PAP :	10.69	2 - 1
88 - pape,n.		
20 - PIK :	9.5	27 - 2
89 - typique,a.	5	
90 - pique,v.	4.5	
21 - KIT :	8.9	6 - 1
91 - quitte,v.		
22 - PYT :	7.64	4 - 1
92 - dispute,n.		
22 - KÒD :	7.58	3 - 1
93 - code,n.		
23 - KUD :	6.85	1 - 1
94 - coude,n.		
23 - KID :	6	1 - 1
95 - liquide,a.		
24 - PIP :	5.81	1 - 1
96 - pipe,n.		
24 - TÂP :	5.52	3 - 1
97 - tempe,n.		
24 - PÈT :	5	2 - 1
98 - peinte,a.		
25 - KAP :	4.12	4 - 1
99 - cap,n.		
26 - KIP :	3.68	1 - 1
100 - équipe,n.		
26 - KÔD :	3	4 - 1
101 - féconde,a.		
26 - PÔP :	3	1 - 1
102 - pompe,n.		
27 - TÂD :	1	2 - 1
103 - tende,v.		
TOTAL	753	-103

(1) MDG : nombre de mots différents dans le Français Général

MDF : nombre de mots différent dans le Français Fondamental.

(2) U : valeur d'usage.

CONSONNE INITIALE :

TABLEAU 2a

RANG	USAGE	K	T	P
		CONTEXTES		
1	600		T u T	
2	419		T i K	
3	270		T i T	
4	258		T Y D	
5	162	K ɔ̃ T		
6	152		T ɛ T	
7	105		T ɑ̃ T	
8	53	K ɑ̃ T		
9	43		T ɔ D	
10	31	K u T	T ɔ̃ B	P i D
11	25	K ɛ T	T i P	
12	24	K ɔ̃ K K u P		
13	20	K o T		
14	18			P ɑ T

TABLEAU 2 b

RANG	USAGE	CONTEXTES		
15	16	K ɑ T		
16	15		T A K T i G	
17	12	K ɔ K	T ɑ P	
18	11	K u P		
19	10			P ɛ T P ɑ̃ T
20	9			P ɑ P P i K
21	8	K i T		
22	7	K ɔ D		P y T
23	6	K u D K i D		
24	5		T ɑ̃ P	P i P P ɛ̃ T
25	4	K ɑ P		
26	3	K ɔ̃ D K i P		P ɔ̃ P
27	1		T ɑ̃ D	

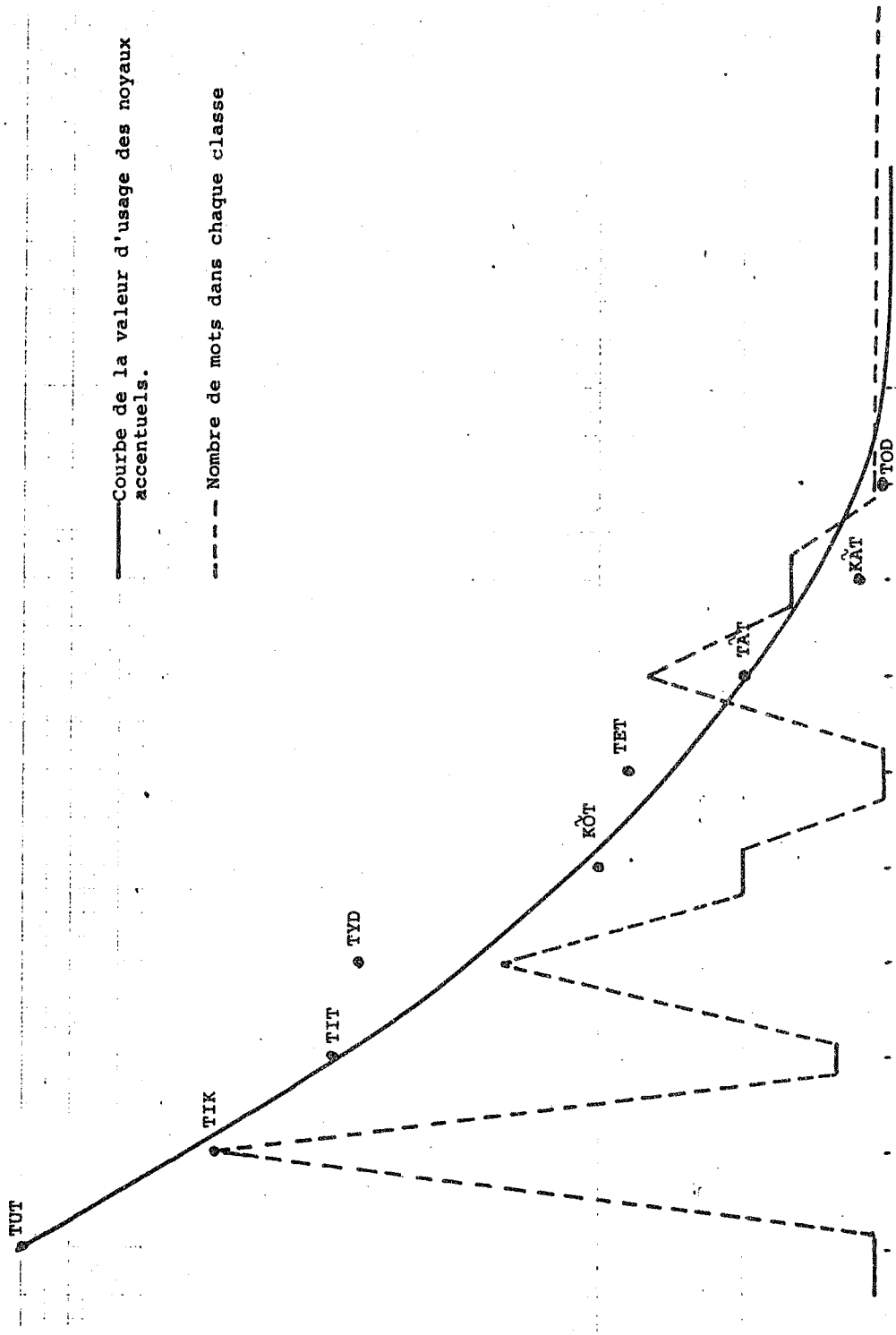


Figure 1 : Valeur d'usage des noyaux accentuels et des mots dans le Français Fondamental

recours à un modèle stochastique de reconnaissance. D'autre part, les contraintes définies sur cet univers sont plus représentatives de la langue réelle.

Afin de simplifier la présentation de notre technique d'analyse, nous avons limité momentanément les recherches au noyau accentuel précédé et suivi d'une occlusive, plus exactement aux suites du type :

{P, T, K} V {P, T, K, B, D, G} = Ocs V ocsz (occlusive sourde + voyelle + occlusive sourde ou sonore).

B - Présentation du matériau linguistique

Le Français Fondamental comprend 103 mots dont le noyau accentuel potentiel présente une structure de la forme ci-dessus définie, c'est-à-dire sept fois moins que le Français Général (ou français défini par la totalité du lexique).

On compte 41 noyaux accentuels différents du type Ocs V Ocsz. Si on range les items du lexique en fonction de la structure du noyau accentuel (tableau 1), le nombre de mots dans chaque classe permet de définir les contraintes sur les syllabes atones. On remarque que même dans les classes les plus fournies (par exemple : —— TIK : 25 items), le taux de redondance est considérable. Cette constatation nous amènera, pour rendre compte des contraintes dans ce contexte, à proposer un modèle particulier (voir p.13-14).

Le classement par valeur d'usage fait apparaître une forte dissymétrie dans la distribution statistique des noyaux accentuels : les noyaux qui commencent par /T/ sont largement en tête et représentent 92% de la valeur d'usage de l'ensemble(tableau 2).

Mais cette dissymétrie est compensée par la distribution des items : les classes de noyaux les plus fréquents comprennent également le plus grand nombre de mots (fig. 1). En conséquence, les mots qui appartiennent aux classes les plus fréquentes seront difficiles à reconnaître, tandis que les noyaux accentuels les moins fréquents, qui apportent le plus d'information, conduisent, paradoxalement, à une reconnaissance immédiate des mots qu'ils contiennent. C'est la raison pour laquelle les modèles stochastiques sont d'un faible secours dans le cadre de la reconnaissance automatique.

Les contraintes à la fin du groupe intonatif sont définies à partir du tableau 3 qui contient les seules combinaisons possibles au sein du noyau accentuel, dans le Français Général et dans le Français Fondamental.

/K/	P	T	K	B	D	G
i	⊕ 3.27	+	⊕ 9.5		+ 31.28	
e						
ɛ		⊕ 19.97	+		+	+
a	⊕ 18.69	⊕ 12.11	+		+	
ɑ						
ɔ	+	+	+		+	
o						
u	+					
y	+	⊕ 7.54	+			
ø		+				
ɔe						
ɑ̃	+	⊕ 10.93			+	
õ	⊕ 3	+			+	
ɛ̃		⊕ 5				

Tableau 3 a : Distribution des consonnes occlusives en Français Fondamental (⊕ chiffre) et en Français Général (+ et ⊕).

/K/	P	T	K	B	D	G
i	⊕ 3.68	⊕ 1			⊕ 6	
e						
ɛ		⊕ 25.24				
a	⊕ 4.12	⊕ 16	+	+	+	
ɑ						
ɔ	+	+	⊕ 12.93	+	⊕ 7.58	
o		⊕ 20.91			+	
u	⊕ 11.41	⊕ 31	+		⊕ 6.85	
ø						
œ						
ɑ̃	+	⊕ 53.95			+	
õ		⊕ 162.18	⊕ 24.71	+	⊕ 8	
ɛ̃		+				
y	+ 24	+	+	+		

TABLEAU 3 b : + Français Général
⊕ Français Fondamental

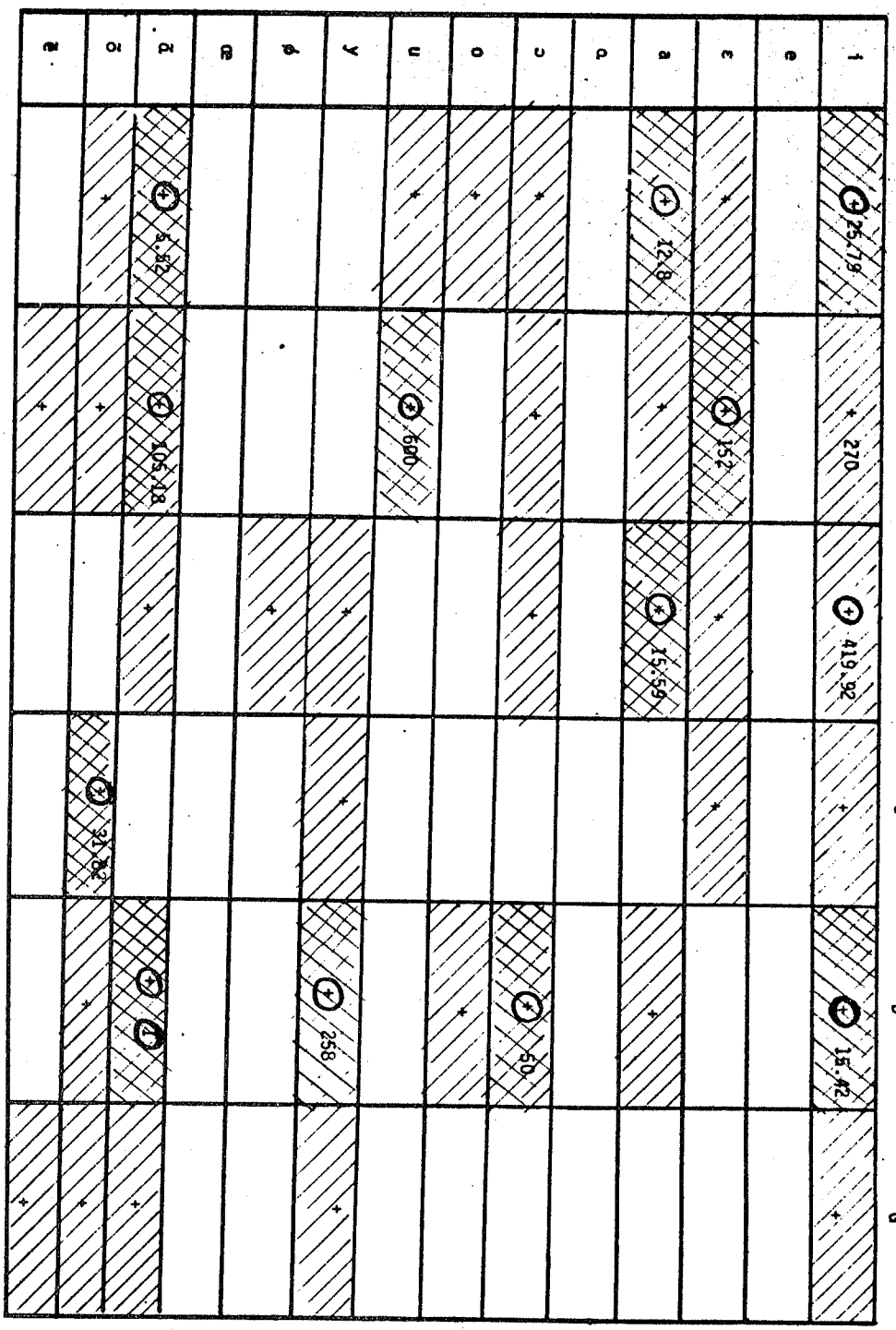

/ T /
P
T
K
B
D
G
ι
ε
α
ρ
ο
υ
γ
φ
β
ζ
σ
ε
25.79
270
419.92
12.8
152
15.59
15.42
690
50
258
4.82
105.18
21.82

TABLEAU 3 c

On voit que pour le Français Général les possibilités de combinaison sont peu nombreuses :

T V {P, T, K, B, D, G} = 39 sur 84 soit 46%
K V {P, T, K, B, D, G} = 33 sur 84 soit 39%
P V {P, T, K, B, D, G} = 30 sur 84 soit 35%

Les possibilités de combinaison sont réduites de plus de la moitié en Français Fondamental où les contraintes phonologiques sont donc très fortes :

K V {P, T, K, B, D, G} = 17 sur 84 soit 20%
T V {P, T, K, B, D, G} = 14 sur 84 soit 16%
P V {P, T, K, B, D, G} = 10 sur 84 soit 10%

Cette comparaison montre de quel poids peut peser la situation et les contraintes qui lui sont associées sur la reconnaissance de la chaîne sonore.

C - Contraintes phonologiques et structure syntaxique

Vicens (10) souligne qu'on doit classer le vocabulaire selon certaines caractéristiques afin d'éviter de parcourir tout le vocabulaire à la recherche d'un item. Il classe son vocabulaire en fonction des suites constrictives + voyelles. Nous avons, en ce qui nous concerne, opéré un double classement :

1° en fonction de la structure du noyau accentuel,
2° en fonction des catégories grammaticales (noms, adjectifs, verbes, etc...)

Ex. NOMS : 2 - TIK
2. politique
4. critique

ADJ. : 1 - TUT
1. toute

2 - TIK
3. politique

Les contraintes phonologiques sont évidemment liées aux structures syntaxiques (11) : l'hypothèse syntaxique que fait l'auditeur sur la phrase limite le nombre des choix possibles et augmente les contraintes phonologiques.

Nous nous sommes limité dans une première étape aux phrases représentées par la structure :

$$\underbrace{\text{Dét} + \text{Nom} + \begin{Bmatrix} V \\ Cop \end{Bmatrix}}_{SN} + \underbrace{\langle \text{Dét} \rangle \begin{Bmatrix} ADJ \\ \langle \text{Nom} \rangle \end{Bmatrix}}_{SV}$$

où Dét peut être réécrit comme : Dét→(Préar)+Ar+(Postar). C'est-à-dire en fonction des possibilités du lexique choisi : "toutes les boutiques, les cinquantes boutiques, {la, les, une, des} boutique(s)".

La structure syntaxique SN + SV, telle qu'elle vient d'être précisée, est toujours formée de deux groupes intonatifs au moins (6), car SN est toujours porteur d'un intonème progrédient (sauf dans le cas de pronominalisation).

EX. Les domestiques / comptent / les casquettes
Tout / est incertitude
La conquête / est une politique.

Par conséquent les points d'ancrage où seront définies prioritairement les contraintes seront constitués par la fin du Syntagme Nominal, qui est un Nom, par la fin du Syntagme Verbal, qui est un Adjectif ou un Nom et, éventuellement, par la dernière syllabe du Verbe si le Groupe Nominal (GN) qui le suit est plurisyllabique (premier exemple ci-dessus).

Enfin, si on analyse les relations sémantiques qui existent entre GN1 (Groupe Nominal Sujet) d'une part, et GN2 (Groupe Nominal Complément) et S. ATTR. (Syntagme Attributif) d'autre part, on remarque que les contraintes sont telles que pour un GN1 quelconque du Français Fondamental dont le noyau accentuel est de la forme TIK, on ne trouve sous GN2 et S. ATTR. que 27 noyaux différents sur :

- 41 attestés en Français Fondamental, soit 65%
- 102 en Français Général soit, 26%
- 252 possibilités théoriques de combinaison soit, 10%.

Les chiffres montrent que les hypothèses syntaxiques que l'auditeur est susceptible d'établir sur une phrase sont en nombre limité. Afin d'augmenter les chances de succès, un système de reconnaissance pourrait construire deux hypothèses syntaxiques, l'une à partir de SN, l'autre à partir de SV. La décision serait prise s'il y a convergence des résultats, après rétroaction aux niveaux phonologique et acoustique.

Notre propos n'était pas de développer ici un modèle syntaxique pour la reconnaissance de la parole. Nous avons fait allusion rapidement aux structures syntaxiques et aux relations sémantiques pour montrer la place des contraintes phonologiques dans un modèle de reconnaissance et leur lien avec les autres niveaux linguistiques.

On en est arrivé aujourd'hui à un stade de la recherche dans le domaine de la reconnaissance automatique où paradoxalement l'identification acoustique est, d'un certain point de vue, meilleure que la performance de l'oreille et où les résultats sont loin d'approcher la performance de l'auditeur humain. L'auditeur humain, en effet, pour compenser l'information somme toute grossière fournie par l'oreille, a recours aux contraintes phonologiques associées à des structures de haut niveau (syntaxiques, sémantiques), pour l'essentiel non marquées acoustiquement. Broadbent (12) a montré le rôle capital joué, dans la perception, par l'étage neurologique où la décision est prise en fonction du taux d'information contenu dans la parole et des liens de dépendance qui structurent les éléments de la phrase.

D - Technique d'analyse des contraintes.

On distinguera trois types de contraintes :

1° les contraintes absolues liées à la structure linguistique de la langue considérée :

Ex. Absence de É devant consonne en français

2° les contraintes dues aux lacunes fortuites : B peut apparaître après K :

Ex. Combe
Mais elle n'apparaît pas après KU : KUB n'existe pas. Rien n'empêche cependant que KUB provenant d'une langue étrangère par exemple, devienne un mot français. Cet item supprimerait une lacune fortuite.

Dans une analyse linguistique, ces lacunes ne constituent pas des contraintes. Pourtant, il est économique de les considérer comme telles en reconnaissance de parole car elles limitent sérieusement le nombre des combinaisons possibles ; cette économie compense largement la complexité des règles ad hoc qui peuvent rendre compte de ces lacunes.

3° les contraintes induites par la situation, dont nous avons déjà parlé.

I - Forme des règles de neutralisation

Chomsky présente la redondance lexicale sous forme de règles contextuelles de réécriture du type :

$$X \rightarrow Y \ / \ W \text{ —— } Z$$

où X ne contient aucun élément redondant et où Y restitue les éléments redondants dans le contexte W—Z.

$$(1)\ /+\ \text{Cons}/ \rightarrow \begin{bmatrix} + \text{ Stident} \\ + \text{ Continu} \\ + \text{ Aigu} \\ \text{etc...} \end{bmatrix} \quad / \#\text{——} \quad \begin{bmatrix} + \text{ Cons.} \\ - \text{ Vocal} \end{bmatrix}$$

Le trait Consonantique suffit à définir /s/ dans les groupes initiaux st, sp, ..., puisqu'aucune autre consonne n'est possible devant une occlusive. Mais cette procédure suppose que la redondance a été déjà identifiée. Or c'est précisément le but que nous nous proposons : dégager les traits redondants et définir les contraintes qui pèsent sur la distribution des unités.

Nous partons par conséquent de la forme phonétique pleine, caractérisée par tous les traits. Les règles, à ce niveau, doivent donc être la réciproque des règles de redondance lexicale de Chomsky.

Soit, en Français Général, la distribution des consonnes occlusives après K+V (tableau 3b). On voit que G n'apparaît pas après K+V où l'on ne rencontre que D et B. Si on définit G comme Compacte et {B, D} comme Non Compactes, on voit que seul le trait Non Compact associé au trait Voisé peut apparaître après K+V ; ce trait est donc redondant :

$$(2)\ [-\ \text{Compact}]\ /\ \underline{\begin{bmatrix} + \text{ Voisé} \end{bmatrix}} \rightarrow [0\ \text{Compact}] / +\text{Compact.Voy}\text{———}$$

Ce type de règle présente deux avantages :

a) le premier terme indique immédiatement quels sont les traits et les consonnes réalisés dans le contexte.

b) et puisque ce trait est neutralisé, on peut induire de la règle les suites interdites : K+V+G et celles qui sont autorisées : K+V+{B,D}.

Une règle de Chomsky présenterait cette redondance sous la forme :

$$(3)\ \begin{bmatrix} + \text{ Inter} \\ + \text{ Voisé} \end{bmatrix} \rightarrow [-\ \text{Compact}]\ /\ +\ \text{Compact.Voy}\ \text{———}$$

Ce type de règle présente un inconvénient : le premier terme (l'archisegment) est trop abstrait pour être utilisé efficacement dans un domaine comme la reconnaissance dont les unités de référence sont des éléments concrets. En effet, le ou les traits qui définissent l'archisegment devraient être testés avant qu'on puisse attribuer à ce dernier les traits redondants donnés par la règle. Tâche difficile, quand on sait que les traits du modèle de Jakobson ne sont

P	P	T	K	B	D	G
i	⊕ 5.81	R 12	⊕ 9.5	R 5	+ 31.28	R 4
e				R 5		R 4
ɛ	R 8	⊕ 10.97	R 7	R 8,R 5, R 6	R 6	R 6 ,R 4
a	⊕ 10.69	⊕ 18.11	R 13	R 5, R 6	R 6	R 6, R 4, R 13
ɑ				R 5		R 4
ɔ	R 3, R 9	R 3, R 9	R 3	R 5, R 3	R 3	R 3, R 4
o	R 2, R 3	R 2, R 3	R 2, R 3	R 2, R 5, R 3	R 2, R 3	R 2, R 4, R 3
u	R 3	R 3	R 3	R 5, R 3	R 3	R 3, R 4
y	R 10, R 17	⊕ 7.64	R.7	R 10,R 5	R 17	R 4
ø				R 5		R 4
ɔe				R 5		R 4
ɑ̃	R 19	⊕ 10.93	R 7, R 13	R 5, R 18	R 19, R 18	R 13, R 4, R 18
ɔ̃	⊕ 3	R 23	R 23	R 5, R 18	R 23, R 18	R 23, R 4, R 18
ɛ̃		⊕ 5	R 7	R 5, R 18	R 18	R 4, R 18

TABLEAU 4 a : Distribution des consonnes et règles de neutralisation en Français Fondamental.

/ K /

	P	T	K	B	D	G
i	⊕ 3.68	⊕ 1	R 11	R 5	⊕ 6	R 4
e				R 5		R 4
ɛ	R 8	⊕ 25.24	R 7	R 8, R 5, R 6	R 6	R 6 ,R 4
a	⊕ 4.12	⊕ 16	R 13	R 5, R 6	R 6	R 6, R 4, R 13
ɑ				R 5		R 4
ɔ	R 9	R 9	⊕ 12.93	R 5	⊕ 7.58	R 4
o	R 16	⊕ 20.91	R 7	R 5.	R 16	R 4
u	⊕ 11.41	⊕ 31	R 7	R 5	⊕ 6.85	R 4
ø				R 5		R 4
œ				R 5		R 4
ɑ̃	R 19	⊕ 53.95	R 13	R 5	R 19	R 13, R 4
ɔ̃	R 21	⊕ 162.18	⊕ 24.71	R 21, R 5	⊕ 3	R 4
ɛ̃	R 1	R 1	R 1, R 7	R 1, R 5	R 1	R 1, R 4
y	+ 24	R 10	R 7	R 5	R 10	R 4

TABLEAU 4 b

/T/	P	T	K	B	D	G
i	⊕ 25.79	+ 270	⊕ 419.92	R 4', R 5	R 4'	+ 15.42
e				R 5		R 4
ɛ	R 8	⊕ 152	R 7	R 8, R 5, R 6	R 6	R 6, R 4
a	⊕ 12.8	R 14	⊕ 15.59	R 5, R 6	R 6, R 14	R 6, R 4
ɑ				R 5		R 4
ɔ	R 9	R 9	R 15	R 5	⊕ 50	R 15, R 4
o	R 2	R 2	R 2, R 7	R 2, R 5	R 2	R 2, R 4
u	R 16	⊕ 600	R 7	R 5	R 16	R 4
y	R 10	R 17	R 7	R 10, R 5, R17	⊕ 258	R 4
ø				R 5		R 4
œ				R 5		R 4
ɑ̃	⊕ 5.52	⊕ 105.18	R 15	5 20	⊕ 1	R 15, R 4
õ	R 22	R 22, R 14	R 22, R 7	⊕ 31.82	R 14	R 4
ɛ̃	R 1	R 1	R 1, R 7	R 1	R 1	R 1, R 4

TABLEAU 4 c

pas toujours représentés de façon univoque dans la chaîne sonore : comment tester le trait Consonantique dans la règle (1) par exemple quand on peut confondre /s/ et /i/ en reconnaissance de la parole ?

En revanche, si nous présentons la neutralisation sous la forme :

$$(4)\quad \begin{bmatrix} + \text{ Strident} \\ + \text{ Continu} \\ - \text{ Voisé} \\ \text{etc...} \end{bmatrix} \; / \; \left[\frac{+ \text{ Conson.}}{} \right] \rightarrow \begin{bmatrix} 0 \text{ Strident} \\ 0 \text{ Continu} \\ 0 \text{ Voisé} \end{bmatrix} / \# \; — \; \begin{bmatrix} + \text{ Conson.} \\ - \text{ Vocal.} \end{bmatrix}$$

nous savons que dans le contexte de la règle n'apparaît que le segment défini dans le premier terme, segment concret qui peut être identifié grâce aux traits par lesquels il se réalise.

II - Nombre de règles et économie

On suppose que les voyelles ont été préalablement reconnues. On peut, bien sûr, à l'aide d'un nombre de règles limité rendre compte des contraintes qui pèsent sur la distribution des voyelles dans le contexte Ocs — Ocsz. Mais ces règles devraient être précédées de la reconnaissance des consonnes, ce qui paraît impossible tant que les voyelles ne sont pas identifiées. On peut, nous verrons, éviter ce cercle vicieux. (voir p.12).

Les règles doivent donc définir les contraintes sur la distribution des unités les plus difficiles à reconnaître, les consonnes.

a) Pour le Français Général, 19 règles pour 753 mots rendent compte de 60% de possibilités de combinaison non réalisées.

b) Pour le Français Fondamental, 23 règles pour 100 mots rendent compte de plus de 80% de possibilités de combinaison non réalisées (tableau 4).

Ce nombre de règles est relativement limité quand on pense que pour le Français Fondamental en particulier on doit expliciter par des règles ad hoc des lacunes qui ne sont pas forcément systématiques. Les règles de neutralisation représentent donc une économie considérable pour la reconnaissance de la parole.

III - Projection des règles dans la structure acoustique

Notre première tâche a consisté à analyser les transitions des deuxième et troisième formants (fig. 2) et les

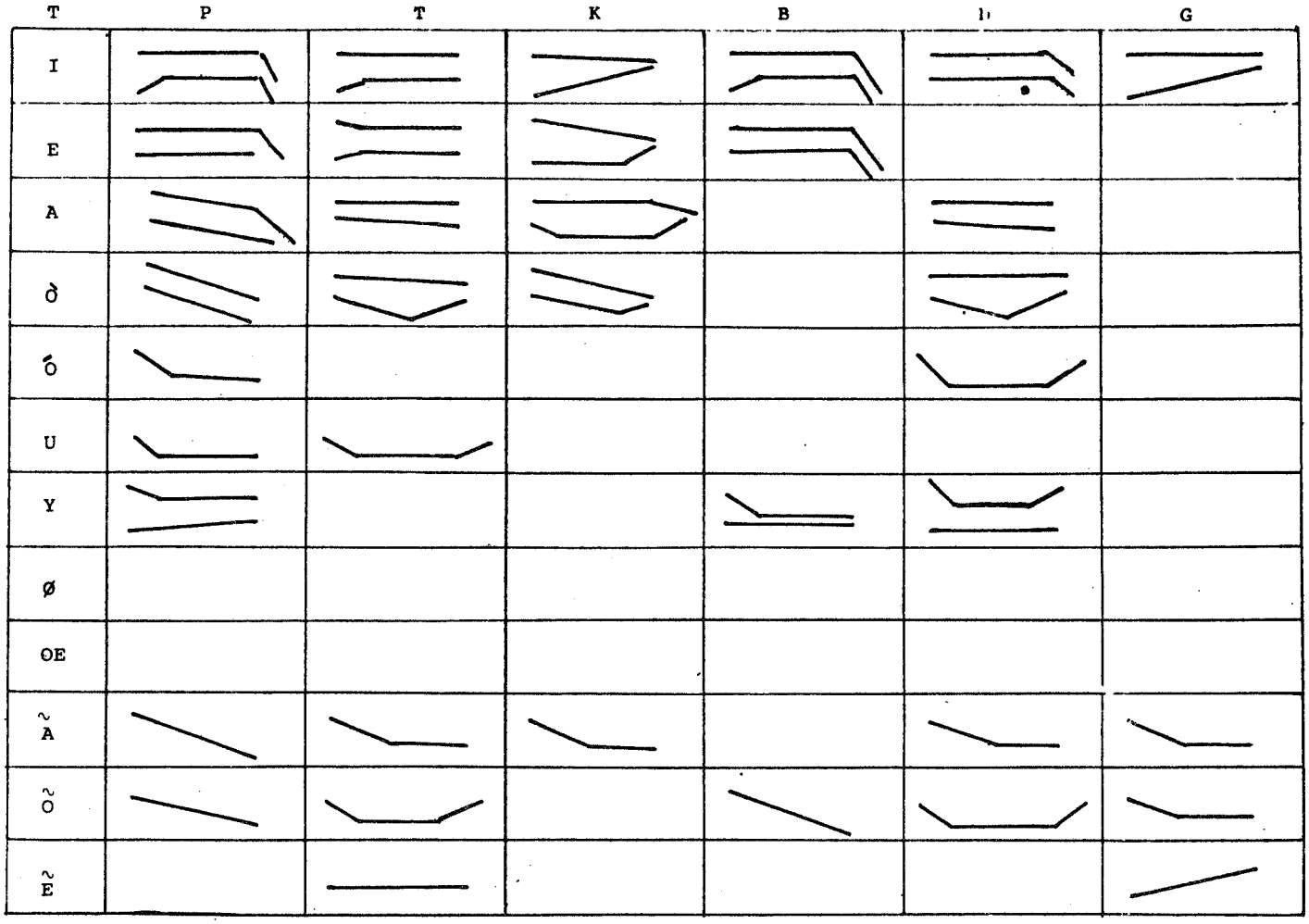


Figure 2a : transitions de F2 et F3 ou de F2 seul

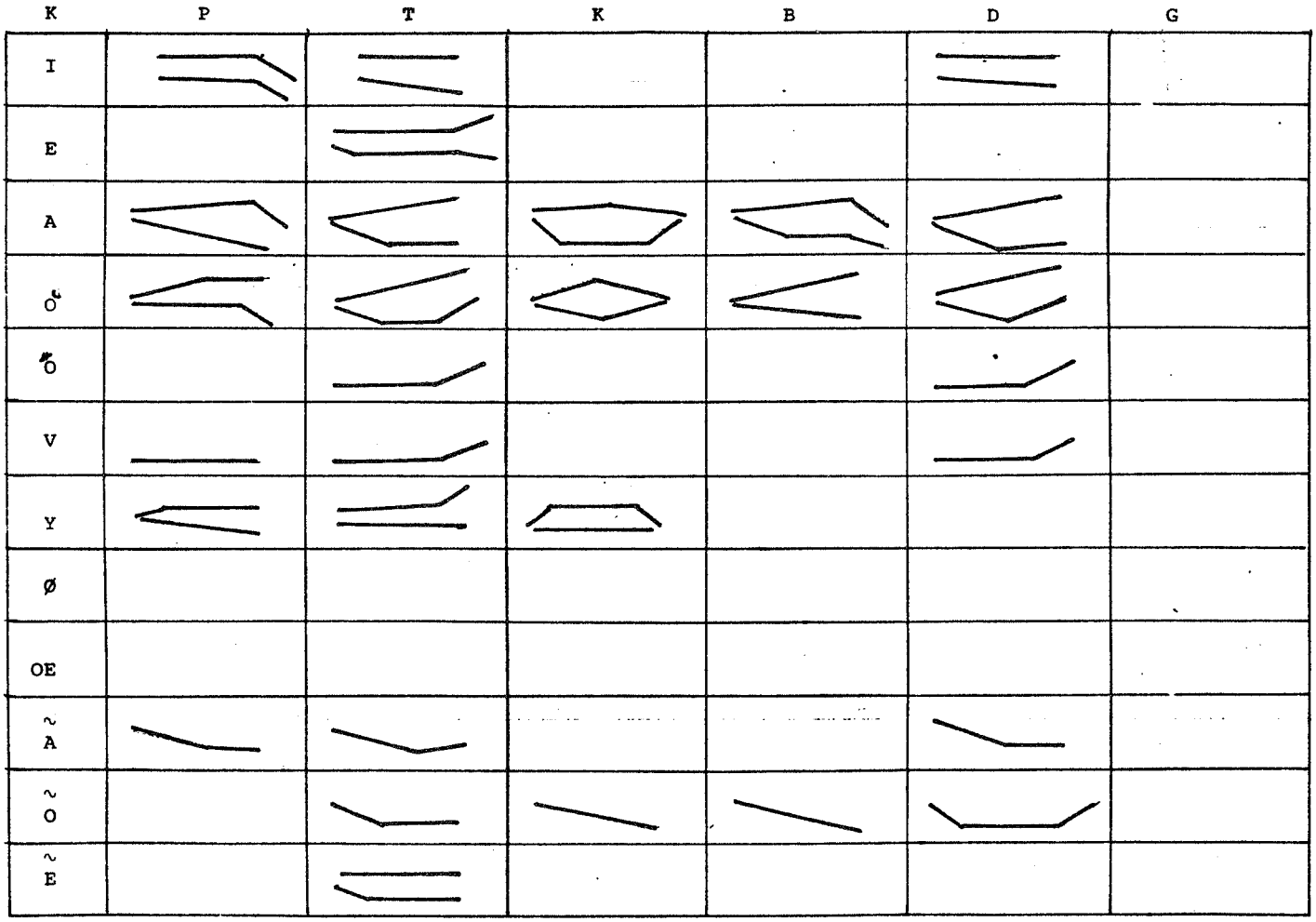


Figure 2 b : transitions de F2 et F3 ou de F2 seul

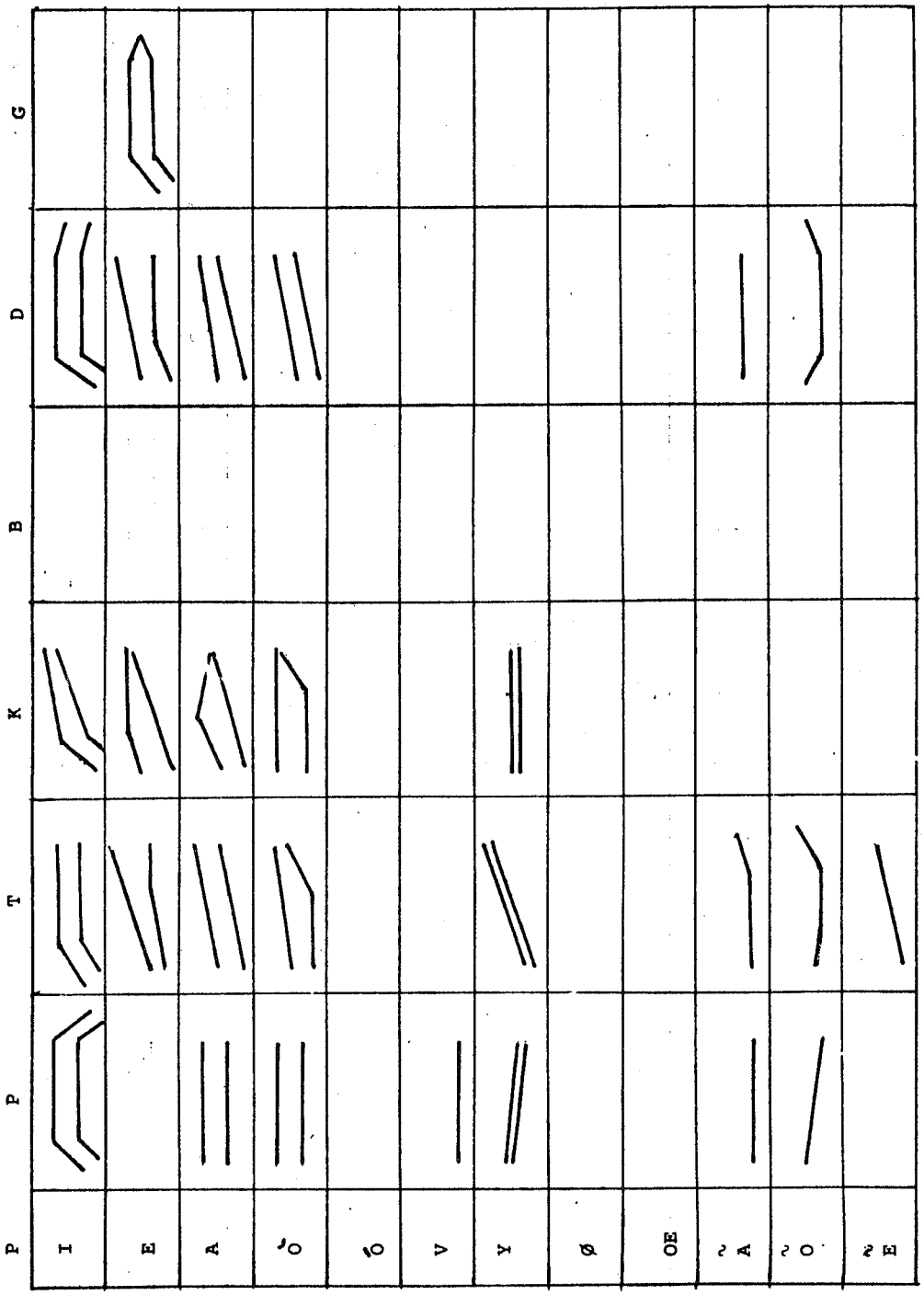


Figure 2c : transitions de F2 et F3 ou de F2 seul

autres indices acoustiques des lieux et modes d'articulation des occlusives dans le contexte choisi.

Toutes les combinaisons n'étant pas réalisées, les transitions formantiques ne sont pas toutes représentées sur la figure.

Les règles de contraintes sont projetées au niveau acoustique dans des algorithmes de reconnaissance qui utilisent les indices analysés.

1° Français Général :

Soit l'algorithme (fig. 3) pour la reconnaissance des consonnes dans le contexte : X — I — # . On essaie d'abord de reconnaître la consonne qui précède la voyelle, car elle est mieux articulée que la dernière.

Chaque test doit porter sur l'indice le plus marqué. Il se trouve que l'indice constant, le mieux défini - la transition négative de F2 et F3 - autorise une première dichotomie T, K/P. La consonne P peut donc être identifiée dès le premier test (13).

On a vu que la redondance, dans les règles, était traduite en termes de traits acoustiques. Chaque décision dans l'algorithme est accompagnée d'une étiquette qui représente le trait réalisé par l'indice choisi : la transition négative de F2 et de F3 autorise une dichotomie entre la consonne Grave : P et les consonnes Non Graves : T, K.

Les traits sont de la sorte introduits à partir d'indices concrets.

On a déjà tenté d'utiliser -sans grand succès- les traits dans la reconnaissance de la parole (14). Cet échec relatif provenait du fait qu'on attribuait aux traits un contenu acoustique trop abstrait. Notre procédure, qui lie la définition du trait à sa réalisation contextuelle devrait être plus efficace ; elle permet en outre d'intégrer les règles de neutralisation dans les algorithmes de reconnaissance.

Les indices par lesquels se réalisent les traits peuvent varier d'un algorithme à l'autre : les indices varient en effet avec le contexte.

Ainsi, les transitions, pour passer de T à la voyelle (fig. 2a) ont des formes qui diffèrent en fonction à la fois du timbre de la voyelle et de la nature de la consonne subséquente.

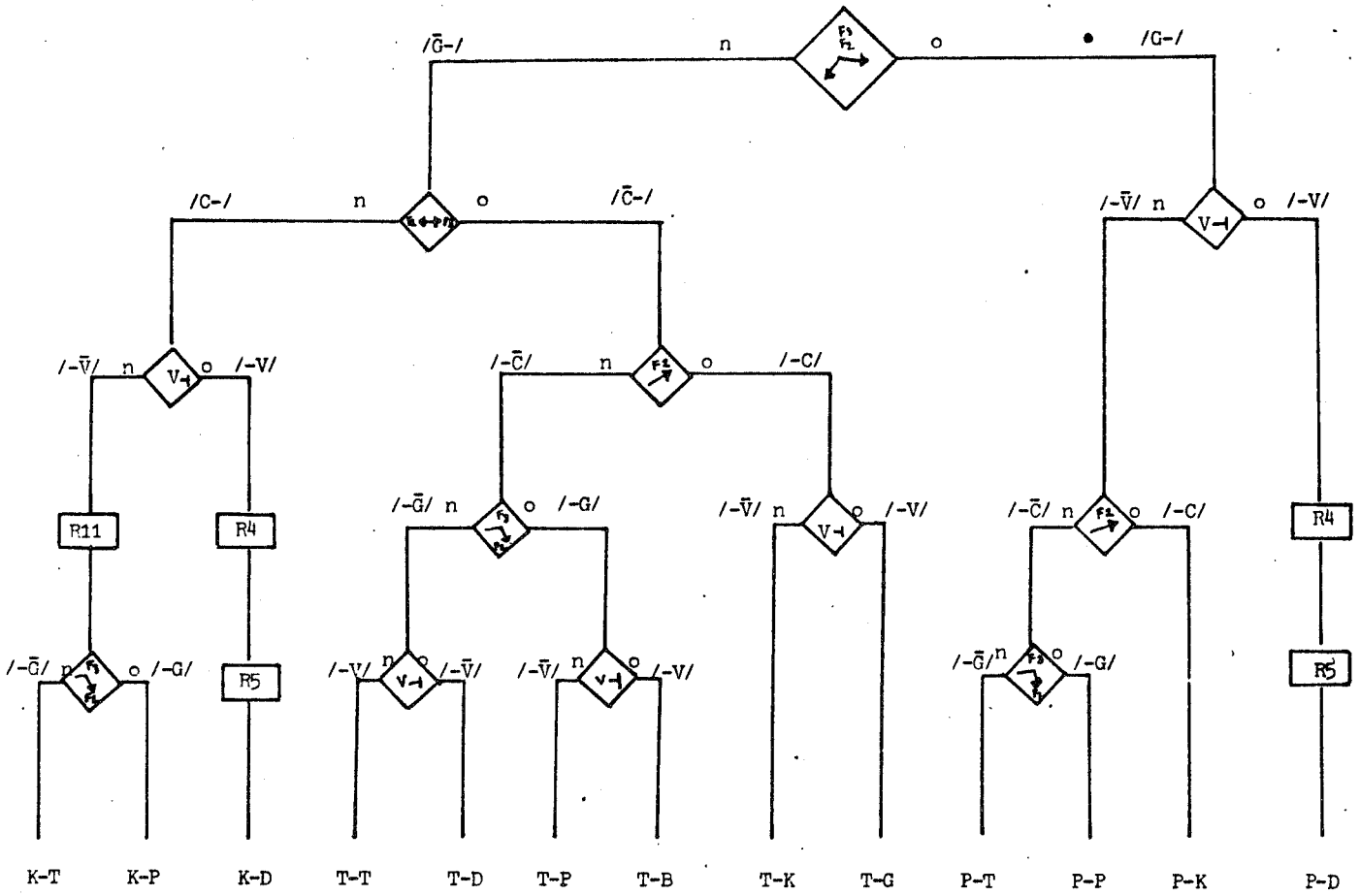


FIG. 3_ ALGORITHME DE RECONNAISSANCE DES CONSONNES DANS LE CONTEXTE I POUR LE FRANCAIS GENERAL

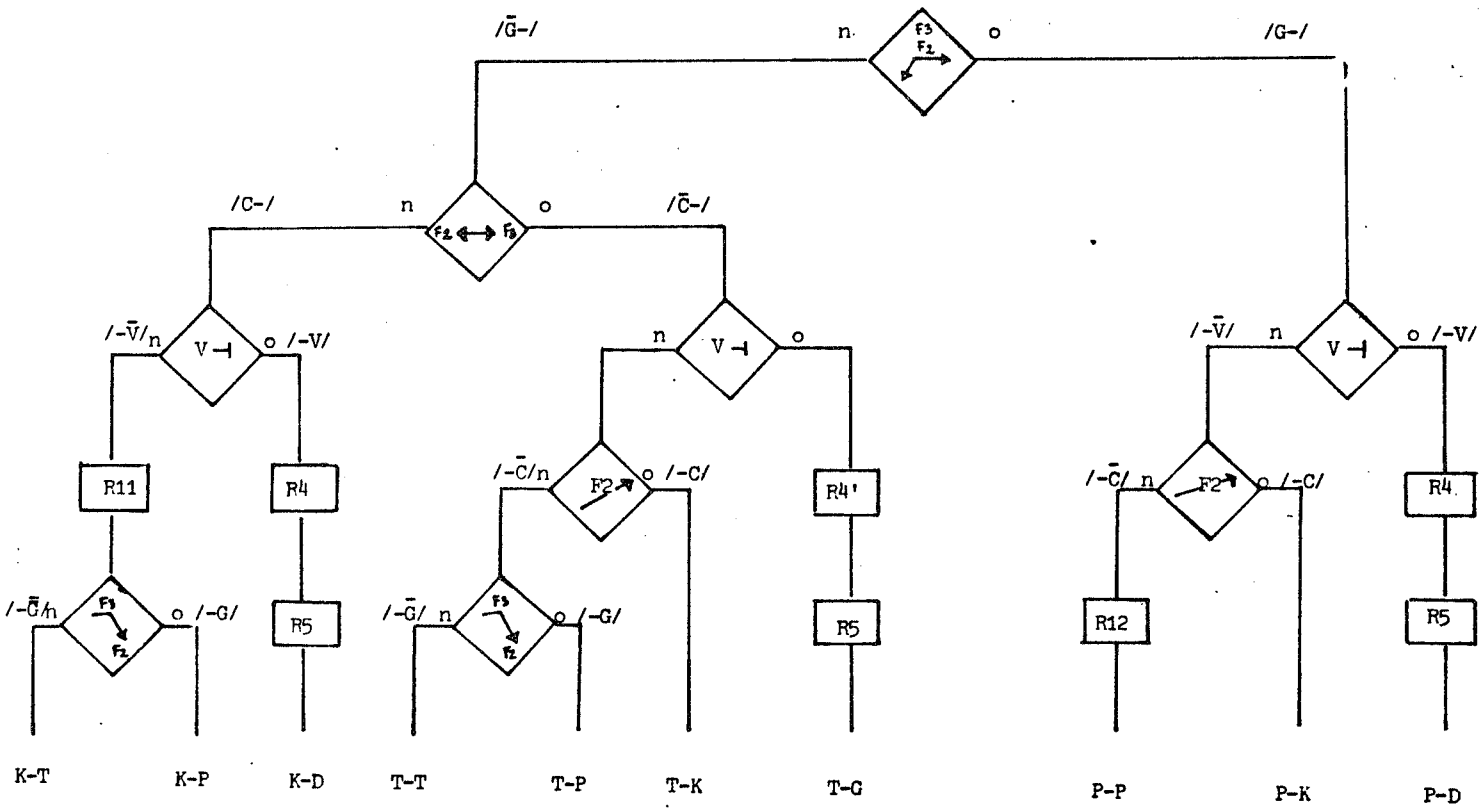


FIG. 4 - ALGORITHME DE RECONNAISSANCE POUR LE FRANCAIS FONDAMENTAL : VOYELLES I

En conséquence, la dichotomie T, K/P est assurée devant I par un test sur la transition négative de F2 et de F3, tandis qu'elle le sera devant Ò par un test sur la transition positive de F2. Dans le premier cas, on identifie, par la transition, le trait GRAVE, dans le deuxième cas on identifie le trait NON GRAVE ; devant Ò, en effet, la transition positive de F2 est le seul indice constant et il permet d'identifier T, K ; dans PÒ en revanche, la transition est tantôt droite, tantôt descendante.

Une fois que P a été éliminé, l'écartement progressif de F2 et de F3 devient un bon indice pour identifier le trait Non Compact. Le test sur la condition opposée pour identifier le trait Compact ne donne pas de résultats corrects car la convergence des formants sur les consonnes compactes K, G n'est pas toujours évidente.

Il convient de remarquer que l'écartement progressif de F2 et de F3 qui est un indice sûr pour l'identification de T dans le sous-ensemble {T, K}, ne l'est plus dans le sous-ensemble {P, K, T} car le degré d'écartement de F2 et de F3 sur P est ambigu. Cet exemple montre que l'identification des consonnes doit se faire dans un certain ordre imposé par leur structure acoustique.

Dans l'exemple choisi, une fois que la consonne initiale a été reconnue, s'il n'y a pas de contraintes après la voyelle -après TI, par exemple, où toutes les occlusives peuvent apparaître- on est obligé de tester chaque indice avec un taux d'erreur proportionnel au nombre de tests effectués. Après KI, en revanche, trois consonnes seulement sont attestées : KIT, KIP, KID. L'introduction des règles de neutralisation dans l'algorithme supprime certains tests et contribue par conséquent à diminuer le taux d'erreur.

Après le test sur l'indice de voisement, la reconnaissance de D est automatique car R4 et R5 neutralisent respectivement les traits Grave et Compact après KI si la consonne est voisée.

La règle R11 qui rend le trait Non Compact redondant, quand il est associé au trait Non Voisé, permet de supprimer le test sur l'indice de compacité. De ce fait, le test pour l'identification du trait GRAVE dans le sous-ensemble {P, T} devient plus efficient et plus économique : il ne porte en effet que sur la transition négative de F3 ; cet indice qui est nettement marqué, eût été ambigu en présence de K qui offre également une transition de F3 négative.

2° Le Français Fondamental

En Français Fondamental, le nombre des contraintes est très élevé ; l'intervention des règles de neutralisation doit diminuer de façon drastique le taux d'erreur (fig. 4).

Quatre règles seulement permettent d'identifier de façon automatique la consonne sonore après la voyelle I lorsque l'indice de voisement a été reconnu. Après KI et PI où seules deux occlusives sourdes sont possibles au lieu de trois (respectivement T/P et P/K), les règles R5 et R4 suppriment les tests sur l'indice de compacité dans un cas et sur l'indice de gravité dans l'autre. On remarque d'autre part que l'indice qui permet de séparer P/K après PI devient plus performant car les transitions de F2 de ces deux consonnes sont opposées.

En Français Général où les trois consonnes P, T, K, sont réalisées après PI, il fallait tester une transition intermédiaire, caractéristique de T, dont la présence accroissait le risque d'erreur puisque dans ce cas l'indice de compacité était relativement moins marqué.

Si on choisit d'autres voyelles que I, les contraintes sont encore plus fortes, notamment pour le Français Fondamental (fig. 5). Dans le contexte de È, par exemple, deux tests et 3 règles devraient permettre une identification correcte des consonnes ; seules trois suites sont en effet possibles avec la voyelle E dont la fréquence d'occurrence dans la chaîne est pourtant très élevée (7,2% des voyelles, Rang : 3) : KÈT, TÈT, PÈT.

Dans le cas où toutes les possibilités de combinaison Ocs V Ocsz sont réalisées dans le noyau accentuel, un algorithme nécessite 5 bits d'information pour reconnaître seulement 18 formes différentes ; l'introduction des contraintes ramène ce taux à 1,8 bits pour la reconnaissance de 4,2 noyaux accentuels en moyenne (tableau 5).

Les contraintes phonologiques réduisent la quantité d'information nécessaire, et diminuent en même temps le risque d'erreurs proportionnel au nombre de tests ; leur présence permet aux algorithmes de fonctionner -avec le minimum d'information- comme des codes à détection d'erreurs.

IV - Reconnaissance des voyelles

L'identification des consonnes à l'aide des règles de neutralisation suppose la reconnaissance préalable des voyelles. Celle-ci doit être facilitée, nous l'avons dit, par le fait que les voyelles du noyau accentuel à la fin du groupe intonatif sont longues. La longueur et la proéminence accen-

FIG. 5 - ALGORITHME DE RECONNAISSANCE DES CONSONNES DANS LE CONTEXTE È POUR LE FRANCAIS FONDAMENTAL

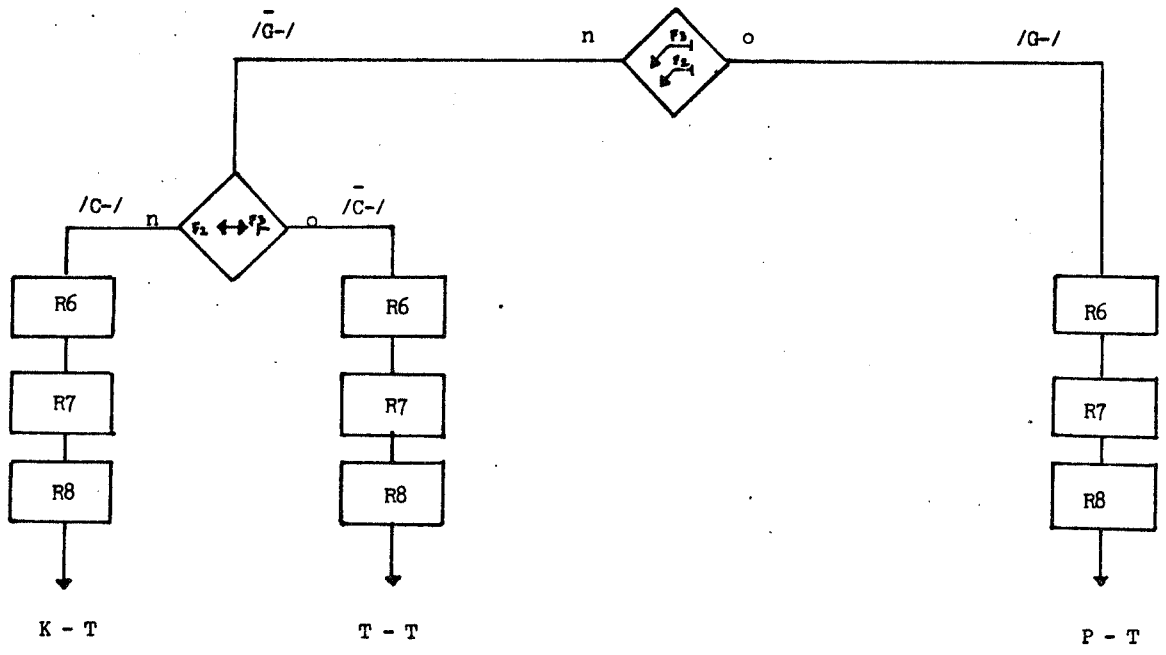


TABLEAU 5

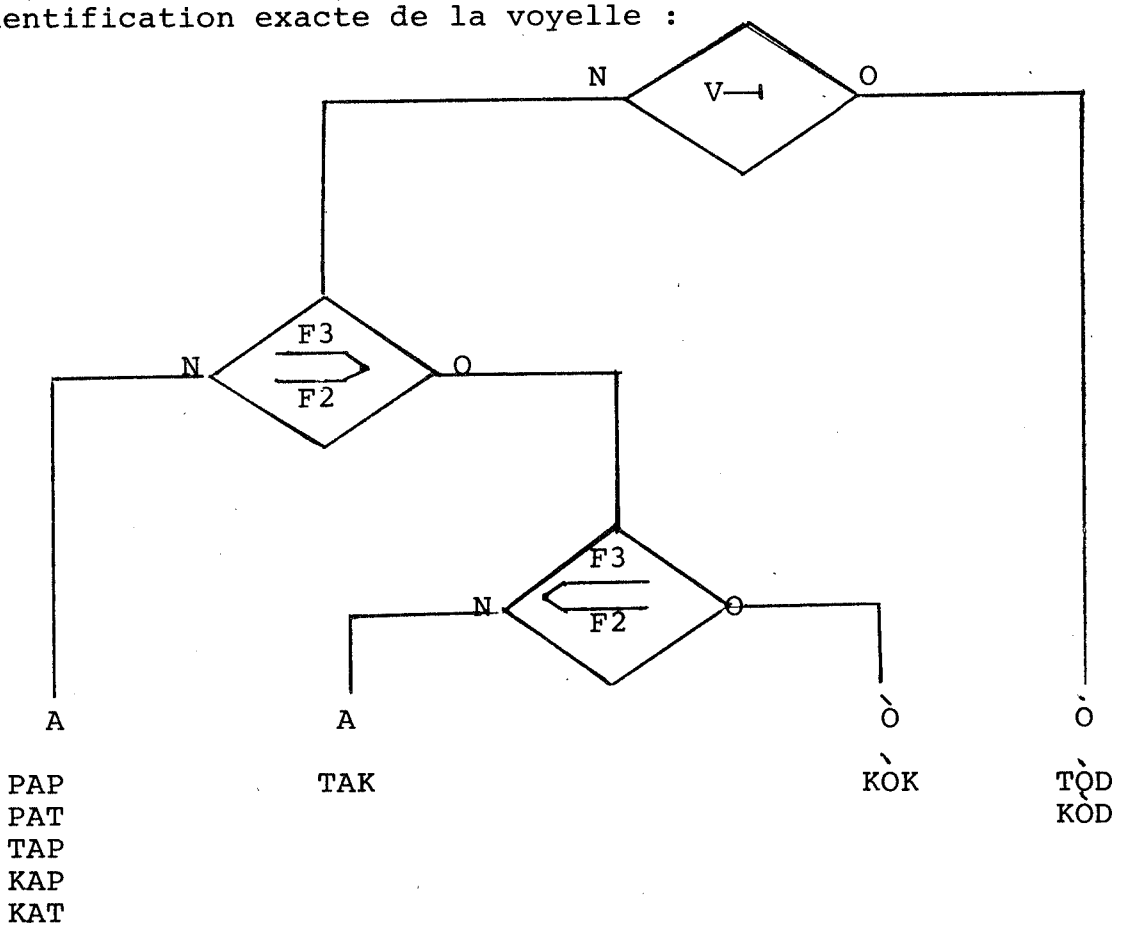
Contexte V.	N. de noyaux	N de Bits
I	10	3.2
A	6	2.7
A	5	2.5
O	5	2.5
U	4	2
Y	4	2
E	3	1.7
O	3	1.7
O	1	O
E	1	O
Moyenne	4.2	1.84

Taux d'information dans les algorithmes compte-tenu des contraintes phonologiques

tuelle diminuent l'effet du contexte au centre de la voyelle.

En conséquence, les zones de recouvrement pour un même locuteur devraient être réduites au minimum. Pour le sujet M. R., le seul risque de confusion se situe entre È, A, Ò. Etant donné que ces voyelles apparaissent partiellement dans des contextes différents, un test complémentaire sur les consonnes peut permettre de lever l'ambiguïté.

Soient les deux candidats possibles A et Ò qui entrent dans les noyaux accentuels PAP, PAT, TAP, TAK, KAP, KAT et TÒD, KÒK, KÒD, l'algorithme suivant devrait permettre une identification exacte de la voyelle :



On utilise ainsi certaines contraintes distributionnelles pour décider de la nature de la voyelle. Cette méthode est intéressante en ce sens qu'on est obligé de parcourir le contexte consonantique qu'on identifie en même temps que la voyelle.

V - Les contraintes sur les syllabes non accentuées

Les contraintes sur les voyelles et les consonnes sont bien plus fortes en syllabe atone qu'en syllabe accentuée.

La difficulté en français pour définir ces contraintes -dans la perspective d'une reconnaissance par phrase- provient du fait qu'à l'intérieur du groupe intonatif les syllabes accentuées se désaccentuent et se trouvent au même niveau de proéminence que les syllabes atones :

Ex. : Toutes les boutiques /ont été pillées

Dans cette phrase, la syllabe accentuable des mots soulignés a une structure phonologique qui relève des lois de distribution que nous avons définies pour les noyaux accentuels. Mais aucun indice acoustique ne permet de distinguer ces syllabes des syllabes atones (bou dans boutique) dont la forme est régie par d'autres lois.

Dans ces conditions, comment définir les contraintes dans la phrase ?

La meilleure solution consisterait à émettre une hypothèse qui permettrait 1) une segmentation en mots, 2) l'utilisation des contraintes spécifiques de chaque position du mot.

Mais il semble que cette procédure ne puisse pas être mise en oeuvre si l'auditeur n'a pas d'autres points d'ancrage que le noyau accentuel.

Une méthode élégante, en complément à l'étape acoustique de la reconnaissance, pour pallier cette difficulté, est l'étude des régularités sur les transitions d'état.

A l'intérieur du groupe intonatif les voyelles sont brèves. Fortement influencées par l'entourage consonantique, elles n'ont pas le temps d'atteindre leur régime permanent. Les zones de dispersion et de recouvrement sont alors très larges. Par ailleurs, dans cette position, les consonnes sont le plus souvent sous-articulées. Il est vain, à notre avis, d'attendre du système de détection acoustique qu'il puisse identifier chacune des voyelles et des consonnes dans les syllabes inaccentuées.

Etant donné les fortes contraintes phonologiques et syntaxiques qui pèsent à l'intérieur du groupe intonatif, la recherche des voyelles et leur réduction à trois catégories devraient suffire à identifier la chaîne inaccentuée ou du moins à y créer des points d'appui pour une hypothèse syntaxique.

Un test sur les traits Compact et Aigu permet pour le locuteur M. R. et pour le corpus considéré, de réunir les voyelles dans les classes (15) :

- C (Compactes) : A, $\tilde{A}$
- $\overline{C}$ A (Non Compactes + Aiguës) : I, É, È, Y, Ø, OE, $\tilde{E}$
- $\overline{C}$ $\overline{A}$ (Non Compactes + Non Aiguës) : U, Ó, Ò, $\tilde{O}$.

Le calcul des transitions d'état, analogues à une dérivée première, à partir de la combinaison de ces deux traits conduit, dans les conditions de notre corpus, à la reconnaissance de la chaîne inaccentuée sans avoir à identifier les consonnes.

Soit :

$C = a$

$\overline{C}\ \overline{A} = \overline{a}$

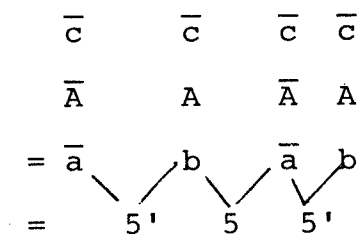
$\overline{C}\ A = b$

et les transitions :

aa = 2	$b\underline{a}$ = 1'
$a\underline{b}$ = 1	$\underline{ba}$ = 5
$a\overline{a}$ = 3	$\underline{a}a$ = 6
bb = 4	$\overline{a}a$ = 3'
$\overline{a}b$ = 5'	

Nous obtenons l'analyse :

Toutes les boutiques



Soit TIK le noyau accentuel reconnu à la fin du premier groupe intonatif ; soit x le nombre des voyelles identifiées dans le groupe grâce aux pics d'intensité par exemple.

A quoi correspond la chaîne ?

Etant donné les structures syntaxiques admises, le noyau TIK appartient à un nom à choisir dans une liste de huit

items (tableau la). Chacun de ces noms peut rentrer dans l'une des suites :

la — ; les — ; une — ;
des — ; tou(s,tes) les — ;
les cinquante —.

Au total 48 suites différentes et une combinatoire relativement complexe. Le calcul des transitions d'état pour le mot boutique précédé du Déterminant donne les résultats suivants :

1 la boutique	= 35'
2 les boutiques	= 55'
3 une boutique	= 55'
4 des boutiques	= 55'
5 toutes les boutiques	= 5'55'
6 les 50 boutiques	= 41'35'

Ce calcul fait apparaître une confusion possible entre les suites 2, 3 et 4. Le même calcul pour chacun des huit mots de cette classe montre que la seule confusion possible a lieu entre politique et domestique.

Toutes les autres suites sont distinctes. L'ambiguïté, on le voit, est tout de même réduite au minimum :

EX. Toutes les	boutiques	= 5'55'
" "	critiques	= 5'44
" "	pratiques	= 5'1'1
" "	domestiques	= 5'55'4
" "	politiques	= 5'55'4
" "	gymnastiques	= 5'41'1

A cette étape de la reconnaissance, on fait intervenir l'Analyseur syntaxique qui construit une hypothèse à partir des mêmes tests sur chacun des groupes intonatifs afin de lever l'ambiguïté.

La procédure d'analyse par les transitions d'état pourrait être appliquée à des suites plus complexes. Elle revient à réduire la chaîne inaccentuée à un système de trois voyelles dont on calcule la combinatoire ; elle annule donc l'effet du recouvrement des voyelles.

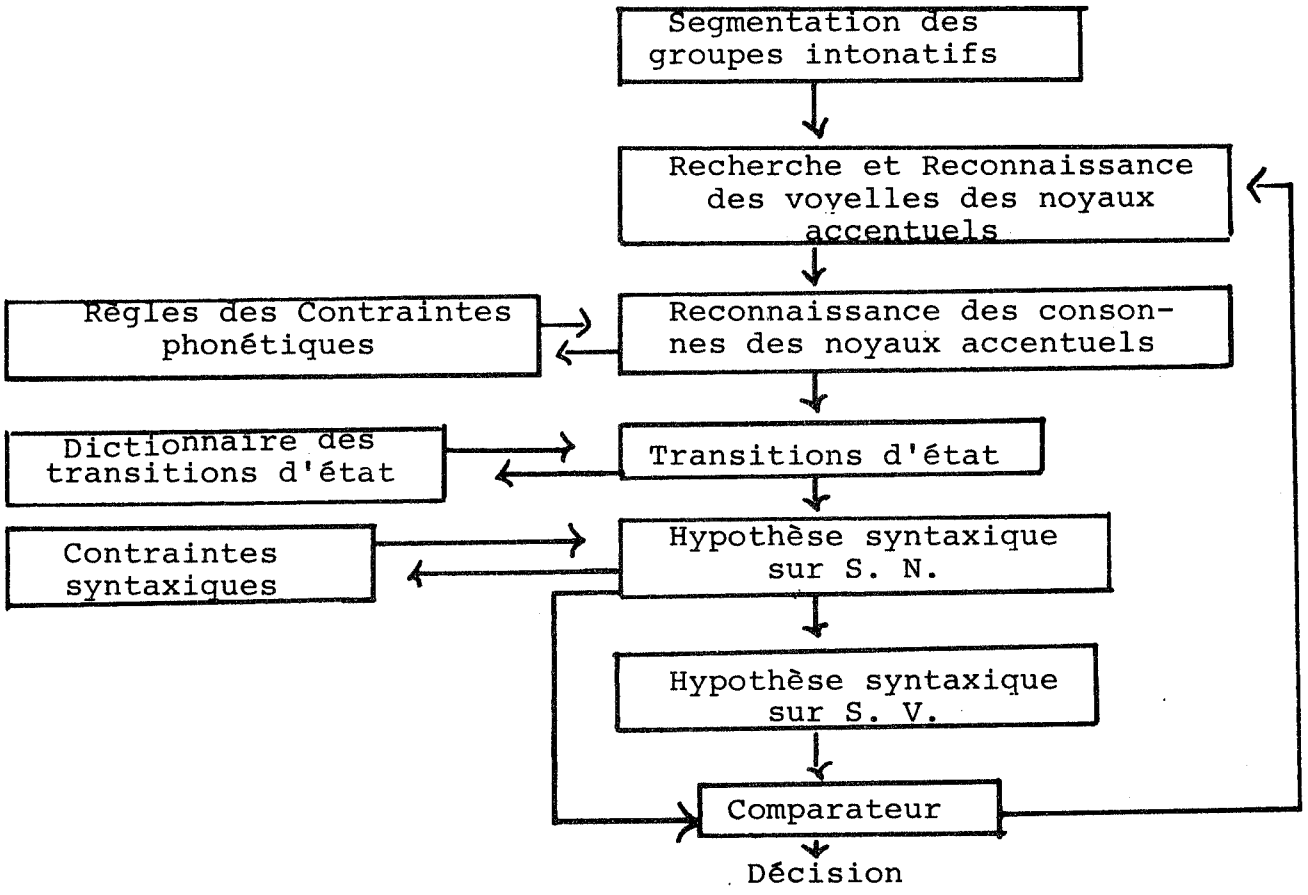
Le modèle d'application des contraintes phonologiques que nous proposons ne se substitue pas, comme nous l'avons dit, à l'étape acoustique de la reconnaissance. Il doit la compléter et la rendre plus efficiente.

Conçu pour fonctionner dans un système de reconnaissance par poursuite de formants, nous avons, dans les algorithmes fait référence aux transitions de formants comme

indices d'identification des consonnes.

Cette référence n'est qu'une façon -propre au phonéticien- de représenter la structure phonique des segments de parole ; elle n'implique aucun jugement quant à l'efficacité des différentes méthodes de reconnaissance acoustique. Avec les adaptations nécessaires, ce modèle peut se prêter à tout système susceptible de fournir les paramètres d'identification des consonnes.

L'organigramme suivant résume les étapes de la procédure dont nous venons de donner un rapide aperçu :



BIBLIOGRAPHIE

(1) J.S. Liénard, *Analyse, synthèse et reconnaissance automatique de la parole*, Thèse de Doctorat d'Etat. 1972 .

(2) M. Mlouka, *Reconnaissance automatique de la parole : une expérience de reconnaissance par mots*, Thèse de Doctorat de 3e cycle. 1974.

(3) R. De Mori, A descriptive technique for automatic speech recognition, *IEEE Transactions*, vol. Av. 21, n° 2, 1973 P.89 - 100 ; Research for a grammar of speech, *Preprints of the Speech communication seminar*, Stockolm 1974, vol. 3, P. 233-238.
W.J. Hess, A pitch-synchronous, digital feature extraction system for phonemic recognition of speech, *ibid* P. 201-213.

(4) J.P. Haton, Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole, Thèse 1974.

P. Vicens, *Aspects of speech recognition by computers*, Thèse 1969.

(5) C'est la méthode utilisée par Carré pour rechercher les formants et les transitions sur notre corpus.

(6) A. Di Cristo, Recherches sur la structuration prosodique de la phrase française, communication présentée aux *Journées d'Etudes de Toulouse.*

(7) G. Faure, Contribution à l'étude du statut phonologique des structures prosodématiques, *Studia Phonetica*,3, P. 93-100.

(8) M. Rossi, L'intonation dans les phrases transformées par permutation, *Linguistics*, 103, 1973, P. 64-94.

(9) A.Juilland,*Dictionnaire inverse de la langue française*, Mouton, La Haye. 1965.
Nous avons préféré cet ouvrage à celui de Gougenheim car il constitue un outil approprié à l'analyse distributionnelle du noyau accentuel. Il comporte en outre des renseignements statistiques fondamentaux (rang, fréquence d'occurrence, valeur d'usage).

(10) Vicens, *op.cit.*,P. 94.

(11) K. Stevens, Segments, features and analysis by synthesis, in J.F. Kavanagh and I.G. Mattingly, *Language by ear and by eye*, Cambridge, M.I.T. Press, 1972.

(12) Broadbent, *Perception and Communication*, Pergamon Press, New York, 3e éd. 1969.

(13) Il est souhaitable de faire porter le test sur des indices complémentaires. Nous simplifions à dessein la présentation car notre propos est de montrer les liens qui existent entre les règles de contraintes et la réalité acoustique.

(14) Voir par exemple, J. Wiren, H.L. Stubbs, Electronic binary system for phonemic classification, J. A. S. A. 28, 8, 1956, P. 1 082- 1 091.

(15) Une méthode simple consisterait à faire le rapport de F1 et de F2 sur un seuil respectif S1 et S2 choisi en fonction du locuteur. Mais d'autres méthodes sont également possibles, voir G. Fant, *Acoustic theory of speech production*, Mouton, La Haye, 2e éd. 1970, P. 215-225 ; Distinctive features and phonetic dimensions, STL.Q.P.S.R. 2-3, 1969.
En ce qui concerne le modèle des transitions d'état, voir F.C. Hennie, *Finite - State models for logical machines*, John Wiley, New York, 1968, Ch. I, II.

THEME 2

ANALYSE ET PERCEPTION

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

LE VECTEUR DELTA COMME INDICE PHONETIQUE ET SON APPLICATION A LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

J.J. MARIANI, J.S. LIENARD, G. RENARD

L.I.M.S.I. (C.N.R.S.) - B.P. 30 - 91406 ORSAY

RESUME

La notion de diphonème, utilisée jusqu'à présent en synthèse, est étendue à l'analyse phonétique et à la reconnaissance automatique de la parole. Les transitions spectrales sont caractérisées par un ensemble de valeurs variant entre -1 et +1, formant le vecteur DELTA. Celui-ci représente l'évolution spectrale dans un intervalle de quelques dizaines de millisecondes. Sa forme reflète conjointement l'évolution en amplitude et l'évolution en fréquence des zones formantiques. L'application de cette notion à la reconnaissance automatique permet d'utiliser une méthode particulière de reconstitution des phrases continues, à partir de "jalons" reconnus avec une forte probabilité.

SUMMARY

The notion of diphone still only used in speech synthesis is extended to phonetic analysis and to automatic speech recognition. Spectral transitions are characterized by a set of values ranging from -1 to +1 which forms the DELTA vector. This vector represents the spectral evolution during a few centiseconds interval. Its envelope reflects the speech formant variations in both magnitude and frequency. The application of this notion to speech recognition permits the use of a particular method to reconstruct continuous sentences from defined marks which are recognized with a great probability.

LE VECTEUR DELTA COMME INDICE PHONETIQUE ET SON APPLICATION A LA RECONNAISSANCE DE LA PAROLE

J.J. MARIANI - J.S. LIENARD - G. RENARD

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (L.I.M.S.I. - C.N.R.S.)

B.P. 30 - 91406 ORSAY

I - INTRODUCTION

L'expérience acquise par notre équipe en synthèse et en reconnaissance de la parole (bib 1, 2) nous a amenés à poser l'hypothèse suivante : l'information phonétique de la parole est tout entière contenue dans le "squelette phonétique", forme (Gestalt) composée de l'ensemble des bruits d'explosion ou de friction, des traces de formants, et plus généralement de tous les éléments acoustiques reflétant le mouvement du conduit vocal. Ce squelette, perçu habituellement de manière globale, peut cependant, lorsque la parole a été articulée soigneusement (intelligibilité au niveau phonétique), être décomposé en éléments physiquement distincts, décelés par autocorrélation du spectrogramme selon un délai de quelques dizaines de millisecondes (bib 2, 3). Le message est donc segmentable objectivement en éléments transitoires ("phonatomes", "diphonèmes"), dont la pertinence est reconnue en synthèse, mais qui n'ont guère été utilisés, jusqu'à présent, dans les domaines de la perception phonétique ou de la reconnaissance automatique.

Pour décrire le spectrogramme en termes d'évolution temporelle, nous avons imaginé la notion de "vecteur delta". Nous nous proposons, après avoir défini cette grandeur, de montrer comment elle reflète certains traits phonétiques dégagés depuis longtemps par les phonéticiens, et d'examiner son application à la reconnaissance automatique de la parole (*).

II - DEFINITION ET PROPRIETES DU VECTEUR DELTA

II - 1. Calcul des composantes du vecteur DELTA

Nous admettrons dans la suite que le spectrogramme A est numérisé sur K canaux (sorties de filtres redressées et intégrées, voies FFT, etc.) et N échantillons (pas de temps $\tau \simeq 10$ ms). L'élément a_{ij}

(*) Ce travail a été en partie effectué dans le cadre du contrat DRME n° 73.164

représente l'amplitude spectrale à l'instant i ($i = [1, N]$) et à la fréquence j ($j = [1, K]$). Le spectre instantané (colonne i du tableau A) est le vecteur $S_i = [a_{i1}, \ldots, a_{ij}, \ldots, a_{iK}]$.
Il s'agit d'associer à l'instant i l'évolution du spectre dans une fenêtre de θ ms entourant cet instant.

Le terme d'"évolution" doit se référer à la psychophysiologie de la perception ; mais nous ne savons pas grand'chose de la perception de l'intensité sonore dans des conditions aussi complexes que celles du message parlé. En première approximation nous pourrions adopter une notation quasi-logarithmique : quand, sur l'intervalle θ, l'intensité spectrale passera du simple au double, l'évolution prendra une valeur positive x ; inversement une valeur -x caractérisera une valeur moitié, -2x une valeur égale au quart, etc.

Mais la progression logarithmique ne nous satisfait pas entièrement pour les écarts d'intensité importants : il est bien connu qu'un son de 80 dB ne paraît pas deux fois plus fort qu'un son de 40 dB lorsqu'on peut effectuer librement la comparaison, et que pour obtenir une sensation 10 fois plus forte il suffit d'augmenter le niveau de 30 dB environ (au sujet du "paradoxe de l'intensité fractionnée", voir bib 4). D'autre part, lorsque l'intensité spectrale est nulle ou très faible, on peut être conduit à des formes indéterminées, alors que Δ doit dans ce cas indiquer une variation nulle. Ces raisons, ajoutées à la simple commodité du calcul, nous ont fait adopter l'expression suivante :

$$\delta_{ij} = \frac{a_{(i+\frac{\theta}{2},j)} - a_{(i-\frac{\theta}{2},j)}}{a_{(i+\frac{\theta}{2},j)} + a_{(i-\frac{\theta}{2},j)} + \beta_j}$$

β_j représente le bruit de fond moyen dans la voie j observé en l'absence de signal utile.

La valeur de δ_{ij} est comprise entre +1 et -1. La figure 1 montre la variation de δ_{ij} en fonction du rapport $r = a_{(i+\frac{\theta}{2},j)} / a_{(i-\frac{\theta}{2},j)}$ pour deux valeurs différentes de l'émergence par rapport au bruit de fond. On voit ainsi que la variation de δ_{ij}, quasi-logarithmique pour $r < 10$, est asymptotique au-delà, et que δ_{ij} tend vers zéro quand le signal dans la voie j tombe au dessous du seuil β_j.

II - 2. Propriétés fondamentales du vecteur DELTA

Nous appelons vecteur delta Δ_i à l'instant i l'ensemble des composantes δ_{ij}.
Ce vecteur possède quelques propriétés remarquables :

- insensibilité au niveau moyen du signal : des séquences A et A' telles que A' = kA fournissent des vecteurs DELTA identiques. Il est donc inutile d'effectuer une normalisation en niveau ou une compression du niveau instantané du signal.
- insensibilité au timbre de la parole : certaines voix sont plus riches que d'autres dans la zone aiguë du spectre ; un même locuteur, parlant plus ou moins fort, produit un spectre moyen diffé-

rent dans son équilibre entre les diverses zones du spectre. Ces phénomènes disparaissent dans le vecteur Δ , qui traduit l'évolution spectrale indépendamment du niveau moyen dans chaque voie

- prise en compte de la composition spectrale du bruit de fond : les variations spectrales sont atténuées lorsque la valeur du signal devient du même ordre de grandeur que le bruit de fond dans la zone fréquentielle considérée.

La figure 2 représente un spectrogramme schématique, ainsi que les spectres et les vecteurs DELTA à divers instants. On voit que la forme du vecteur DELTA fournit plusieurs informations :

- l'évolution en amplitude spectrale et caractérisée par son sens, positif(amplitude croissante) ou négatif (amplitude décroissante), et par son intensité. Un spectrogramme évoluant uniquement en amplitude fournit des pôles simples.
- l'évolution en fréquence est caractérisée par deux pôles opposés, séparés par un passage à zéro. Une montée en fréquence se traduit par une pente positive au passage à zéro.

De plus, l'aire comprise entre le vecteur DELTA et l'axe fréquentiel représente à chaque instant l'évolution spectrale dans son ensemble. On peut donc, en calculant

$$\frac{1}{K} \sum_{j=1}^{K} |\delta_{i,j}|$$

obtenir une fonction d'instabilité spectrale, comprise entre 0 et 1, dont les maxima marquent les moments de transition et les minima les moments de stabilité (fig. 2).

III - LE VECTEUR DELTA COMME INDICE PHONETIQUE

III - 1. Etude à partir des spectrogrammes schématisés pour la synthèse

Nous allons considérer le vecteur delta obtenu lors d'un maximum d'instabilité et rechercher les corrélations avec les indices acoustiques connus, à partir de représentations schématiques empruntées à P. DELATTRE (bib 5).

a) Voyelles isolées

Nous admettons qu'une voyelle n'est pas perçue lors de son état stable, mais seulement lors de son établissement ou de son extinction. Donc une voyelle isolée est caractérisée par les transitions silence-voyelle et voyelle-silence. Il est clair que les maxima du vecteur DELTA correspondent dans ce cas aux maxima du spectre (fig. 3a), et l'on peut y retrouver la notion de formant.

b) Diphonèmes fricative-voyelle
Prenons l'exemple de | sɛ | (fig. 3b) le vecteur DELTA possède les caractéristiques suivantes :
- maxima positifs correspondant à l'apparition des formants de la voyelle suivante
- zone négative aiguë correspondant à la disparition du bruit (turbulence) de la fricative
- au voisinage des maxima positifs on trouve des passages à zéro qui sont dus aux glissements de fréquence des formants.

c) Diphonèmes liquide-voyelle
Soit | lɛ | (fig. 3c). Les maxima positifs correspondent aux formants de la voyelle ; ils sont accompagnés de maxima négatifs de moindre amplitude. Le sens des transitions formantiques est indiqué par la pente autour des zéros du vecteur delta.

d) Diphonèmes nasale-voyelle
Prenons l'exemple de | nɛ | (fig. 3d) le voisement apparaît ici comme une zone négative dans les premières valeurs de DELTA. La discontinuité des liens fournit des maxima positifs et négatifs bien distincts tout au long de l'axe de fréquence. Cependant deux phénomènes sont confondus : la transition d'amplitude entre le 2e formant de nasalité et le 3e formant vocalique, et la transition fréquentielle de ce dernier.

e) Diphonèmes plosive-voyelle
Prenons | dɛ | (fig. 3e) le vecteur DELTA reflète bien les évolutions en amplitude, mais nous rencontrons un problème voisin du précédent : la transition d'amplitude masque en partie l'évolution fréquentielle. Pour résoudre ce problème nous sommes amenés à considérer séparément l'évolution en amplitude et l'évolution en fréquence, qui n'ont pas lieu exactement au même instant. Ce point fait l'objet de nos études actuelles.

III - 2. Etude à partir de spectrogrammes réels

La phrase "Le gamin est parti à l'école" a été analysée au moyen d'un banc de 10 filtres échelonnés en fréquence de manière logarithmique, et échantillonnés toutes les 8 ms. Le vecteur DELTA a été calculé pour chaque échantillon, sur un intervalle θ de 20 ms. La figure 4 représente les vecteurs DELTA obtenus aux instants d'instabilité maxima (les segmentations excédentaires ont été conservées). On remarquera les points suivants :
- les diphonèmes se répartissent très souvent en deux classes : ceux pour lesquels le vecteur DELTA est entièrement positif et ceux pour lesquels il est entièrement négatif. On peut rapprocher ceci de la notion d'ouverture-fermeture en phonétique articulatoire.
- l'évolution en amplitude masque en général l'évolution en fréquence, si l'on caractérise celle-ci par les zéros de DELTA. Ceci tient

à la quantification très grossière adoptée en fréquence. Cependant l'évolution en fréquence est décelable dans la dissymétrie du vecteur DELTA au voisinage de ses maxima.
- la première composante de DELTA possède toujours une forte valeur lors de l'attaque des sons non voisés, et une valeur plus faible, ou nulle, pour les sons voisés (sauf /ga/).

Tous ces éléments nous incitent à penser que le vecteur DELTA représente correctement la transition phonétique. Nous allons maintenant appliquer cette notion à la reconnaissance automatique de la parole.

IV - APPLICATION DU VECTEUR DELTA A LA RECONNAISSANCE AUTOMATIQUE

IV - 1 . Reconnaissance par mots (bib 6)

Une première expérience de reconnaissance a porté sur les mots. Cette reconnaissance utilise la segmentation (bib 3) d'un mot dont on ne prend que les vecteurs DELTA lors des instants instables, détectés à l'aide d'une courbe de stabilité. Les vecteurs DELTA sont alors calculés comme représentant la variation entre deux spectres situés l'un 20 ms avant l'instant d'instabilité, l'autre 20 ms après.

Une comparaison dynamique entre vecteurs DELTA trouvés et vecteurs DELTA de référence à partir d'une distance d'ordre 1 (somme des valeurs absolues des différences), et ceci pour l'ensemble des vecteurs des mots, donne le candidat-mot le meilleur.

Cette méthode a fourni les résultats suivants, pour un locuteur (mots dissyllabiques, une seule passe d'apprentissage) :

Taux de reconnaissance sur				20	mots	différents		:	96%
"	"	"	"	60	"	"	"	:	94%
"	"	"	"	100	"	"	"	:	91%

Des essais inter-locuteurs sont en cours.

IV - 2. Reconnaissance analytique

La reconnaissance analytique est ici employée dans le but de reconnaître le plus exactement possible une phrase continue, phonème par phonème, au fur et à mesure de sa production, et de ne transmettre que les phonèmes reconnus, pour synthétiser au bout d'une chaîne de transmission le message prononcé.

Le principe de cette reconnaissance est basé sur la pertinence des éléments phonétiques que sont les diphonèmes. Dans un stade final, il faudra donc identifier un diphonème parmi un millier de diphonèmes possibles. Pour nos premiers essais, nous avons employé un dictionnaire d'une trentaine de diphonèmes de référence et nous traitons actuellement des phrases avec une centaine de diphonèmes de référence.

Comme en reconnaissance par mots nous utilisons une courbe d'instabilité spectrale qui est tout simplement la somme, en fonction du

temps, des valeurs absolues des 32 composantes du vecteur DELTA. Cette courbe nous permet de détecter les instants d'instabilité maximale du spectre.

Le vecteur DELTA alors calculé comme étant la variation entre le spectre situé à mi-distance de l'instant stable du premier terme du diphonème et de l'instant instable, et le spectre situé à mi-distance de l'instant instable et du second terme du diphonème. Cet écart de durée variable permet de détecter plus exactement le vecteur DELTA lors de transitions rapides.

Ces vecteurs DELTA, calculés sur 100 niveaux, sont alors comparés, à l'aide d'une distance d'ordre 1, aux vecteurs DELTA de référence, définis dans un dictionnaire. Ces derniers ont été établis par segmentation automatique, corrigée parfois à la main, à partir de diphonèmes prononcés isolément (donc beaucoup plus lentement qu'en parole continue).

L'établissement de ce dictionnaire soulève le problème suivant : faut-il affiner l'algorithme de segmentation du texte à reconnaître afin d'éliminer les segmentations parasites, qui sont cependant l'expression de l'existence physique de certains phénomènes (explosion de plosives, silence suivant une fricative), ou inclure dans le dictionnaire ces phénomènes acoustiques ?

Après quelques essais, nous nous sommes rangés à la deuxième solution.

Pour deux locuteurs, chacun ayant prononcé normalement deux fois une même phrase : "Paul a caché un vase dans le salon", et chacun ayant établi un dictionnaire des diphonèmes composant cette phrase, les résultats ont été les suivants :

	D_1	D_2
$P_1 + P_1'$	56%	45%
$P_2 + P_2'$	20%	26%

P_1, P_1' : phrases prononcées par le locuteur 1

P_2, P_2' : phrases prononcées par le locuteur 2

D_1 : dictionnaire prononcé par le locuteur 1

D_2 : dictionnaire prononcé par le locuteur 2

Le taux de reconnaissance varie donc plus en fonction de la bonne élocution du sujet (d'où bonne segmentation) que suivant l'identité entre l'auteur du dictionnaire de référence et le locuteur. Ceci met à jour la relative invariance du vecteur delta suivant le locuteur. Les résultats eux-mêmes sont légèrement meilleurs que ceux obtenus simultanément par comparaison des spectres normalisés aux instants stables,

ceci étant surtout sensible, bien sûr, lors de la reconnaissance interlocuteurs.

Les taux de reconnaissance obtenus peuvent paraître faibles. Mais il y a théoriquement deux fois plus de chances de faire une erreur sur une reconnaissance aux instants instables que sur une reconnaissance aux instants stables, puisqu'un seul paramère définit deux phonèmes dans le premier cas, alors qu'un paramètre définit un seul phonème dans le second cas. Il est donc remarquable de détecter deux diphonèmes l'un à la suite de l'autre tels que le deuxième membre du premier soit similaire au premier membre du second. A fortiori, le fait que trois diphonèmes s'enchaînent parfaitement prête au diphonème du milieu une probabilité d'exactitude qui est grande. Cette remarque nous conduit à placer des jalons (diphonèmes réputés corrects) au sein de la phrase continue, et à effectuer la concaténation des diphonèmes détectés entre ces deux jalons suivant la méthode de cheminement présentée dans la figure 5.

Grâce à cet algorithme de concaténation nous avons obtenus les résultats suivants (les jalons sont soulignés) :

Phrase P_1 Transcription phonétique	p ɔ l a k a ʃ e ɛ̃ v a z d ã l ə s a l õ
Reconnaissance par D_1	p ɔ l a k a ʃ e ɛ̃ ɛ̃ ɛ̃ ɛ̃ ɛ̃ v a z d ã l ə s a l õ õ
Reconnaissance par D_2	p ɔ l a k a ʃ . e ɛ̃ ɛ̃ ɛ̃ v a z d ã l ə s a l õ õ
Phrase P'_1 Transcription phonétique	p ɔ l a k a ʃ e ɛ̃ v a z d ã l ə s a l õ
Reconnaissance par D_1	p ɔ l a k a ʃ e e ɛ̃ ɛ̃ v a a z d ã l ə s a l õ õ
Reconnaissance par D_2	p ɔ l a k a ʃ . e ɛ̃ ɛ̃ v a a z d ã l ə s a ? ɛ̃ ɛ̃

On voit ainsi qu'il est possible, par cette méthode purement acoustique, de reconstituer presque parfaitement la phrase prononcée par un locuteur à partir du dictionnaire des diphonèmes prononcés isolément par le même ou par un autre locuteur.

Cette possibilité est directement liée à la reconnaissance basée sur les diphonèmes. Elle est voisine de la démarche que semble adopter un auditeur humain, qui reconstitue la phrase entendue au fur et à mesure qu'il intègre certains points forts ou "jalons". Elle présente ainsi l'avantage de pouvoir décrypter une phrase par fragments, sans en attendre la fin, ce qui permet d'envisager son application à la transmission de parole continue en temps légèrement différé.

V - CONCLUSION

Nous nous sommes efforcés de montrer que les transitions phonétiques étaient utilisables sur le plan de l'analyse et de la reconnaissance automatique. Pour cela nous avons défini la notion de vecteur DELTA, qui reflète la variation spectrale sur un intervalle de quelques dizaines de millisecondes.

Cette notion tend à lier dans une même grandeur les dimensions fondamentales de temps, d'amplitude et de fréquence. Elle est directement issue de la théorie de la Forme.

Les premiers résultats observés en reconnaissance nous semblent confirmer la pertinence de ces idées tout en ouvrant la voie à des possibilités nouvelles. Il reste cependant beaucoup de points à approfondir, notamment dans les relations du vecteur DELTA avec les grandeurs phonétiques usuelles.

VI - BIBLIOGRAPHIE

1. E. LEIPP, M. CASTELLENGO, J. SAPALY, J.S. LIENARD - Structure physique et contenu sémantique de la parole - Colloque sur la parole organisé par le GALF à Grenoble - avril 1967 - La Revue d'Acoustique n° 3-4, 1968

2. J.S. LIENARD - Analyse, synthèse et reconnaissance automatique de la parole - Thèse d'Etat - Université Paris VI - avril 1972

3. J.S. LIENARD, M. MLOUKA, J.J. MARIANI, J. SAPALY - Real-time Segmentation of Speech - Speech Communication Seminar - KTH, Stockholm - août 1974

4. S.S. STEVENS, H. DAVIS - Hearing, its psychology and physiology - J. WILEY and sons - N.Y. LONDON - 1938

5. P. DELATTRE - From acoustic cues to distinctive features - Phonetica - 18 : 198-230 (1968)

6. M. MLOUKA - Reconnaissance automatique de la parole : une expérience de reconnaissance par mots - Thèse de 3e cycle - Université Paris VI - octobre 1974

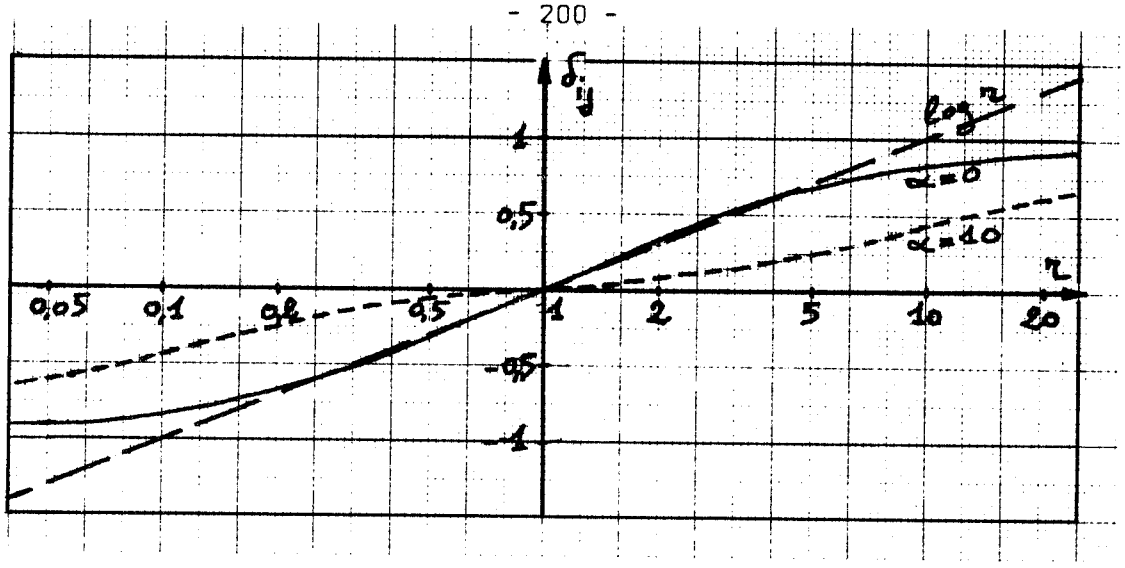


Fig 1 - **Calcul des composantes fréquentielles du vecteur delta**

r représente la quantité $a_{(i+\frac{\theta}{2},j)} / a_{(i-\frac{\theta}{2},j)}$

α représente l'amplitude relative du bruit de fond : $\alpha = b_j / a_{(i-\frac{\theta}{2},j)}$

On a alors $\delta_{ij} = \frac{r-1}{r+1+\alpha}$

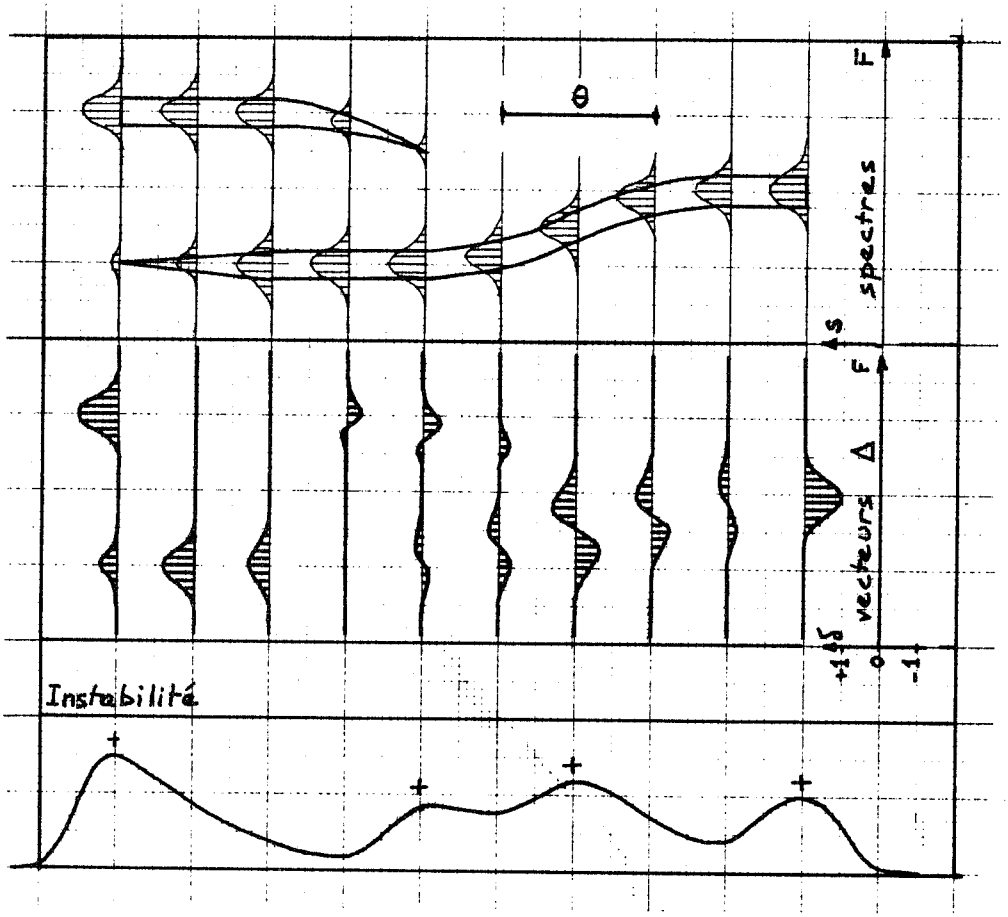


Fig 2 - **Définition du vecteur DELTA sur un spectrogramme schématique**

La courbe d'instabilité spectrale (aire absolue du vecteur delta à chaque instant) est également représentée. Ses maxima sont marqués par des +.

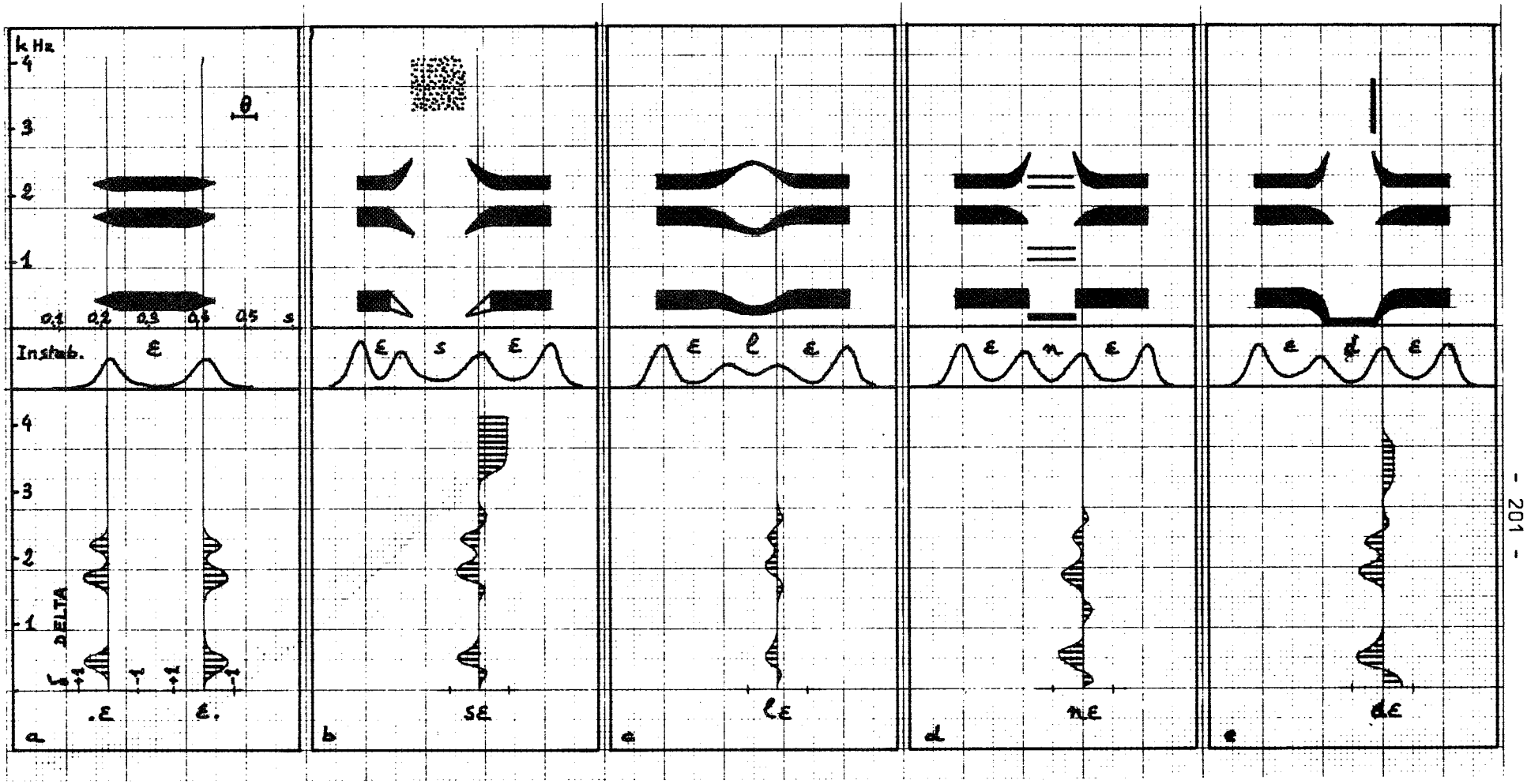


Fig 3 - Le vecteur DELTA associé à quelques transitions phonétiques schématisées par P.DELATTRE en vue de la synthèse

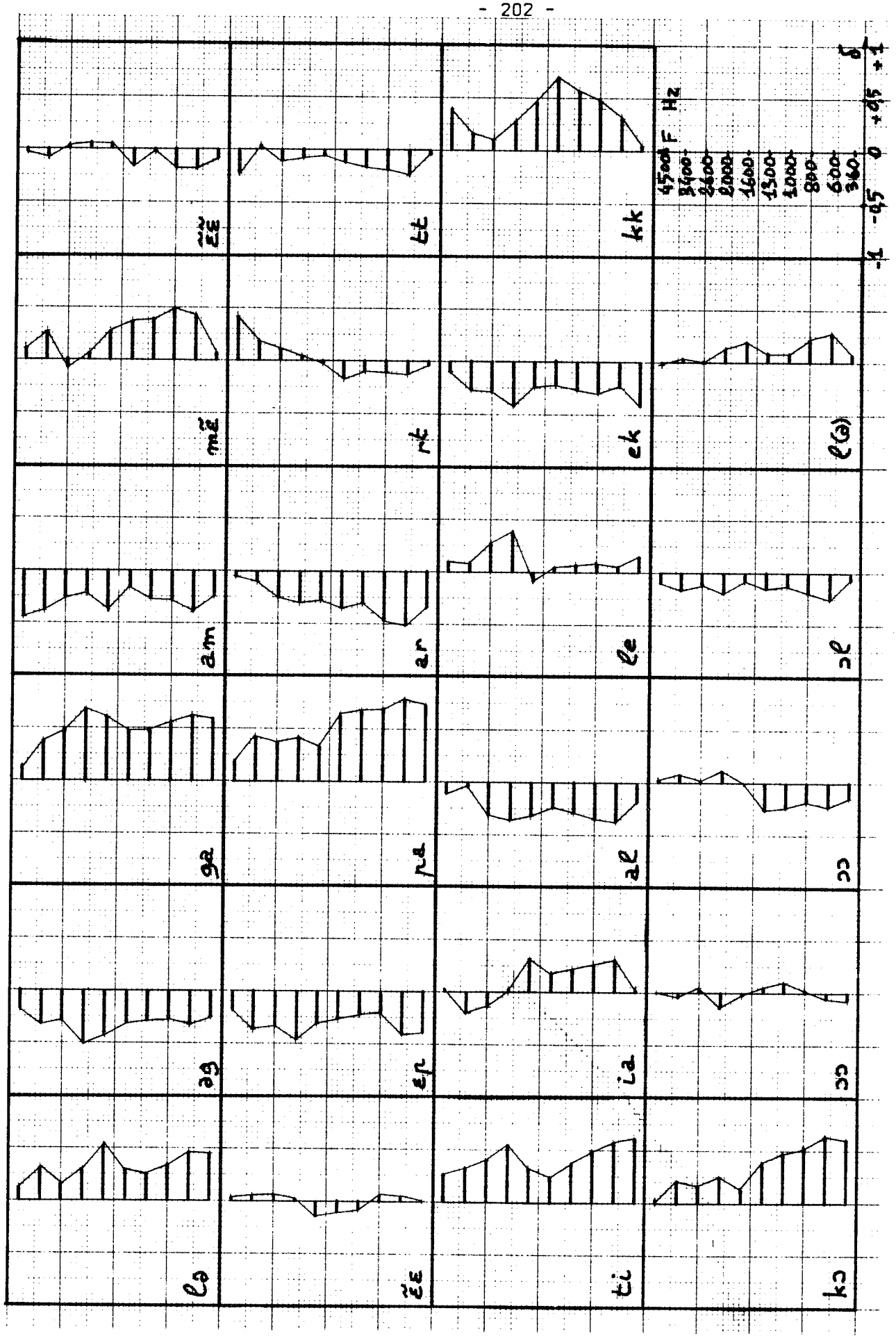


Fig 4 - Vecteurs DELTA établis par analyse (banc de 10 filtres) aux instants d'instabilité maxima

Phrase : "le gamin est parti à l'école"

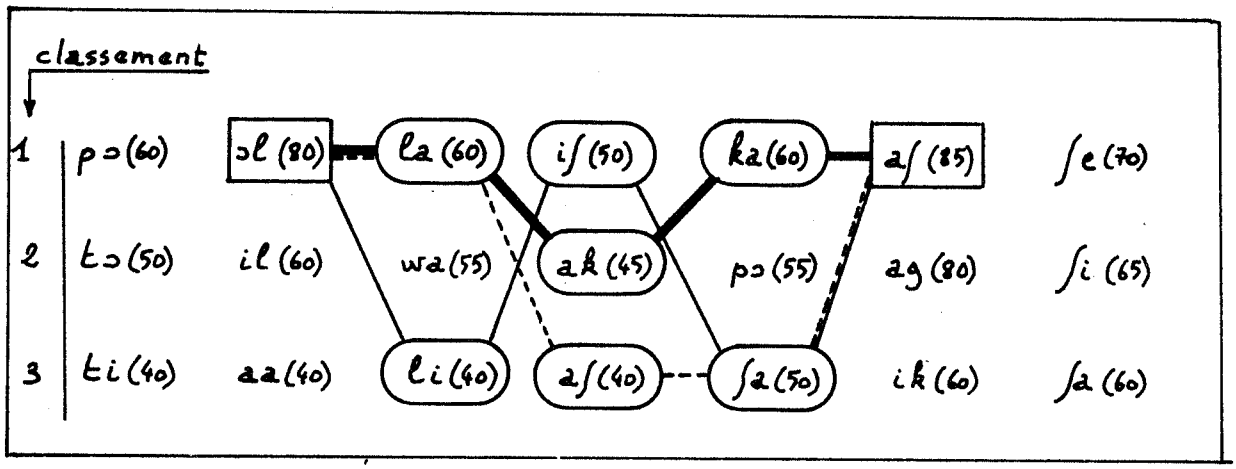


Fig 5 - **Fonctionnement de l'algorithme de concaténation des diphonèmes**

chemin A (trait épais) ɔlakaʃ note=165

chemin B (trait interrompu) ɔliʃaʃ note=140

chemin C (trait fin) ɔlaʃaʃ note=150

Les jalons trouvés sont encadrés dans des rectangles. Les diphonèmes-candidats sont classés de haut en bas par probabilité décroissante. Ne figurent dans cet exemple que les trois premiers candidats, suivis de leurs notes.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

DETERMINATION DE FORMANTS ET DE TRAITS PAR CODAGE PREDICTIF

D. TUFFELLI

E.N.S. D'ELECTRONIQUE ET DE RADIOELECTRICITE - GRENOBLE

INSTITUT DE MATHEMATIQUES APPLIQUES - GRENOBLE

RESUME

En utilisant la méthode de Markel pour détecter les formants, on doit calculer les coefficients d'autocorrelation, ceux de prédiction et l'erreur normalisée. On étudie ces différents coefficients en vue d'une détection de traits ("parole-silence", "voisé-non voisé", "consonnes voisées"). Quelques résultats de segmentation et de détection de formants sont montrés.

SUMMARY

In using Markel method for formant detection, one must compute the autocorrelation coefficients, the prediction ones and the normalized error. We study these various coefficients for feature detection ("voiced-unvoiced", "speech-silence", "voiced-consonants"). Some results on segmentation and formant detection are shown.

DETERMINATION DE FORMANTS ET DE TRAITS PAR CODAGE PREDICTIF

D. TUFFELLI

E.N.S. D'ELECTRONIQUE ET DE RADIOELECTRICITE - GRENOBLE

INSTITUT DE MATHEMATIQUES APPLIQUES - GRENOBLE

1 - INTRODUCTION

Le codage prédictif est une technique très générale de traitement du signal /3/ qui a été appliquée dans de nombreux domaines. En particulier, il a été utilisé en analyse synthèse de la parole /1/ et /2/ où d'excellents résultats ont été obtenus. En reconnaissance l'emploi de cette technique permet au niveau du prétraitement d'effectuer une détection des formants. Un processus de reconnaissance de la parole (au niveau acoustique et phonétique) peut se décomposer schématiquement en une compression d'informations (analogue souvent à celle d'un système d'analyse-synthèse) et une phase de reconnaissance complexe qui dépend étroitement des paramètres détectés à l'analyse. Dans ce rapport, nous décrirons les différents résultats obtenus à partir d'une analyse par codage prédictif : choix et calcul des paramètres, détection des formants, et de différents traits qui pourront être utilisés ultérieurement pour une reconnaissance de la parole.

2 - MODE OPERATOIRE

a) Enregistrement et acquisition du signal

- Enregistrement en ambiance peu bruyante
- Filtrage passe haut 300 HZ et passe bas 4000 HZ
- Fréquence d'échantillonage 10 KHZ
- Conversion analogique - digitale sur 10 bits.

b) Traitement du signal

- Méthode de Markel appliquée au signal non préaccentué
- Analyse asynchrone
- Fenêtre de Hamming de largeur 25 msec progressant par pas de 6 msec

c) Contrôle des résultats

A partir de :
- Sonagrammes
- Examens du signal temporel
- Ecoutes directes d'une fraction du signal par conversion digitale - analogique
- Ecoutes de synthèses obtenues avec le synthétiseur à formants de l'E.N.S.E.R.G.

3 - DETECTION DES PARAMETRES

La reconnaissance des sons "voisés" est effectuée avec un certain succès à partir des formants /5/ à /8/. On a donc cherché à détecter ces paramètres. Mais, en utilisant des techniques de codage prédictif, le calcul des formants implique le calcul des coefficients d'autocorrélation et de prédiction sur le signal de parole ainsi que le calcul de l'erreur normalisée. On a cherché à utiliser ces différents coefficients pour effectuer des détections de traits acoustico-phonétiques.

Au niveau du calcul des coefficients d'autocorrélation, on a retenu comme paramètre le premier coefficient d'autoccorélation /4/ qui est égal à l'intégrale du spectre de puissance pondéré par un cosinus.

$$R1 = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} P(\omega) \cos(\omega T)\, d\omega$$

$$T = \frac{1}{Fe}$$ Fe : Fréquence d'échantillonage

En fait, on normalise par rapport à l'énergie (courbe RO) et on utilise r1 = R1/RO pour la détection du trait de "voisement", r1 est compris entre -1 et +1 (courbe R1). On a choisi des seuils de 0.85 et 0.65 pour la détection du caractère "voisé" ou "non voisé". On reporte la décision à une étape ultérieure si r1 est compris entre 0.65 et 0.85. De plus, la décision "voisé" est prise systématiquement si l'énergie RO est supérieure à un seuil. Pour la décision "parole-silence", deux seuils sont utilisés. Si l'énergie RO est inférieure à un seuil assez faible pour les sons "non voisés" ou à un seuil un peu plus élevé pour les sons "voisés" l'échantillon est classé "silence". Lorsqu'il existe une composante continue assez forte dans le signal, la décision "silence" est prise après un test sur la valeur des coefficients d'autocorrélation (qui sont alors tous de valeur élevée).

Au niveau du calcul des coefficients de prédiction, on a retenu comme paramètres : l'erreur normalisée, le premier coefficient de prédiction et la somme des valeurs absolues des coefficients de prédiction. On utilise d'abord un filtre d'ordre 6. On a tracé la courbe représentant l'erreur normalisée calculée avec 6 coefficients de prédiction (courbe EN). L'erreur normalisée varie peu à partir de 6 coefficients de prédiction. On étiquette l'échantillon comme "silence" si l'erreur normalisée est supérieure à 0.85, comme "non-voisé" si cette erreur est supérieure à 0.30 et comme "voisé" si elle est inférieure à 0.06. On peut utiliser également pour la détection "voisé-non-voisé" le premier coefficient de prédiction, ce coefficient représente, d'après la propriété bien connue des équations polynomiales, la somme des poles de la fonction de transfert. Si les poles de bande passante étroite sont en moyenne situés en basse fréquence, ce coefficient sera plus grand que si les poles sont situés en haute fréquence.

Par ailleurs, on a relevé souvent une relation assez surprenante entre l'énergie du signal et la somme des valeurs absolues des coefficients de prédiction (courbe RO et $\sum |a_k|$). On a déjà vu comment varie le premier coefficient de prédiction, quant aux autres, ils représentent la somme des produits n à n des poles, cette somme sera d'autant plus faible en valeur absolue que les bandes passantes des poles sont larges.

On essaiera d'utiliser un paramètre de ce type pour une segmentation des sons voisés.

A ce niveau de calcul (filtre d'ordre 6), les décisions "parole-silence" et "voisée-non voisée" sont définitives. Pour classer les échantillons non encore étiquettés on se sert d'une agglomération des différents critères précédents (on a représenté par la courbe D1 le résultat d'une agglomération de critères pour la décision de "voisement"). Le résultat final de la classification est représenté par la courbe D2.

Au niveau du calcul des formants, on utilise 14 coefficients pour les sons voisés et 6 seulement pour les sons non voisés. Différents essais ont été effectués pour essayer de déterminer, au moyen de la pente de l'erreur normalisée, le nombre de coefficients de prédiction minimum. Ceci a conduit assez souvent à des valeurs de formants erronées pour les consonnes voisées. Les meilleurs résultats ont été obtenus avec 14 coefficients (nombre maximum essayé). On a conservé comme paramètre possible la variation relative de l'erreur normalisée entre 6 et 10 coefficients en vue de détecter les consonnes voisées.

Pour le calcul des formants, on utilise différentes contraintes /9/. Les plages à priori de F1, F2, F3 sont en HZ : (180, 1100) (510, 3500) (1750, 4000). En fait, dans notre expérience, ces plages sont variables en fonction du premier coefficient d'autocorrélation du signal préaccentué (coefficient représenté par la courbe R1P) ceci pour permettre essentiellement la distinction des sons O et OU d'une part (l'énergie est alors concentrée en basse fréquence) et I d'autre part. Bien que les valeurs des bandes passantes calculées soient trop grandes lors des transitions (comparées à des valeurs théoriques /10/) on utilise également des contraintes sur les bandes passantes des formants. En particulier, pour éliminer toute recherche inutile de formants dans une zone mal définie, la bande passante du pole correspondant à la plage du deuxième formant doit être inférieure à 400 HZ. Si l'on n'arrive pas à trouver de pole acceptable pour F3, on réduit la borne supérieure de la plage de F2 et on recommence une recherche de F2 et F3. Les résultats sont présentés par les courbes F1, F2, F3.

Cette détection est purement locale. On ne tient aucun compte des échantillons précédents ou suivants. On compte utiliser dans une deuxième étape (après la première étape de détection locale et avant tout lissage) la continuité des formants pour trouver les formants dans les échantillons pour lesquels aucune décision n'é été prise, pour éliminer les points aberrants et pour corriger des transitions trop rapides.

La méthode de Bernouilli et celle de Bairstow sont utilisées pour la résolution de l'équation polynomiale. Puisque les poles d'un échantillon à l'autre sont relativement stables, on conserve les poles Z_i de l'échantillon précédent pour résoudre l'équation de l'échantillon courant. On effectue un changement d'axe $Z' = Z - Z_i$ avec le schéma de Horner puis un changement de variable $Y = 1/Z'$. On applique Bernouilli au polynome en Y résultant, il y a convergence vers le pole le plus près de Zi. On obtient ainsi une valeur grossière de la racine que l'on affine ensuite avec la méthode de Bairstow.

4 - CONCLUSION

Le système décrit ici n'a pas été exploité de manière intensive et de nombreux points devront être améliorés et approfondis. La rapidité de calcul sera grande et proche du temps réel avec un calculateur spécialisé /11/.

Nous avons cherché à mettre en évidence des paramètres utilisables pour la reconnaissance issus directement d'un traitement du signal par codage prédictif. Les résultats déjà obtenus sont très encourageants.

BIBLIOGRAPHIE

/1/ ATAL B.S., HANAUER S.L.
Speech analysis and synthesis by linear prediction of the speech wave
J.A.S.A. 50, 637 - 655 (1971)

/2/ MARKEL J.D., GRAY A.H., WAKITA JR.H.
Linear prediction of speech theory and practice
S.C.R.L. Monograph n° 10 (1973)

/3/ MENDEL J.M.
Discrete techniques of parameter estimation
Marcel Dekker, INC. New-York (1973)

/4/ SERIGNAT J.F.
Travaux sur le vocoder à autocorrelation
Etude et simulation d'un vocoder à prédiction linéaire
Thèse docteur-ingénieur. Grenoble (1974)

/5/ TUBACH J.P.
Reconnaissance de la parole
Thèse doctorat ès-sciences. Grenoble (1970)

/6/ FACCA R.
Méthode de segmentation et d'analyse par traitement direct du signal vocal.
Application à la classification et la reconnaissance des consonnes.
Thèse docteur-ingénieur Toulouse (1974)

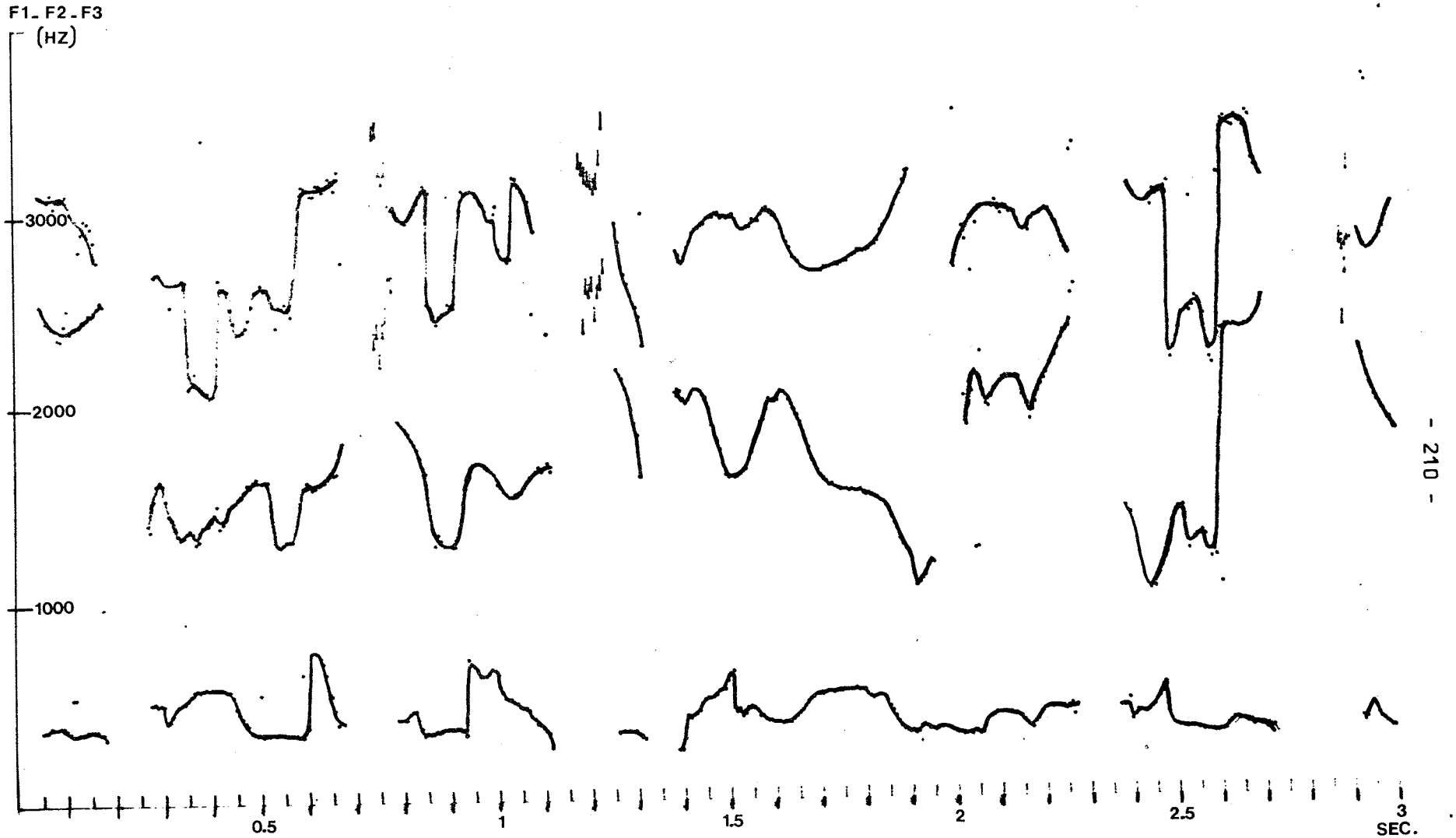
/7/ MCCANDLESS S.
An algorithm for automatic formant extraction
IEEE transactions on acoustics, speech and signal processing
ASSP - 22, 135-141 (1974)

/8/ CARRE R.
Contribution aux études sur l'analyse et la synthèse de la parole. Rôle et importance des formants.
Thèse doctorat ès-sciences Grenoble (1971)

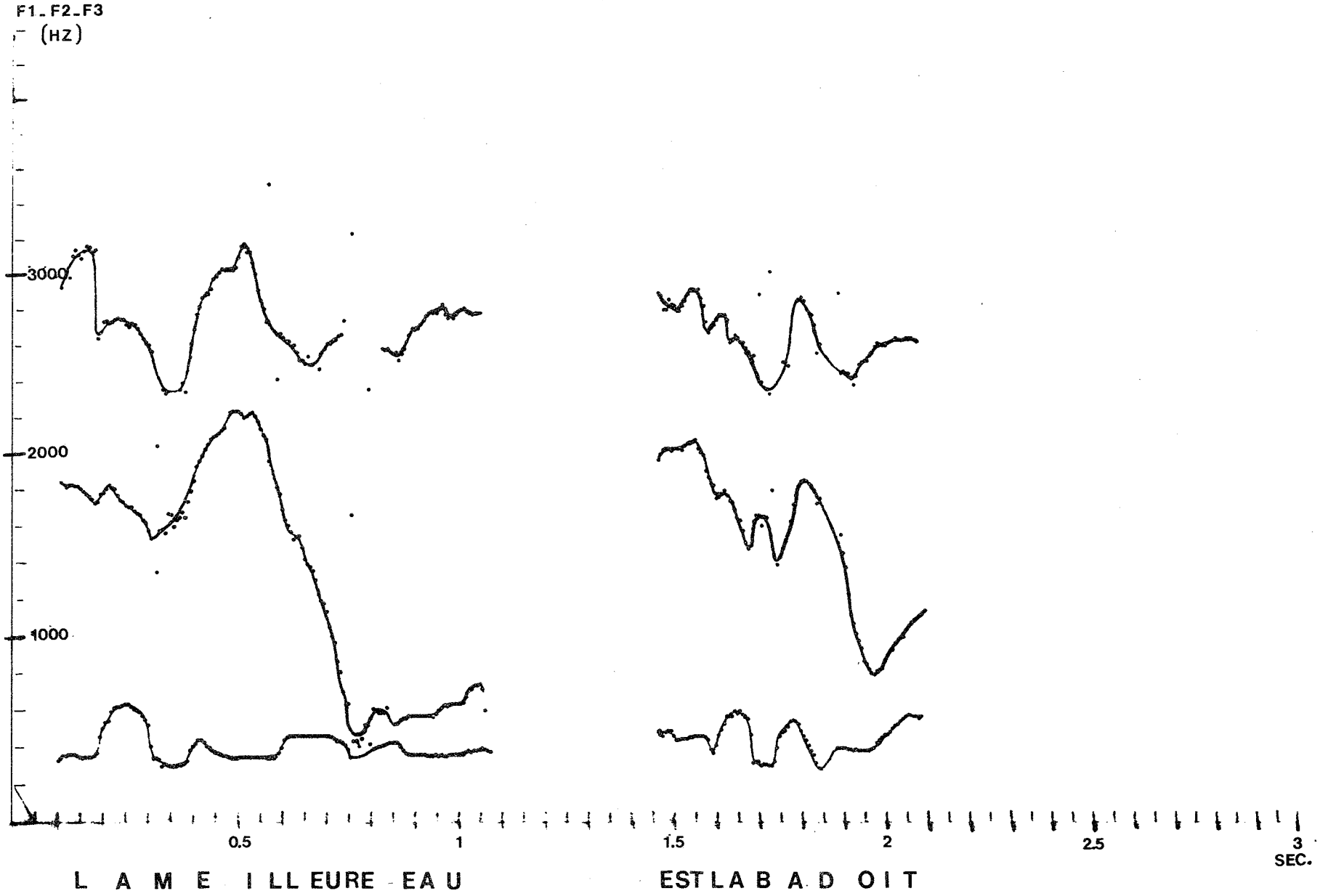
/9/SCHAFER R.W., RABINER L.R.
System for automatic formant analysis of voiced speech
J.A.S.A. 47, 634-648 (1970)

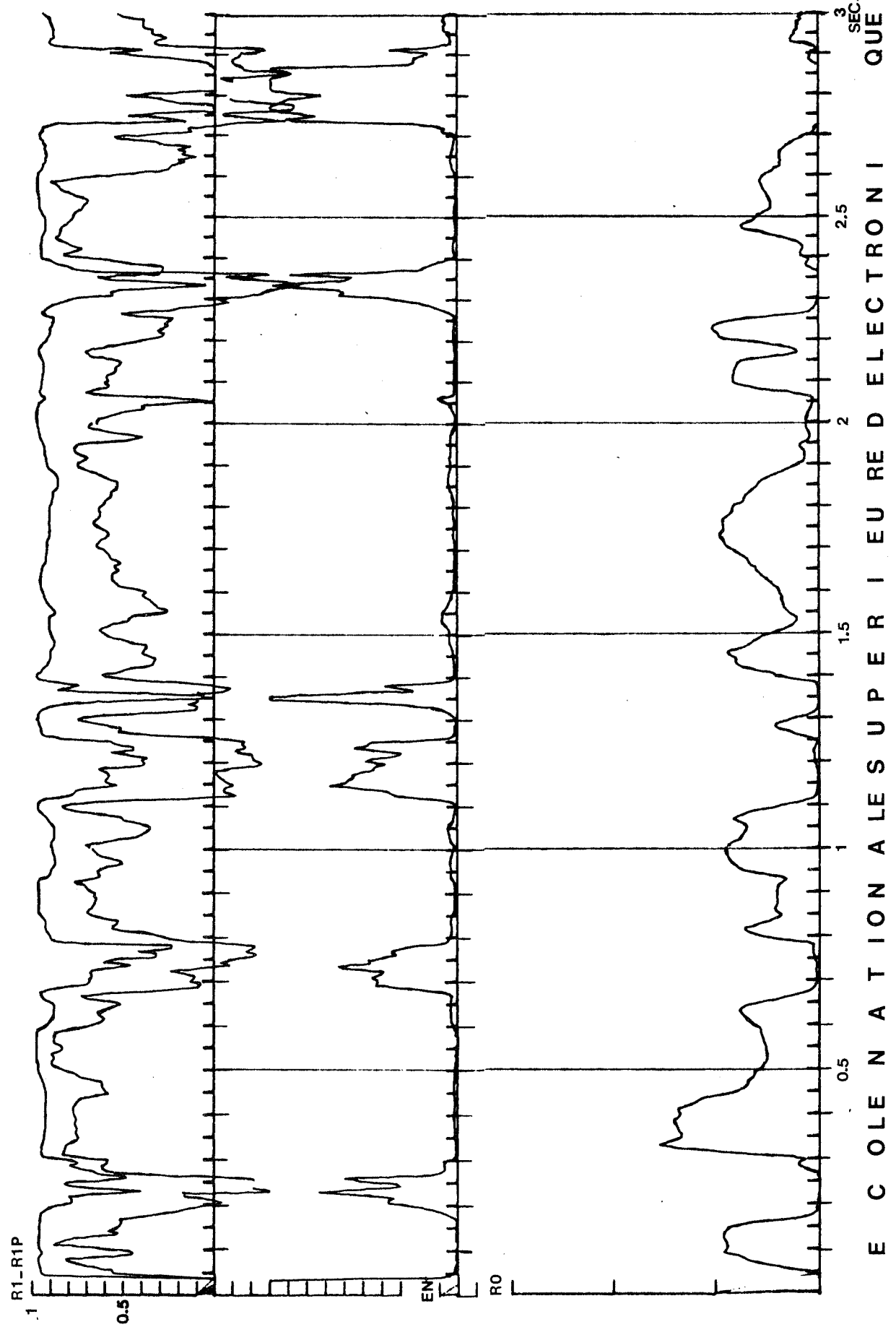
/10/ FANT G.
Vocal tract wall effects, losses, and resonance bandwidths
STL - QPSR 2-3, 28-52 (1972)

/11/ DEGRYSE D., SERIGNAT J.F., CARRE R.
A special speech processor
Conference on speech communication and processing.
Paper K5. BOSTON (1972).

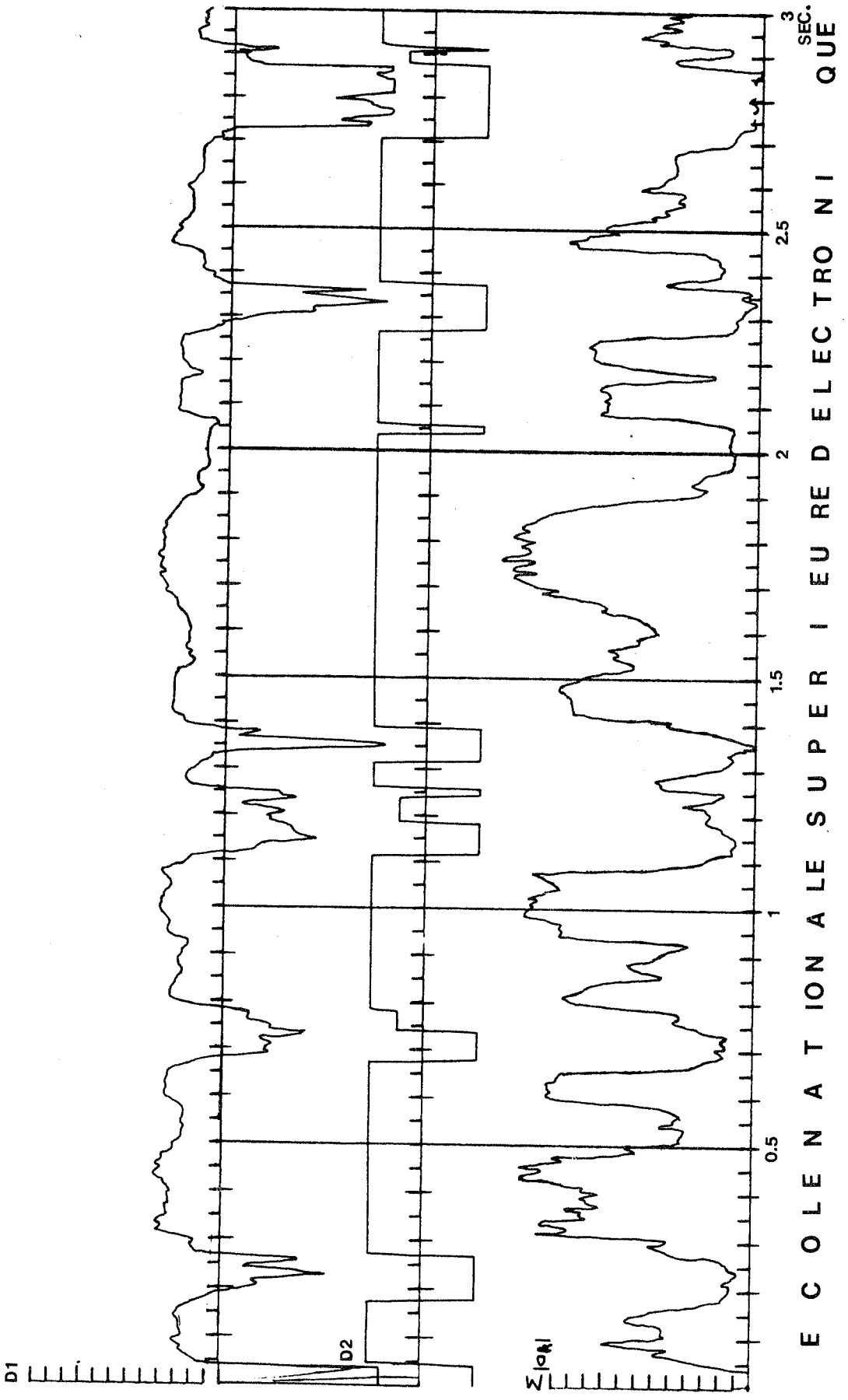


E C O L E N A T I O N A L E S U P E R I EU RE D EL EC TRO N I QUE


F1_F2_F3
(HZ)
3000
2000
1000
0.5
1
1.5
2
2.5
3
SEC.
L A M E I LL EURE - EA U
EST LA B A D O I T



E C O L E N A T I O N A L E S U P E R I E U R E D E L E C T R O N I Q U E


D1
D2
Σ|ORI|
0.5
1
1.5
2
2.5
3
SEC.
E C O L E N A T I ON A LE S U P E R I EU RE D E L EC TRO N I QUE

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

L'APPORT DE L'ANALYSEUR DU PROJET A.R.I.A. SUR QUELQUES EXEMPLES D'ANALYSE PHONETIQUE

D. DOURS* - R. FACCA* - G. MAURAND** - G. PERENNOU***

RESUME A partir des paramètres formantiques et pour diverses représentations, on donne un certain nombre de résultats d'analyse afin de montrer que la probabilité de confusion entre éléments phonétiques est rendue faible.

Dans une deuxième partie, on envisage le problème des semi-voyelles et on montre les transitions qui les caractérisent.

On examine ensuite le cas des consonnes fricatives voisées pour lesquelles la méthode d'analyse fait apparaître clairement la structure formantique.

On termine par l'examen de l'interaction des consonnes plosives avec les voyelles dans toutes les configurations possibles.

SUMMARY Starting from formantic parameters in the case of various representations, a certain number of analysis résults are given, in order to demonstrate that the probability of confusion between phonetical elements is highly reduced.

In a second part, we consider the semi-vowels problem and show the transitions that characterize them.

Then the case of voiced-fricatives is examined : the method of analysis shows clearly their formantic structure.

Finally we examine the interaction of plosive consonnants with vowels in all possible configurations including final position.

* Assistant - Université Paul Sabatier Toulouse
** Professeur - Université du Mirail Toulouse
***Professeur - Université Paul Sabatier Toulouse

L'APPORT DE L'ANALYSEUR DU PROJET A.R.I.A. SUR QUELQUES EXEMPLES D'ANALYSE PHONETIQUE

D. DOURS - R. FACCA - G. MAURAND - G. PERENNOU

1. INTRODUCTION

Nous présentons les résultats de l'analyse des sons produits par la source glottale, à savoir : les voyelles orales, les semi-voyelles, les fricatives sonores. Le cas des voyelles nasales, des consonnes nasales et des consonnes liquides, n'est pas envisagé ici. Nous parlerons par contre des transitions de formants dans le cas des plosives.

Tous ces sons sont susceptibles d'être analysés par la méthode d'identification de la réponse impulsionnelle, développée dans l'analyseur du projet A.R.I.A. |1| - |2|.

2. LES VOYELLES ORALES

Sur les segments voisés, l'analyse produit pour chaque période du fondamental un vecteur paramètre (fréquence, amplitude, phase, et amortissement de chaque formant, énergie moyenne du signal, etc...)

A ce niveau plusieurs problèmes se posent.

- Dans le cas où le segment voisé correspond à la réalisation de plusieurs phonèmes, il faut découper le segment en intervalles correspondants chacun à un phonème.

- Sur chaque intervalle, extraire les valeurs de formants du phonème. Parmi les paramètres fournis par l'analyseur, les fréquences des trois premiers formants, nous semblent bien adaptées à la reconnaissance des voyelles orales. C'est pourquoi dans un premier temps nous n'avons utilisé que ces trois paramètres.

- Compenser les phénomènes de coarticulation quand ils existent.

2.1. Extraction des fréquences des formants

Le découpage en intervalles s'effectue simplement à partir des discontinuités de la courbe d'énergie moyenne du signal. |2| - |3|.

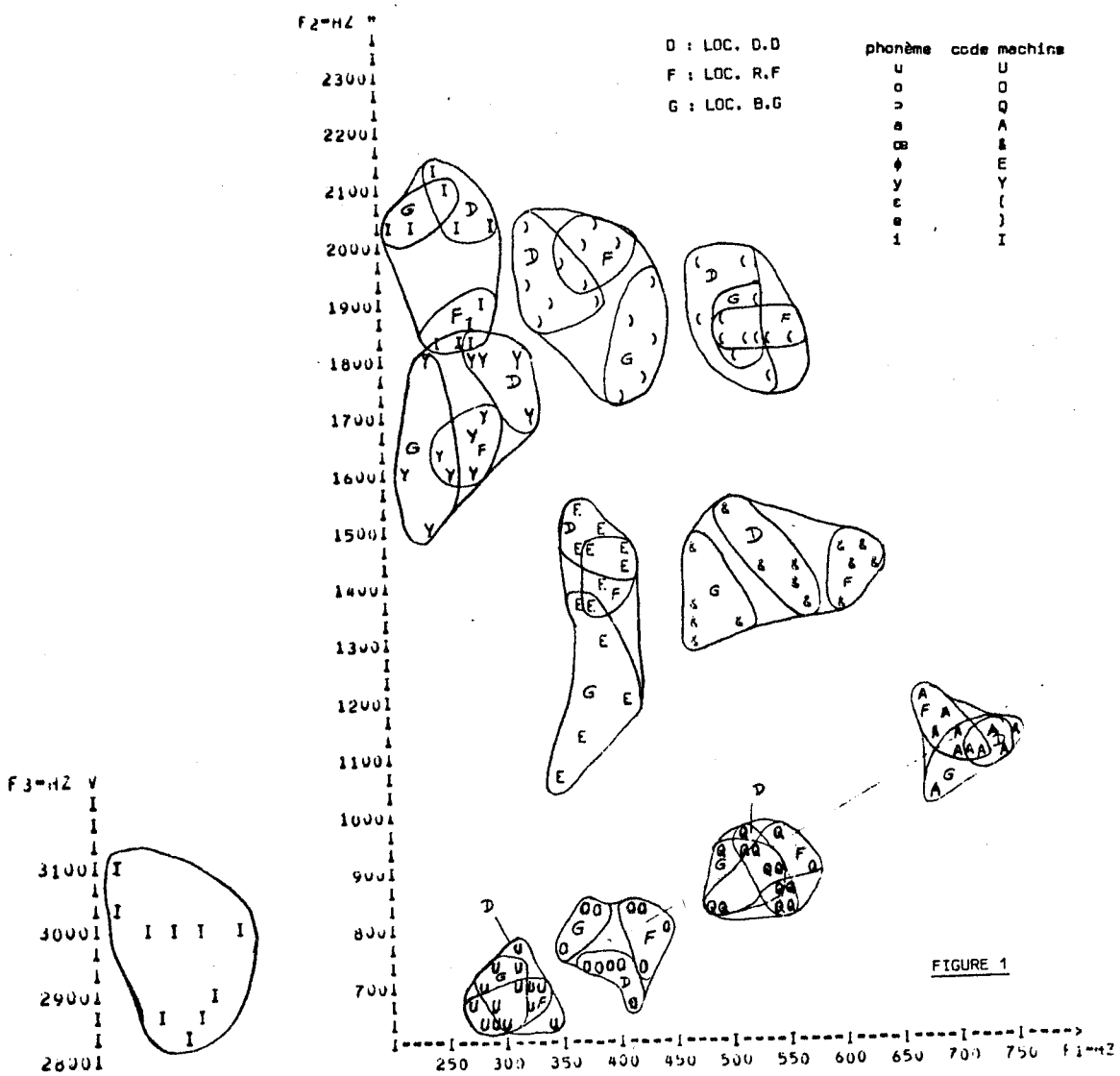


FIGURE 1

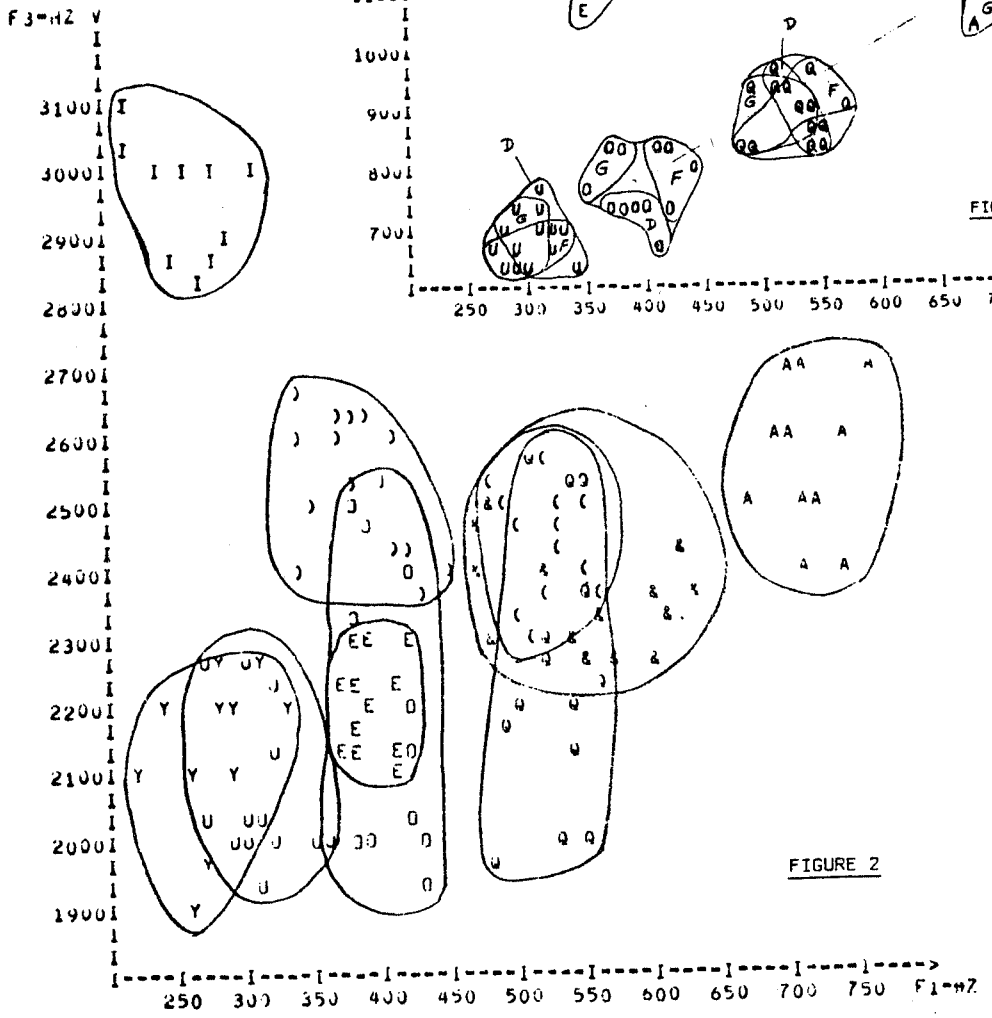


FIGURE 2

Chaque intervalle $[t_i, t_f]$ ainsi déterminé correspond à la réalisation d'un phonème et contient N périodes de fondamental. Chaque formant i est caractérisé par une suite de valeurs : $\{F_{ij}\}$ $j = 1,2,\ldots,N$.

Dans le cas où la suite est monotone, le phénomène de coarticulation n'entre pas en jeu et on extrait les paramètres fréquentiels du vecteur correspondant au maximum d'énergie. Dans l'autre cas, une procédure corrective s'impose, pour compenser le phénomène de la coarticulation. En effet, la coarticulation a tendance à supprimer les zones stables des voyelles et même parfois la réalisation peut être éloignée de la valeur de cible. Nous avons donc développé un algorithme tenant compte de ces phénomènes |2| .

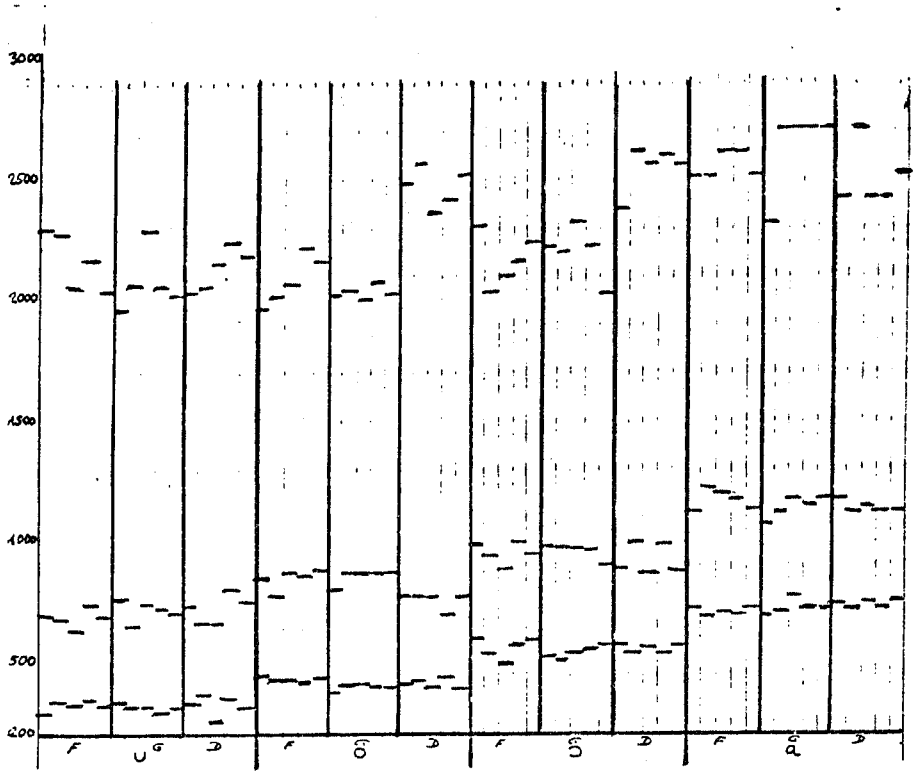
2.2. Représentation des voyelles

Nous avons constitué un échantillon de voyelles issues de cinq contextes différents et prononcées par les trois locuteurs DD-RF-BG (figure 2 bis).

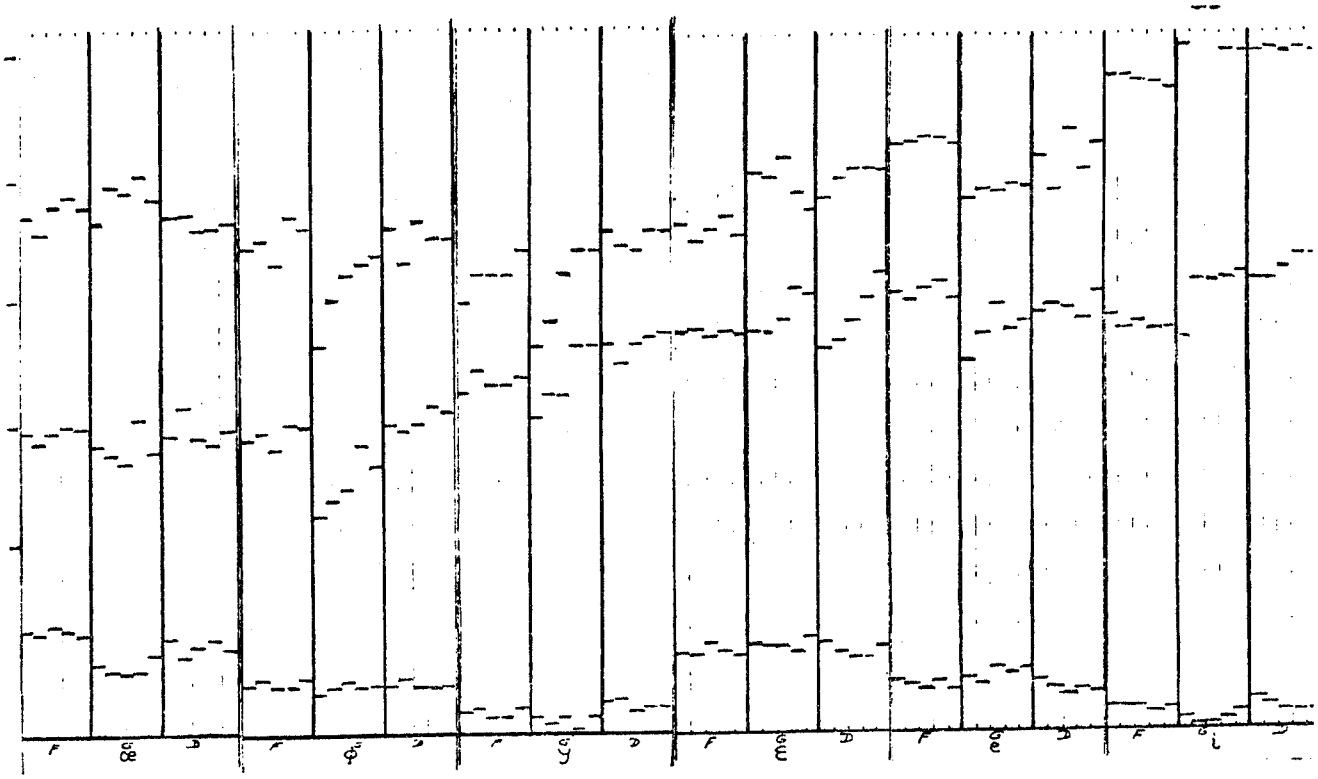
La représentation dans le plan F_1-F_2 met en évidence ce que l'on appelle le triangle vocalique. On constate en effet que les points se groupent dans un triangle approximatif, ayant pour sommets |i|, |u|, le troisième sommet étant donné par le point d'intersection des deux droites sur lesquelles se répartissent deux groupes de voyelles (figure 1).

On peut tout d'abord remarquer que pour un locuteur donné, la répartition des voyelles est telle qu'aucune confusion n'est possible. Il n'en est plus de même pour les trois locuteurs réunis. En effet, certaines confusions apparaissent notamment en ce qui concerne le |y| et le |i| .

Dans le plan F_1-F_3 (figure 2), les confusions sont nombreuses. On peut noter toutefois que la séparation du |y| et du |i| est très nette. Il semble donc qu'il soit important de faire intervenir la fréquence du 3ème formant pour obtenir une meilleure séparation des voyelles.



Représentation des voyelles d'arrière



Représentation des voyelles d'avant labialisées

Représentation des voyelles d'avant non labialisées

Figure 2 bis

Classiquement, on peut à partir des paramètres articulatoires regrouper les voyelles en trois classes (voyelles d'arrière, voyelles d'avant labialisées, voyelles d'avant non labialisées) que nous avons notées respectivement C_1, C_2, C_3 et les différencier entre elles à l'intérieur d'une même classe.

Les fréquences des trois premiers formants étant en relation avec les paramètres articulatoires, nous allons maintenant déterminer des paramètres mixtes basés sur F_1, F_2, F_3 permettant de classer et de reconnaître les voyelles conformément à ce qui est fait du point de vue articulatoire.

2.3. Recherche d'un repère et d'axes discriminants

Les détails techniques de ce qui suit sont exposés dans |7| . Nous nous bornerons ici à donner l'essentiel et les résultats du point de vue phonétique.

2.3.1. Repère discriminant pour les classes C_1, C_2, C_3 .

Pour effectuer la séparation des voyelles en trois classes C_1, C_2, C_3, un examen de la représentation dans le plan F_1-F_2 montre qu'il suffit de séparer C_1 de C_2 et C_3 puis C_3 de C_2 et C_1.

En prenant comme critère à optimiser la largeur de la bande déterminée par deux plans parallèles séparant les classes (encore appelée : facteur de tolérance |8|, nous avons effectivement obtenu de très bons résultats. Sur la figure 3, nous avons donné une représentation des voyelles dans le nouvel espace défini par les axes orthogonaux aux deux bandes optimales, dont les équations exactes sont les suivantes :

$$\psi_1(x) = -178{,}2 - 0{,}61F_1 + 0{,}77F_2 - 0{,}18F_3$$

$$\lambda = 112{,}2$$

$$\psi_2(x) = 2765{,}5 - 0{,}44F_1 - 0{,}67F_2 - 0{,}60F_3$$

$$\lambda = 47{,}46$$

avec $x = (F_1, F_2, F_3)$.

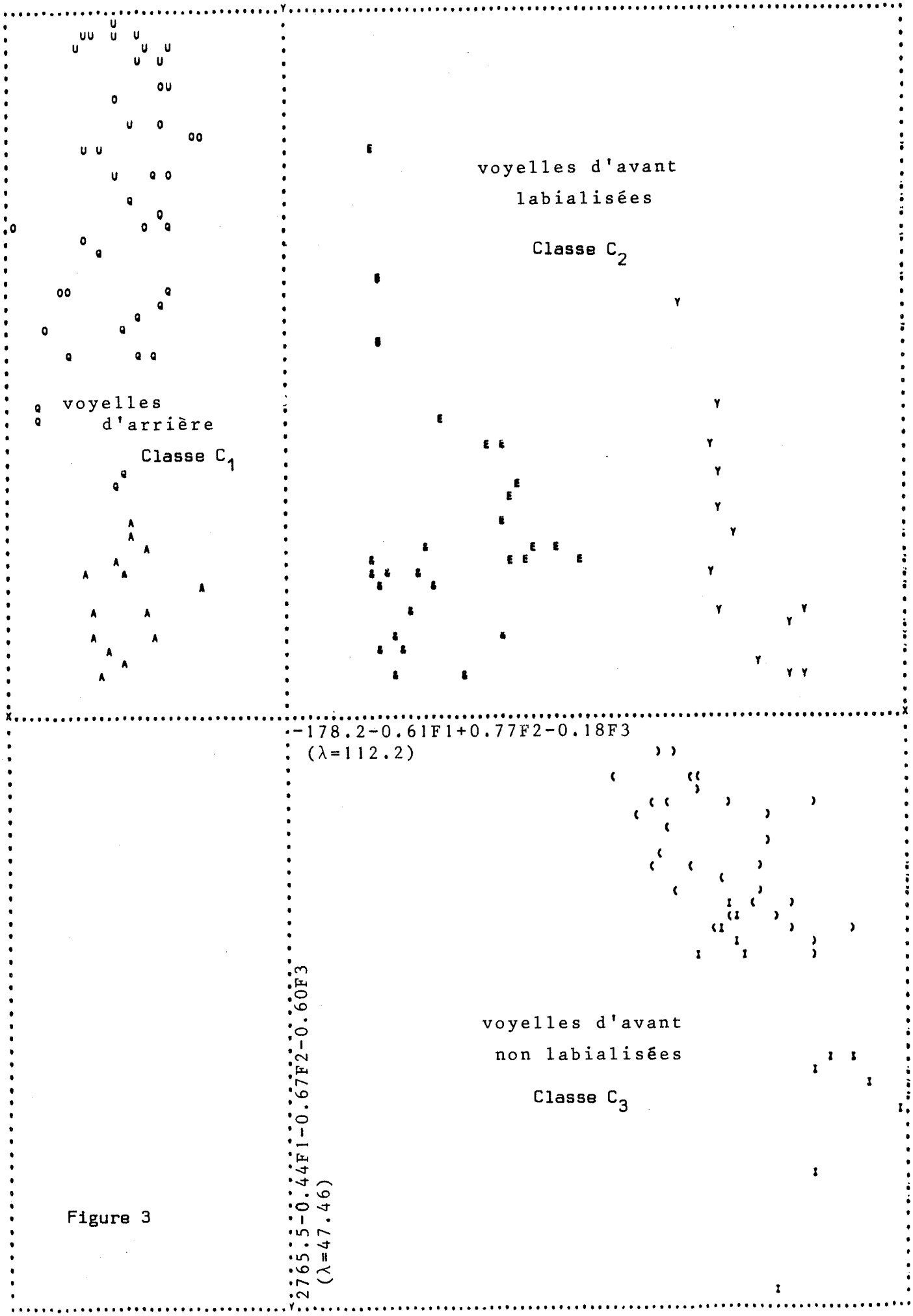

voyelles
d'arrière
Classe C_1
voyelles d'avant
labialisées
Classe C_2
voyelles d'avant
non labialisées
Classe C_3
-178.2-0.61F1+0.77F2-0.18F3
(λ=112.2)
2765.5-0.44F1-0.67F2-0.60F3
(λ=47.46)

Figure 3

Le nombre λ représentant la demi-largeur de la bande à l'optimum est exprimé en Hz.

On observe sur cette figure que chacune des classes est disposée dans un quadrant et quelles sont nettement séparées les unes des autres.

En pratique, deux tests permettent la reconnaissance de ces trois classes.

Si $\psi_1(x) < 0$ et $\psi_2(x) > 0$ la voyelle x est affectée à la classe C_1

Si $\psi_1(x) > 0$ et $\psi_2(x) > 0$ la voyelle x est affectée à la classe C_2

Si $\psi_1(x) > 0$ et $\psi_2(x) < 0$ la voyelle x est affectée à la classe C_3.

2.3.2. Axes discriminants par classes

Nous voyons sur le repère précédent apparaître des zones de confusion sauf dans la classe C_2, dont certaines n'existaient pas dans le plan F_1-F_2. L'examen de ce même plan montre par ailleurs qu'à l'intérieur de chaque classe, les voyelles se rangent approximativement sur un axe de telle manière qu'il semble possible de les y localiser à l'aide d'une seule coordonnée. Nous avons donc cherché, pour chacune des classes, l'application de la forme : $\varphi(x) = W_0 + W_1F_1 + W_2F_2 + W_3F_3$ avec $x = (F_1,F_2,F_3)$ (forme linéaire affine) minimisant la somme des deux critères suivants :

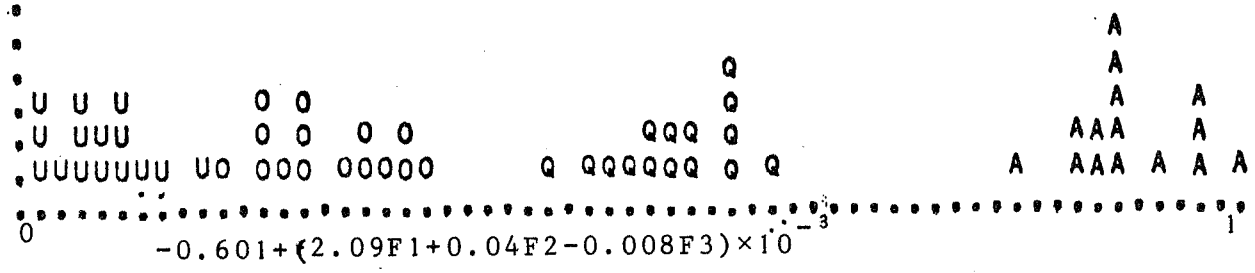
- la somme des variances par classe de voyelles
- la somme des écarts quadratiques $(\varphi(x) - \varphi(y))^2$ des couples $(\varphi(x), \varphi(y))$ intervertis.

Sur la figure 3 bis, on trouvera les trois formes linéaires affines ainsi déterminées, ainsi que la répartition des voyelles sur les axes correspondants. On observera que ces axes n'entraînent en fait aucune interversion.

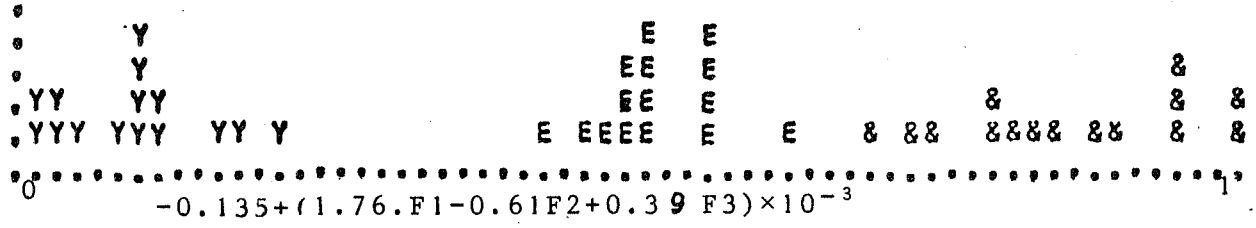
2.3.3. Remarque

Comme nous l'avons vu, deux tests permettent d'affecter une voyelle inconnue x dans l'une des trois classes C_1, C_2 ou C_3. Il suffit ensuite de calculer la fonction $\varphi(x)$ correspondante, pour préciser la voyelle. Cette méthode très simple permet une reconnaissance très rapide des voyelles.

densité relative — voyelles d'arrière



densité relative — voyelles d'avant labialisées



densité relative — voyelles d'avant non labialisées

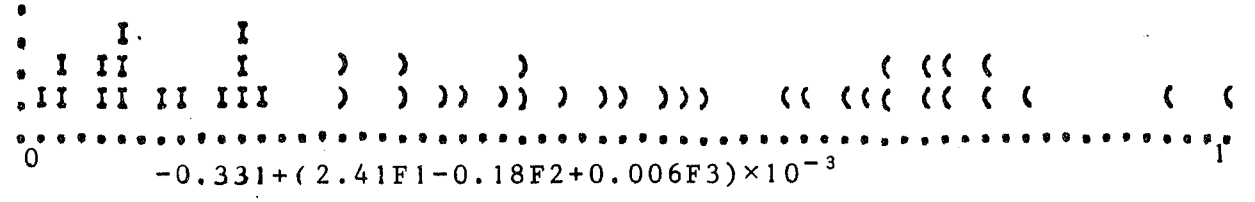


Figure 3 bis

III - LES SEMI-VOYELLES

Les semi-voyelles sont caractérisées par une énergie plus faible que celle des voyelles orales et surtout par des transitions de formants. On peut même dire qu'elles ne sont que transitions de formants.

Les figures suivantes mettent bien en évidence ces transitions. Elles représentent le segment i j $\tilde{5}$ pris dans le mot "papillon" (figure 4) et le mot "oui" (figure 5).

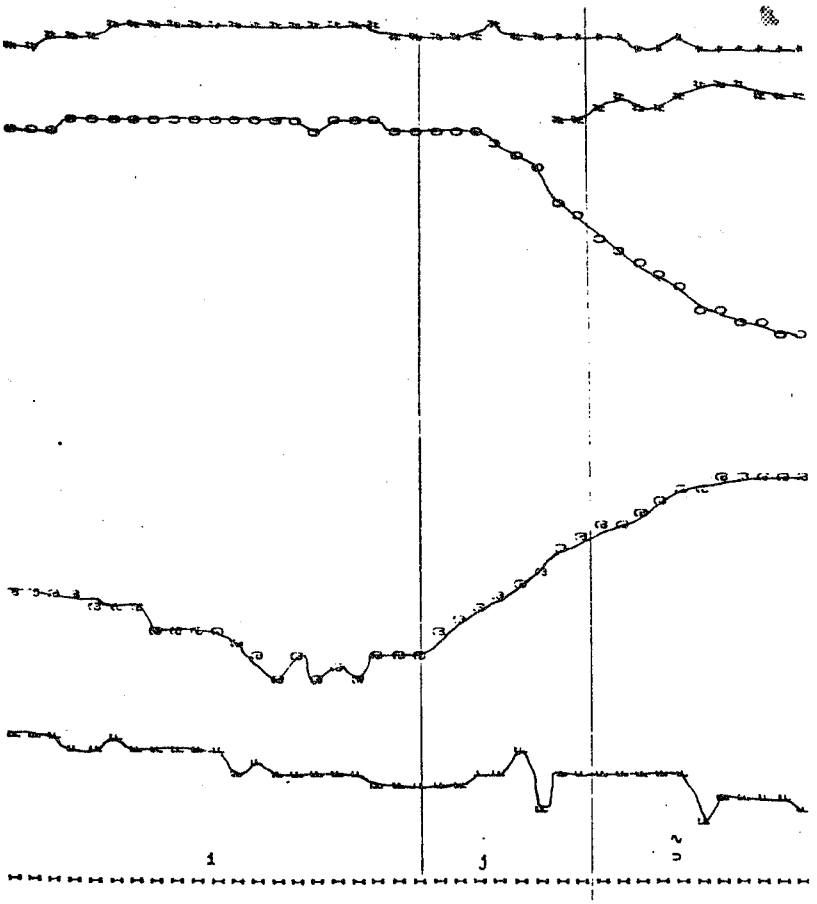


Figure 4 : SONAGRAMME DU SEGMENT [ij$\tilde{5}$] EXTRAIT DU MOT PAPILLON [LOCUTEUR DD].

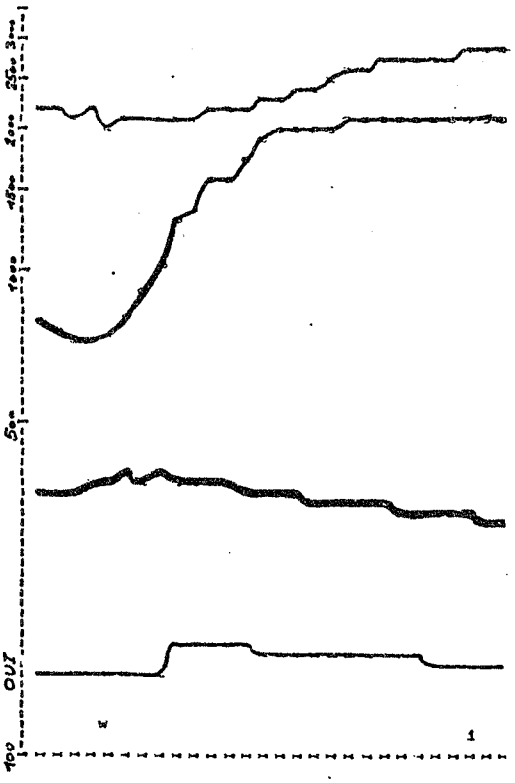


Figure 5 : SONAGRAMME DU MOT "OUI"

IV - FRICATIVES SONORES

Les fricatives sonores sont dues à l'excitation du canal vocal par les cordes vocales auquel s'ajoute un bruit de friction dû au passage d'un flux d'air en un point de resserrement du conduit vocal. Ce bruit de friction est important et masque très souvent la structure formantique de ces sons dans des analyses de type spectral.

Notre but est de montrer que la méthode d'analyse temporelle développée dans l'analyseur du projet A.R.I.A. met bien en évidence cette structure formantique et par là même, permet la distinction des fricatives sonores entre elles. La figure 6 représente la fricative sonore |v| dans le contexte "as-tu-vu".

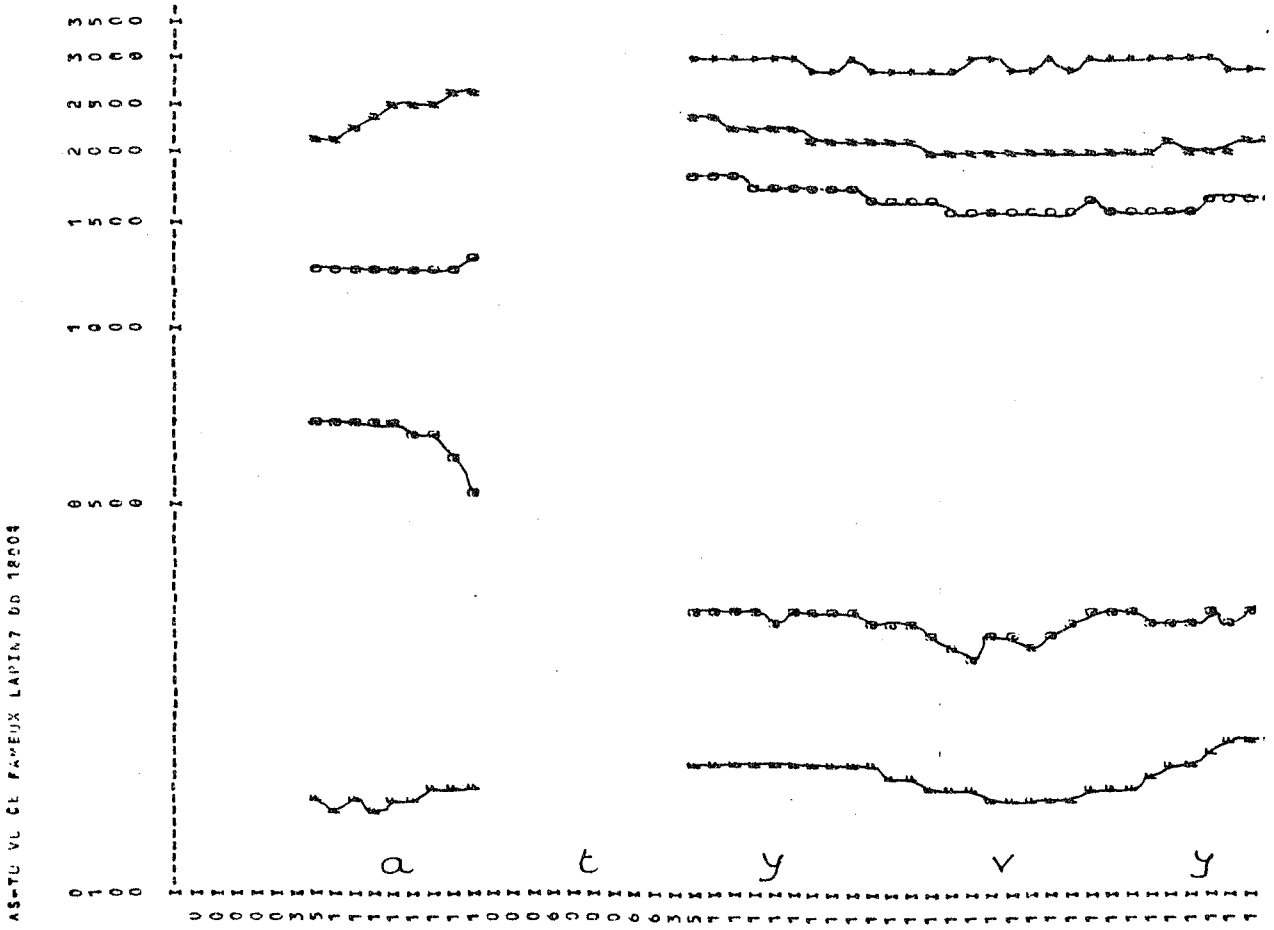


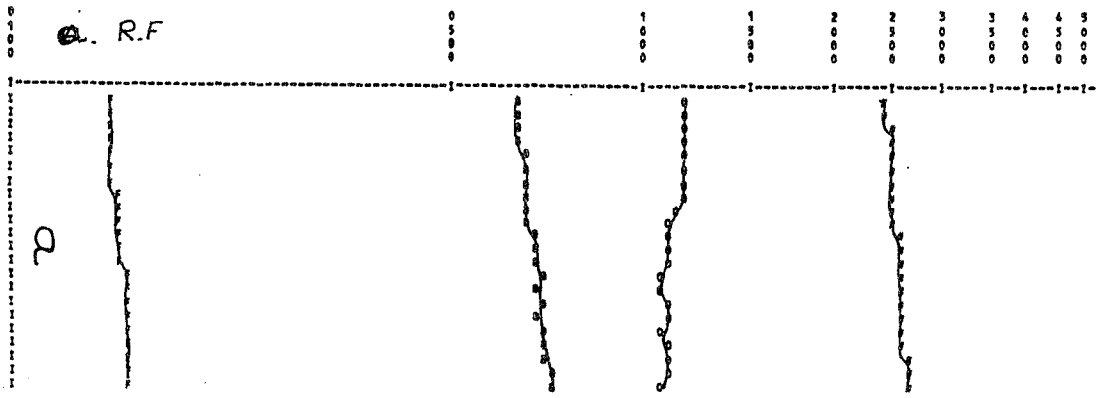
Figure 6

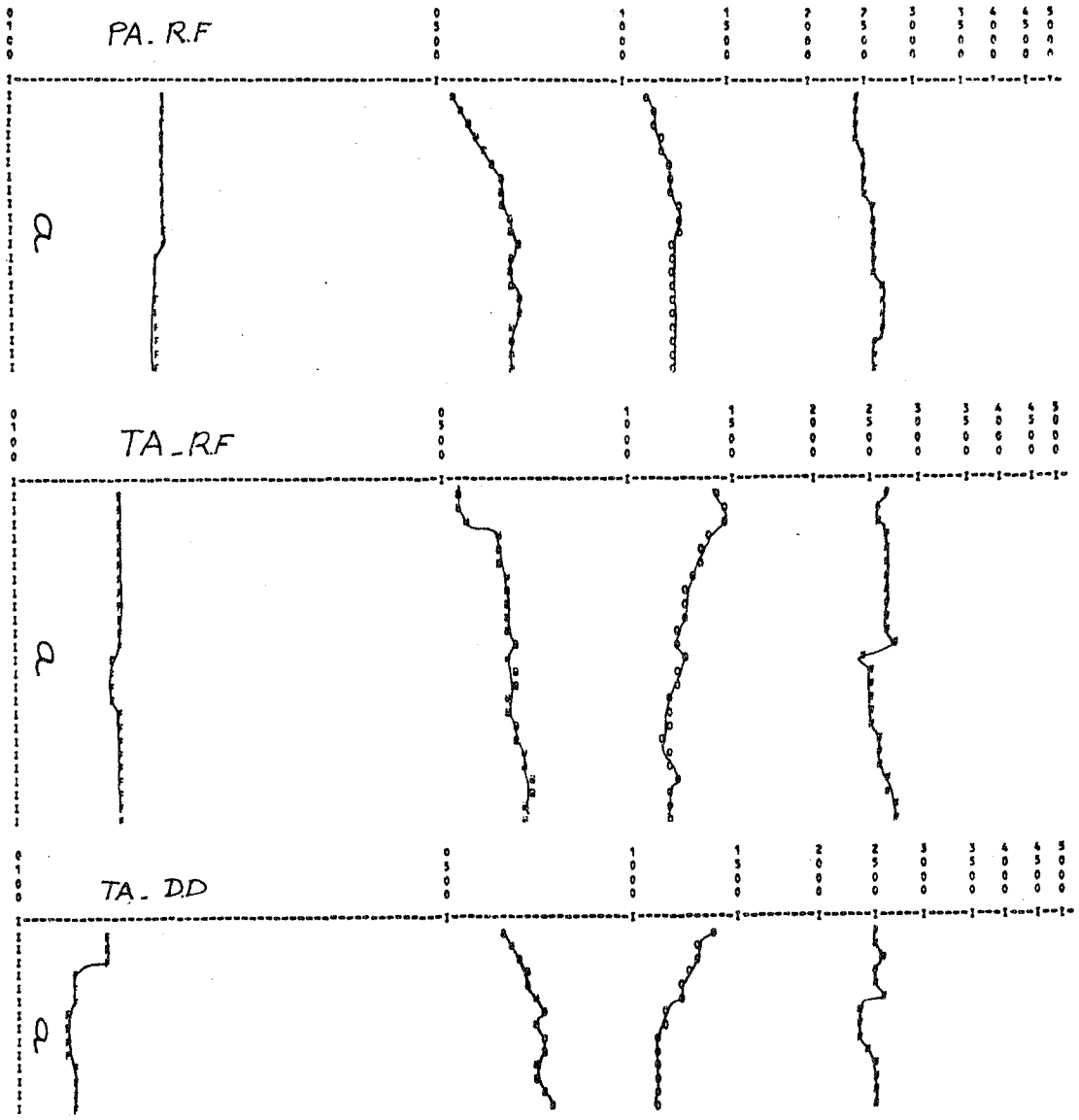
V - TRANSITIONS DE FORMANTS

Les travaux de synthèse réalisés aux laboratoires Haskins ont montré l'importance et le rôle des transitions de formants dans la caractérisation des consonnes. Diverses études |5|, |6| ont montré l'existence d'un ou de plusieurs points de convergence (Locus) de toutes les transitions de formant ayant perceptuellement un même lien d'articulation. Notre propos est de montrer que les transitions de formants sont bien mises en évidence dans le cas d'une analyse effectuée avec l'analyseur du projet A.R.I.A.

Les différentes figures présentées ci-après représentent les logatomes : [PA], [TA]. En l'absence de consonne, les formants F1 - F2 - F3 de la voyelle [a] ont normalement des fréquences moyennes d'environ 700, 1150, 2500 Hz. Mais on voit que dans le cas où la voyelle est précédée des consonnes [P] ou [T], au moment de l'explosion, les formants partent de valeurs de fréquence différentes de la normale. De même dans le cas d'une voyelle suivie de [T] (figure 6), les formants convergent vers des valeurs de fréquence différentes de la normale.

L'intérêt d'une telle analyse réside surtout dans la netteté avec laquelle ces différentes transitions apparaissent. Ceci est très utile dans le cas d'études sur la théorie du "locus" et pour une application à la reconnaissance automatique des plosives basée sur cette théorie.





VI - CONCLUSION

Dans le cas de sons produits par la source glottale, l'analyseur du projet A.R.I.A. fournit des paramètres formantiques dont la qualité essentielle est la grande précision avec laquelle ils sont obtenus.

Nous avons envisagé différents cas d'analyse phonétique qui mettent en relief : d'une part, les possibilités discriminantes des paramètres formantiques aussi bien dans le cas des voyelles orales que des consonnes sonores voisées ; d'autre part, la netteté des transitions de formants que ce soit dans le cas de semi-voyelles ou de consonnes plosives.

BIBLIOGRAPHIE

[1] - DOURS D - FACCA R - PERENNOU G

Analyse temporelle du signal vocal, comparée à l'analyse fréquentielle classique du point de vue de la reconnaissance.

5èmes Journées d'Etude sur la parole - Orasy Mai 74.

[2] - DOURS D - FACCA R

Méthode de segmentation et d'analyse par traitement direct du signal vocal.

Application à la classification et la reconnaissance des voyelles et des consonnes.

Thèses présentées à l'Université Paul Sabatier - Toulouse 74.

[3] - DOURS D - FACCA R

Méthode de segmentation et d'analyse par traitement direct du signal vocal.

Seminaire de linguistique et de reconnaissance de la parole

Université Paul Sabatier - Octobre 74.

[4] - FANT G

Acoustic theory of speech production

Mouton the Hague - Paris 70.

[5] - DELATTRE P - LIBERMAN A.M. - COOPER F.S.

Acoustic loci and transitionnal cues for consonnants.

JASA 27 - 1955 pp. 769 - 774.

[6] - EMERIT E

Nouvelle contribution à la théorie des "locus"

Phonetica 30 : 1-30 - 1974.

[7] - COLLOMB G - PERENNOU G

Recherche de repère et d'axes discriminants pour la représentation des voyelles

Rapport interne - Labo. CERFIA - Avril 1975.

[8] - PERENNOU G

Contribution à l'étude des discriminateurs. Calcul et optimisation.

Thèse de doctorat ès-sciences - Faculté des Sciences de Toulouse 1968.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

LA DETECTION DU VOISEMENT PAR LES PROPRIETES PHYSIQUES RESULTANT DE L'EXCITATION PERIODIQUE DU CONDUIT VOCAL : COMPARAISON STATISTIQUE DE TROIS PROCEDES

Christian ABRY Louis-Jean BOË et Jean-Frédéric ZURCHER
Institut de Phonétique-Grenoble CNET - Lannion

RESUME

Dans le cadre de l'étude du voisement ont été comparés trois types de détection de ce trait : la fonction de présence du fondamental D_1, la comparaison de deux zones spectrales D_2, le nombre de passages à zéro D_3. Sur deux corpus de français (850 sec. de parole continue et 1.188 consonnes en logatomes) les résultats montrent que seule une analyse par segments phonétiques permet de départager D_1, D_2 et D_3. Une classe d'erreurs [ʒ, z] peut être ainsi déterminée uniquement pour D_2, avec diagnostic. La validité de procédures aussi différentes peut-elle amener le phonéticien à reformuler la définition substantielle du trait de voisement ?

SUMMARY

A study of voicing, from the point of view of its detection, has been made with three apparatus : D_1 , D_2 , D_3 . D_1 detects presence or absence of fundamental, D_2 compares two bands of frequencies, D_3 counts zero crossings. From a corpus of 850 sec. of connected speech in French, our gross results are such that an analysis segment by segment (for 1,188 consonants within nonsense bisyllables) has been needed for true evaluation. A cluster of errors [ʒ, z] is so pointed out but for D_2 only. We attempted diagnostic. Such dissimilar procedures could perhaps cause the phonetician to revise his specification of voicing.

LA DETECTION DU VOISEMENT PAR LES PROPRIETES PHYSIQUES RESULTANT DE L'EXCITATION PERIODIQUE DU CONDUIT VOCAL : COMPARAISON STATISTIQUE DE TROIS PROCEDES

Christian ABRY Louis-Jean BOË et Jean-Frédéric ZURCHER
Institut de Phonétique-Grenoble CNET - Lannion

I - INTRODUCTION

Certains termes de classification de la production de la parole sont fournis par la spécification physiologique de la disposition et du fonctionnement des cordes vocales. Ainsi l'opposition *présence/absence* de vibrations laryngiennes reste traditionnellement liée à la définition du trait de sonorité. Récemment encore, CHOMSKY & HALLE[1-2] et LADEFOGED[3] par exemple ont adopté une telle optique. En fait le problème de la caractérisation de ce trait et de sa relation avec une réalité perceptuelle reste complexe : bien que ce ne soit généralement pas le cas, certaines réalisations de parole perçues sonores n'ont pas été reproduites avec une excitation vocale périodique et vice-versa[4-16]. Tout en remarquant qu'il s'agit là d'un cas particulier de production, certains sons de parole synthétique, générés uniquement avec une source de bruit, peuvent être classés dans la catégorie sonore et inversement il est possible d'obtenir une perception de sourdité avec une source harmonique[17-33]. Toutes ces études mettent en évidence l'importance d'autres indices parmi lesquels, dans certains cas, la durée semble occuper une place importante. L'un ou plusieurs d'entre eux interviennent dans la perception de la voix chuchotée et plus couramment dans les cas d'assimilation de voisement.

Afin d'éviter toute ambiguïté, nous utiliserons les appellations suivantes : *sonore/sourd* aux niveaux phonologique ou perceptuel et *voisé/non voisé* pour opposer, sur le plan de la réalisation (ici acoustique), la présence ou l'absence d'excitation périodique dans le signal de la parole.

La détection de ce trait s'est développée successivement dans deux domaines : la synthèse par vocodeur et la reconnaissance automatique de la parole. Les multiples procédures qui ont été élaborées visent toutes en fait à mettre en évidence, dans le signal de la parole, la présence d'une excitation périodique :

. directement, par des procédés de détection de périodicité :
  - calcul direct de cette périodicité[34-35].
  - autocorrélation[36-40] et AMDF (Average Magnitude Difference Function)[39].
  - cepstre[41-42].
  - filtrage inverse par codage prédictif[43-45] et méthode basée sur un critère d'évolution ou de dispersion énergétique[46-48].

. indirectement, en exploitant une des propriétés intrinsèques du signal de la parole :

- Compte-tenu des caractéristiques de la source vocale et de celles de bruit, les sons voisés présentent essentiellement leur énergie dans une bande de fréquences relativement basse alors que c'est le contraire pour les sons non voisés. Pour ceux qui sont produits à la fois par une excitation périodique et une source de bruit, les spectres discret et continu ne se chevauchent pas dans le domaine excursionné par le fondamental. Par exemple pour le français[49] la présence d'énergie, dans une bande de fréquences maximale de 55 Hz à 180 Hz pour une voix d'homme et 85 Hz à 331 Hz pour une voix de femme, correspond à une excitation périodique[50-51]. Ce procédé est assez simple mais sensible aux bruits parasites, il nécessite un réglage précis en fonction du niveau. Pour éviter les inconvénients d'une mesure absolue, le résultat de comparaisons énergétiques a été souvent utilisé pour la décision de voisement[52-56]. Bien que procédant tout à fait différemment, le comptage du nombre de passages à zéro[57-58] et le calcul du rapport de la valeur de la fonction d'autocorrélation correspondant à un décalage de la période d'échantillonnage T_e à l'énergie totale $\frac{C_{xx}(T_e)}{C_{xx}(0)}$[40] exploitent les mêmes propriétés.
- la comparaison de la forme des signaux périodiques avec celle des bruits de friction ou d'occlusion permet aussi une bonne appréciation de la fonction de voisement[59].

Pour un exposé plus détaillé de plusieurs de ces méthodes, on peut se reporter à l'ouvrage de PIROGOV[60] (pp. 209-259).

Souvent les publications font état d'avantages et d'inconvénients relatifs aux méthodes utilisées, mais à notre connaissance il n'existe pas d'études comparatives systématiques. Le travail de LILJENCRANTS[61] unique en son genre, et qui porte précisément sur les trois procédés de détection que nous allons tester ci-dessous, n'est qu'indicatif : il porte sur un corpus de faible taille (6 secondes de parole continue, 2 locuteurs, soit un total de 130 réalisations phonétiques). Et pourtant une telle procédure outre son intérêt évident comme méthode d'évaluation pourrait être utilisée en reconnaissance de la parole (analyse des erreurs).

Dans le cadre de la collaboration entre le Département ETA du CNET de Lannion et l'Institut de Phonétique de Grenoble, nous avons opéré une comparaison entre trois modes de détection, utilisés dans des appareillages qui ont été élaborés à des fins différentes : l'étude des faits prosodiques[62-64] et la synthèse par vocodeur[65]. Dans le premier cas la fonction de voisement est en fait une fonction de présence d'excitation périodique (figure 1), dans le deuxième il s'agit d'un résultat portant sur une comparaison énergétique (figure 2), et dans le troisième du comptage des passages à zéro (figure 3).

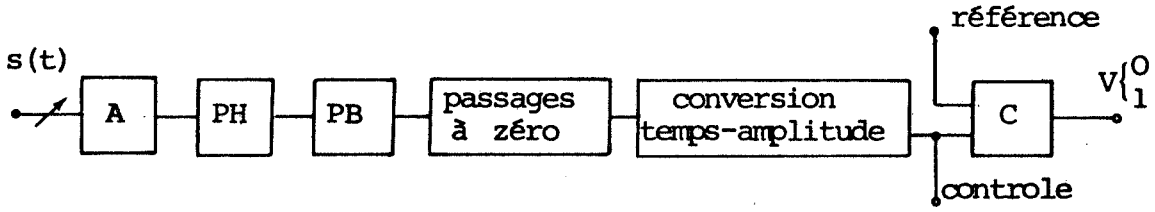


Figure 1 - Mode de détection D_1.

Grâce à la sortie contrôle, visualisée à l'aide d'un inscripteur l'ajustage du gain de l'amplificateur A d'entrée et le réglage manuel du passe-bas PB permettent d'extraire de s(t) le fondamental et lui seul. Le passe-haut PH est utilisé afin d'éliminer éventuellement les signaux parasites. Une conversion *intervalle de temps-amplitude*, déclenchée par les passages à zéro, délivre un signal fonction de la fréquence laryngienne F_ℓ. La comparaison C avec une référence (limite inférieure de F_ℓ) permet d'obtenir la fonction de voisement (V = 1 si $F_\ell \geq 50$ Hz; V = 0 si $F_\ell < 50$ Hz ou en l'absence de signal de parole).

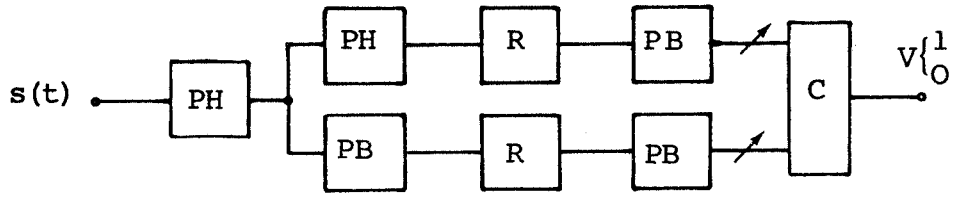


Figure 2 - Mode de détection de D_2.

Après atténuation des bruits parasites (premier PH), les énergies obtenues en de-ça de 700 Hz et au de-là de 1500 Hz, (filtrages PH et PB, redressement R et intégration avec PB) sont comparées par C. Il s'agit donc d'une estimation sur la répartition basse et haute fréquence de l'énergie.

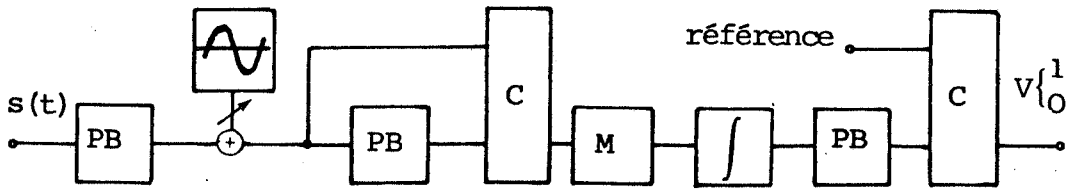


Figure 3 - Mode de détection D_3.

Au signal de la parole, filtré par le passe-bande PB (f_c=100 Hz et 450 Hz), est ajouté, après pondération, une tension sinusoïdale à 10 kHz (ceci afin d'obtenir une décision V=0 en cas d'absence de s(t)). Le premier comparateur C permet de déclencher le monostable M à chaque franchissement de la valeur moyenne (le filtre passe-bas PB a pour fréquence de coupure 250 Hz). Les impulsions (de durée constante = 0,8 ms.) issues de M sont intégrées puis lissées par un deuxième passe-bas (f_c=35 Hz). La tension ainsi obtenue est comparée à une référence. La décision opérée par C délivre la fonction de voisement V.

Les trois modes de détection ont été réalisés sous forme analogique. Ces trois appareils ont été réglés après de très nombreux essais et modifications, nous pouvons estimer qu'ils présentent une bonne illustration de D_1, D_2 et D_3 .

PROCEDURE EXPERIMENTALE ET RESULTATS

Après étalonnage et réglage des trois appareils (ajustage des seuils de déclenchement) nous avons utilisé le montage suivant (figure 4) :

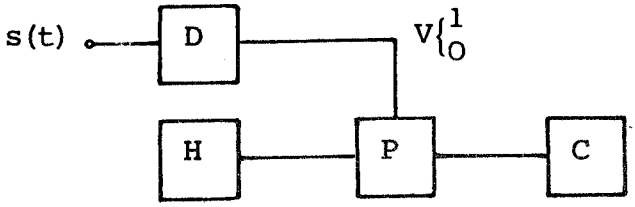


Figure 4 - Montage.

Le compteur C totalise les impulsions délivrées par l'horloge H (1.000 Hz) quand la porte P est ouverte, ce qui est le cas lorsque, à la sortie du détecteur D la fonction de voisement V = 1.

1 - Comparaison globale et qualitative

Pour ce premier test s(t) a été un corpus constitué d'un texte lu par 5 femmes et 5 hommes : nous avons relevé en secondes pour chaque locuteur la durée totale de l'enregistrement et les durées voisées détectées par D_1, D_2 et D_3 (tableau 1). On peut constater qu'en ce qui concerne ces dernières mesures, il existe une différence systématique, bien que peu importante, au détriment de D_2 et D_3 par rapport à D_1; nous l'avons chiffrée par Θ, différence entre les durées affichées par D_1 et D_2 et par D_1 et D_3, exprimée en % par rapport à la durée affichée par D_1. On peut noter qu'en moyenne Θ est peu différent pour les femmes (locuteurs 1 à 5) et pour les hommes (locuteurs 6 à 10). Connaissant $\bar{\Theta}$ (D_1,D_2) et $\bar{\Theta}$ (D_1,D_3) on peut constater que D_2 et D_3 sont en moyenne plus proches l'un de l'autre que de D_1.

Pourtant nous allons voir par la suite que le comportement de D_3 est beaucoup plus proche qualitativement de celui de D_1 que de D_2.

C'est afin de pouvoir interpréter cette mesure globale que nous avons inscrit (oscillomink à jet d'encre SIEMENS) simultanément la fonction de voisement et l'oscillogramme du signal de la parole. Après délimitation et transcription phonétique étroite, nous avons relevé les différences de résultats :

1) entre D_1 et D_2 (ne sont portées sur le tableau 2 que celles qui portent sur toute la durée de réalisation).

locuteur	D_1	D_2	D_3	Θ (D_1,D_2)	Θ (D_1,D_3)
1	48,86	47,20	47,39	3,8	3,0
2	41,75	39,72	40,76	4,9	2,4
fém. 3	44,91	42,51	42,30	5,3	5,8
4	50,06	48,50	49,97	3,1	0,2
5	49,99	48,71	46,83	2,6	6,3
				$\bar{\Theta}$ fem.=3,8	$\bar{\Theta}$ fem=3,5
				$\bar{\Theta}$ =4,9	$\bar{\Theta}$ =3,7
				$\bar{\Theta}$ masc.=5,9	$\bar{\Theta}$ masc=3,8
6	48,47	43,78	46,94	9,7	3,2
7	45,36	45,24	43,78	0,3	3,5
masc.8	58,83	54,90	55,87	6,7	5,0
9	47,71	44,00	44,68	7,8	6,3
10	49,44	47,05	48,93	4,8	1,0

Tableau 1

	sons non voisés							sons voisés											
	p	b̥	t	k	ʀ̥	w̥	ɥ̥	b	d	t̬	g	v	z	s	ʒ	ʀ	l	j	i	ə
$V(D_1)=0$ $V(D_2)=1$	2		1	4	39	2	1	1											
$V(D_1)=1$ $V(D_2)=0$		1			5			10	7	1	3	5	35	4	11	4	1	2	1	1

Tableau 2

Occurrences des différences de détection D_1, D_2

▨ erreur pour D_1 ▨ erreur pour D_2

Les sons qui figurent sur les tableaux 2 (et 3) peuvent être affectés d'un signe diacritique dans la mesure où perçus sourds ou sonores leur oscillogramme présente ou ne présente pas de trace d'excitation périodique. Ainsi [b̥] est bien perçu comme sonore, mais a été réalisé sans excitation vocale, pour [t̬] il s'agit de l'inverse. Les zones hachurées font apparaître symétriquement les "erreurs" respectives.

Pour D_1 elles sont assez faibles (7 au total), pour D_2 relativement plus nombreuses, bien que limitées par rapport au nombre total des réalisations. Nous n'avons pas jugé utile et surtout rentable - vu l'utilisation purement indicative d'un tel tableau - d'exprimer tous les résultats en pourcentages. Nous ne les avons estimés que pour les valeurs les plus élevées soit : [z] (43%), [ʒ] (26%) et [b] (6%) (nous retrouverons plus loin les deux premiers sons en tête du peloton). Pour [R] dévoisé (surtout après une consonne sourde de même syllabe) le nombre d'occurrences n'est pas facilement estimable à cause de sa variabilité d'apparition et cela pour un même locuteur. On peut noter que c'est une "erreur" qui, si elle se révélait suffisamment constante, mériterait d'être exploitée : la détection se ferait alors au niveau du phonème /R/ et non de ses réalisations voisées ou non voisées.
Nous avons constaté d'autre part (ce qui n'est pas consigné dans le tableau) qu'assez souvent $V(D_2) = 1$ pour l'explosion de [k] (21 occurrences, y compris les 4 qui apparaissent dans le tableau). Contrairement à ce que nous avons avancé pour [R] il ne semble pas qu'une utilisation évidente puisse en être faite en reconnaissance.

2) entre D_1 et D_3 (sur le tableau 3 figurent tous les sons qui sont affectés par un manque à détecter portant au moins sur le quart de leur durée).

	sons non voisés			sons voisés													
	t	s	R̥	b	d	tˇ	g	kˇ	v	z	sˇ	ʒ	R	j	a	u	ə	ɑ̃
$V(D_1)=0$ $V(D_3)=1$		1	1	1	3	2	1	1	1	1	3		3			1	2	
$V(D_1)=1$ $V(D_3)=0$	1		3	1	2			1	5	2	2	7	1	5		2	1	

Tableau 3

Occurrences des différences de détection D_1, D_3

▨ erreur pour D_1 ▨ erreur pour D_3

On remarque tout de suite, malgré la prise en compte des erreurs partielles, le peu de signification des valeurs relevées. Celles imputables à D_3 dans ce cas, représentent simplement 0,78% des occurrences de [ɑ] + [ɑ̃], 1,16% de [R] + [R̥], 3,25% de [s] + [sˇ] (essentiellement sur un seul locuteur féminin) et 6,25% de [z] (ce dernier pourcentage sur 80 occurrences seulement).

2 - Comparaison par unité phonétique

Dans ces conditions nous n'avons élaboré un corpus test que pour D_2. Il est bien entendu plus facile d'obtenir systématiquement un corpus de réalisations voisées que de réalisations dévoisées [b, R, w, ɥ, etc.]. D'autre part, comme on peut le remarquer, les erreurs sont

faibles pour les voyelles (voir plus loin le cas de [a]). Ces deux faits nous ont amené à constituer un corpus de 33 logatomes CVCVC (C et V étant identiques pour un même logatome) avec :

C = [b.d.g,v,z,ʒ,m,n.l.R j] et V limité à [i.a.u]

La détection de [R] n'a été testée que pour les réalisations non voisées faisant partie du corpus concernant [R] normalement voisé.

Les logatomes ont été lus deux fois et enregistrés par 6 locuteurs (3 femmes et 3 hommes). Au total chaque consonne a donc été réalisée 108 fois dans 3 positions (initiale, intervocalique et finale) et pour 3 entourages vocaliques; chaque voyelle l'étant 264 fois pour 2 positions et un entourage consonantique.

Les fonctions de voisement issues de D_1 et D_2 ont été inscrites en même temps que l'oscillogramme du signal de la parole. Si l'on met à part le cas de [R] , D_1 affiche systématiquement V = 1 alors que D_2 présente des manques, ce que les résultats précédents laissaient prévoir. Afin de pouvoir les préciser nous avons relevé par position, entourage et pour chaque réalisation, la durée et le temps pendant lequel D_2 aurait dû afficher V = 1 et ne l'a pas fait, ce que nous appelons "manque". En vue d'une meilleure appréciation, nous avons précisé par son, le nombre de réalisations totalement détectées et le nombre de celles qui ont été entièrement non détectées.

Le tableau 4 présente pour chaque son, analysé par D_2 (indépendamment de sa position et de son entourage vocalique) :

- la durée totale des manques exprimée en pourcentage par rapport à la durée totale de toutes les réalisations de ce même son (colonne 1).
- le nombre d'unités qui, tout au long de leur réalisation, ont été correctement détectées voisées (colonne 2) et le nombre de celles pour lesquelles la détection a été totalement manquée (colonne 3); ces deux valeurs sont exprimées par rapport au nombre total de chaque réalisation. Nous donnons ces deux derniers résultats parce qu'il peut ne pas être indifférent pour la reconnaissance-synthèse de connaître la répartition des manques.

La première remarque quantitative qui s'impose est que les pourcentages parfois assez élevés de la colonne 1 sont rachetés par ceux de la colonne 3. Par exemple si la durée des manques de [ʒ] représente 72% du temps de réalisation, il n'y a que 37% de ces sons pour lesquels le voisement aura été véritablement "oublié". Ce sacrifice étant fait, il en reste 51% qui ne présentent qu'un manque partiel et 12% qui n'ont aucun manque. Pour ceux qui présentent un manque partiel, la question est de savoir quel est en moyenne le pourcentage de durée de celui-ci. Il peut être évalué en faisant l'hypothèse que la durée de chaque son est à peu près la même quelle que soit la durée du manque (Hypothèse non infirmée, les sons les moins bien détectés n'étant pas, par exemple, les plus longs ou les plus

	durée des manques	sons totalement bien détectés	sons totalement non détectés
ʒ	72	12	37
z	59	16	17
d	21	44	5
R	19	56	8
g	18	44	4
b	16	48	3
v	15	50	2
j	12	53	3
n	6	59	1
l	4	72	1
m	3	73	0

Tableau 4

Les résultats sont exprimés en pourcentages : pour la durée totale des manques par rapport à la durée totale de réalisation, et pour le nombre d'unités totalement détectées ou non par rapport au total des réalisations. Pour [R] il s'agit de la partie voisée.

brefs). Dans le cas de [ʒ], à partir des relevés du tableau 4, on obtient 69% ce qui correspond tout de même à une moyenne de détection correcte de 31%, soit près du tiers de la réalisation. Comme nous l'avons relevé plus haut, on sait que la sonorité peut être perçue avec 0% de voisement, mais il n'existe pas d'estimation du seuil de durée de voisement nécessaire et suffisant *à lui seul* pour percevoir un [ʒ] un [z], etc., comme sonores. Ce n'est pourtant qu'avec de telles données que l'on pourrait préciser l'impact à la perception d'un manque de voisement dans les pires conditions (l'indice de voisement étant lui seul porteur du trait de sonorité). Ces conditions limites ne semblent pouvoir apparaître qu'expérimentalement. C'est pourquoi on ne devrait tenir compte de cet appauvrissement que dans la mesure où il pourrait amoindrir la résistivité du trait de sonorité (haute selon MILLER & NICELY[66] et autres[67-69]) au bruit dans des conditions *ex vitro*. La première colonne du tableau 4 a servi à ordonner les consonnes selon leur degré croissant de difficultés pour D_2. On constate que cet ordre correspond pratiquement (sauf dans le cas particulier de [R]) à celui que l'on aurait obtenu avec les deux autres colonnes. Connaissant le pourcentage moyen de manques pour une consonne, il est possible de donner une estimation des occurrences totalement détectées ou totalement manquées (figure 5, Annexe).

En prenant comme base la première colonne, on constate un saut très net entre [z] et [d] (même chose pour la colonne 3) et entre [j] et [n] (moins sensible pour les autre colonnes) ce qui aboutit à la tripartition :

[ʒ, z] [d, g, b, v, j] et [n, l, m]

Une analyse de la matrice des différences entre erreurs (figure 6, Annexe : où les valeurs de chaque cellule représentent une distance d satisfaisant aux conditions : $d(x,x) = 0$; $d(x,y) = d(y,x)$; $d(x,y) \leq d(x,y) + d(y,z)$)par les deux algorithmes de classification hiérarchique[70], connexité et diamètre[71], donne deux arbres topologiquement presque identiques, au classement de [j]et [d]près, dans le groupe central; les trois grands **groupes** restant inchangés (figure 7, Annexe).

Les deux classes extrêmes ne sont pas sans naturalité phonétique, soit sur le plan articulatoire :

- fricatives en creux (*groove fricatives*)
- nasales et latérales,

(les têtes de file de la classe centrale étant les occlusives). Mais c'est évidemment sur le plan acoustique, le plus intéressant pour un type de détection tel que D_2, qu'on peut constater que les sons les plus difficiles à détecter voisés [ʒ]et [z], sont précisément les seuls à se caractériser par une source de bruit d'intensité importante au-dessus de 1500 Hz[72]. VOIERS[73] les détermine au moyen du trait de *sibilation*.

Ce classement suggère une possibilité d'utilisation statistique des divergences entre D_1 (ou D_3) et D_2 pour une première partition entre les consonnes voisées :

- bien détectées par D_1 (ou D_3) : [ʒ, z, d, ... n, l, m]
- bien détectées par D_1 (ou D_3) et mal par D_2 : [ʒ, z]
- bien détectées par D_1 (ou D_3) et D_2 : [d,R,g,b,v,j,n,l,m]

Le tableau 5a examine, en complément, la distribution des manques dans les trois positions (initiale, intervocalique et finale) quel que soit l'entourage vocalique. On peut noter que la bipartition [ʒ, z], [les autres sons] est toujours aussi nette, qu'à part [R] l'ordre est à peu près le même que dans le tableau 4, colonne 1, soit [ʒ] et [z] en tête, [n,l,m] à la queue et [d,v,g,j] insérant [b] tantôt en seconde, dernière ou troisième position. La tendance générale s'exprime dans le sous-tableau 5b qui donne le pourcentage de la durée totale des manques pour toutes les réalisations exprimé par rapport à leur durée totale. On constate un net accroissement des erreurs en position initiale. Cette tendance générale ([R] est toujours à part : 17% contre 19% en moyenne, tableau 4, colonne 1) n'est heureusement pas suivie par les plus concernés [ʒ] (65% à l'initiale contre 72% en moyenne) et [z] (51% contre 59%).

Le tableau 6a examine l'incidence sur la détection de l'entourage vocalique : [ʒ] et [z] sont toujours nettement en première position et, à la fin, nous retrouvons [n,l,m], l'ordre intermédiaire étant variable. La tendance générale, résumée dans le sous-tableau 6b, montre que l'assimilation (ou accomodation) vocalique, normale en français (les voyelles y assimilent les consonnes), joue surtout :

- dans le sens d'une amélioration avec [u] (renforcement des

	initiale	intervoc.	finale
b	32	6	13
d	34	14	15
g	26	12	15
v	24	10	13
z	51	63	63
ʒ	65	77	74
m	7	0	1
n	10	5	3
l	7	4	2
R	17	26	16
j	16	9	12

Tableau 5a

Pourcentages de la durée des manques suivant la position

	i	a	u
b	19	23	8
d	23	32	9
g	28	17	6
v	23	14	8
z	64	65	49
ʒ	86	84	48
m	2	3	2
n	5	9	3
l	5	4	2
R	27	25	7
j	6	18	14

Tableau 6a

Pourcentages de la durée des manques suivant l'entourage vocalique

initiale	intervoc.	finale
28	19	19

Tableau 5b

Pourcentages moyens pour l'ensemble des sons suivant la position.

i	a	u
22	27	14

Tableau 6b

Pourcentages moyens pour l'ensemble des sons suivant l'entourage vocalique.

i	a	u
9	13	1

Tableau 6c

Pourcentages d'occurrences de manques totaux ou partiels pour les trois voyelles.

basses fréquences).

- défavorablement avec [a] (voyelle dont les fréquences des 2 premiers formants, à la limite des filtrages utilisés par D_2, 700 Hz et 1500 Hz, font parfois basculer la comparaison en faveur du non-voisement) alors que pour [i], cela se produit seulement dans les cas où l'énergie des 2e et 3e formants est importante par rapport à celle du premier.

[ʒ] et [z] suivent bien cette tendance générale (ce qui n'est pas le cas pour [j] avec [i] : 6% contre 12% en moyenne, tableau 4, colonne 1); on peut noter qu'avec [u] leurs scores s'égalisent (48% et 49%), ce qui représente une très nette amélioration pour [ʒ] ([j] affiche bien entendu un comportement inverse de celui qu'il avait avec [i] : il se place immédiatement après [z]).

A titre indicatif, nous avons relevé les occurrences de manques totaux ou partiels pour les voyelles [i, a, u] (cf. tableau 6c). Si on considère que les résultats du tableau 6b peuvent représenter le pourcentage d'*unités* entièrement ou partiellement manquées (si les durées moyennes des trois classes - non manquées, partiellement manquées, totalement manquées - ne sont pas différentes), on constate que les erreurs se distribuent de la même manière sur les voyelles que sur les consonnes assimilées par ces voyelles ([a] en tête suivi de [i] puis, loin derrrière, [u]). On remarque que ces erreurs sont tout de même moins élevées. Ceci n'est qu'indicatif dans la mesure où le corpus n'a pas été élaboré pour tester les voyelles.

Enfin nous nous proposions d'examiner le comportement de [R] , dont nous avons souvent noté le particularisme. Il n'a été relevé que 7 détections par D_2 de segments non voisés [R] (surtout en finale) et toutes avec [u] . Aucune conclusion ne peut en être tirée, il reste à élaborer un corpus de [R] pour tester véritablement D_2 sur cette possibilité.

CONCLUSION

En résumé, les trois procédés sont de qualité suffisante (cf. les Θ , tableau 1) pour avoir été testés sur le détail de leurs performances. Au vu de celles-ci, D_1 et D_3 présentent un forte équivalence : il n'est pas possible de dégager, en ce qui les concerne, de classes d'erreurs. Par contre D_2 affecte essentiellement la détection des [z] et [ʒ]: environ 60% à 70% de durée de manque et 20% à 40% d'occurrences totalement non reconnues voisées, avec influence de l'entourage et de la position (à deux exceptions près , [z] et [ʒ] précisément, pour lesquels cette dernière n'intervient pas).

Ces résultats confirment et précisent tout à fait ceux de LILJENCRANTS[61] :"The conclusion will be that method [D_2] is inferior to methods [D_3] and [D_1] ...".

On aurait pu s'attendre à ce que les erreurs observées pour D_2 apparaissent aussi pour D_3, puisque tous deux opérent à partir de la répartition fréquentielle de l'énergie. Mais en fait D_3 prend en compte plus précisément la répartition de l'énergie tout au long du spectre : le nombre de passage à zéro est déterminé à la fois par l'énergie des composantes et par leur fréquence[74]. Or les [z] et [ʒ] pour lesquels D_2 présentait des erreurs ont bien sûr une énergie au-delà de 1500 Hz importante, mais celle-ci est largement étalée.

Ces symptômes de D_2 réclameraient un véritable diagnostic avec étude spectrale précise des sons cause d'erreurs. Ainsi, on pourrait peut-être mieux ajuster les fréquences des filtres ou même développer un processus plus élaboré (filtres suiveurs, comparateur multiple, etc.). Mais notre propos est pour l'instant, moins de montrer les limites (voire les impasses) techniques d'un procédé, que d'apprécier l'intérêt de son principe pour la classification des sons.

En réalité le critère utilisé par D_2 ne recouvre-t-il pas d'autres traits déjà utilisés en phonologie? A première vue la classification qu'il opère semble procéder de la définition donnée par JAKOBSON & al.[75] de l'opposition *grave/aigu*, soit à la lettre :"the predominance of one side of the significant part of the spectrum over the other" (ibid., p. 29). Ainsi on pourrait appliquer à [z, ʒ] / [v] les étiquettes *aigu/grave*[69] , ce qui est justifié par l'analyse des erreurs. En réalité, pour D_2 cette dimension n'est pas dissociée de *compact/diffus* (cf. les erreurs sur [a] et sur l'explosion de [k]). Il est bien évident que si le trait *grave/aigu* peut rendre compte grossièrement d'une certaine classification par les erreurs - mais celles-ci sont minimes - la majeure partie des résultats donne en réalité une mesure statistique (appréciable pour le phonéticien) du recoupement des dichotomies obtenues par D_1 et D_2. On peut ainsi mesurer la vérité de la proposition : il est possible de faire une décision sur le voisement par une comparaison des énergies pondérées en deçà de 700 Hz et au-delà de 1500 Hz. Disons-le franchement, ces résultats ne sont pas assez mauvais pour que l'on soit amené à rejeter D_2 comme autre manière de parvenir au classement *voisé/non voisé*. D'autre part, et complémentairement, D_3 nous permet de tester la proposition : il n'existe pas de sons voisés (en français, bien entendu) qui présentent une énergie supérieure à 1500 Hz importante et pour lesquels celle-ci ne soit pas étalée. Il resterait à conduire un ensemble de tests pour savoir si l'appareil perceptuel humain opère à partir de la *répartition* de l'énergie pour détecter le trait de sonorité. Si c'était le cas l'intérêt des phonéticiens pour de tels systèmes serait grandissant : ceux-ci restent en effet toujours beaucoup plus préoccupés par le vol des oiseaux que par celui des avions.

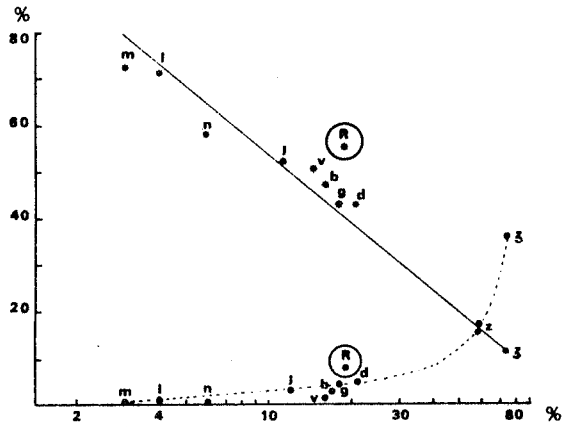


Figure 5

* : % d'unités totalement bien détectées
● : % d'unités totalement non détectées

en fonction du pourcentage de la durée des manques.

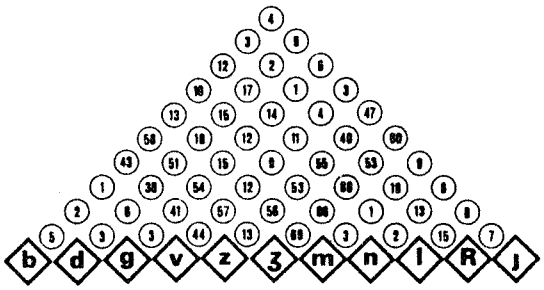


Figure 6 - Pyramide
des distances entre erreurs

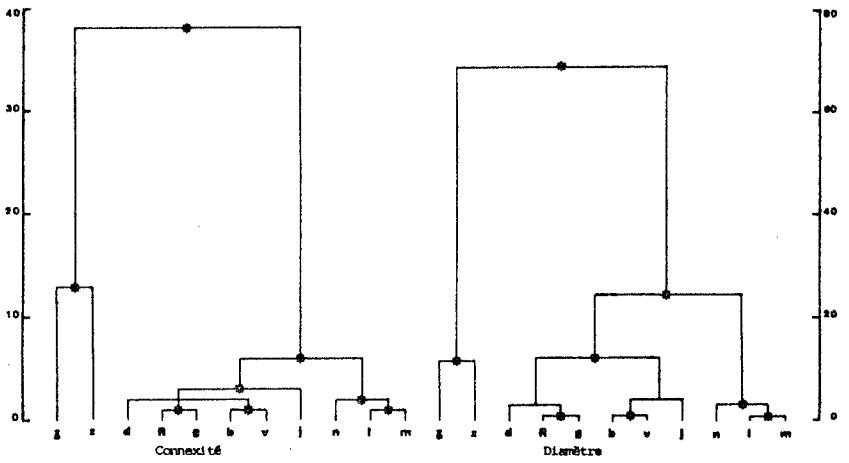


Figure 7 - Classement hiérarchique
par les différences d'erreur.

[1] CHOMSKY (N.) & HALLE (M.), 1968, *The Sound Pattern of English.* - Harper & Row Pub. New-York.

[2] CHOMSKY (N.) & HALLE (M.), 1973, *Principes de phonologie générative.* - 1° & 4° part. "The Sound Pattern of English". Seuil, Paris.

[3] LADEFOGED (P.), 1971, *Preliminaries to Linguistic Phonetics.* - The University of Chicago Press, Chicago, London.

[4] DENES (P.), 1955, *Effect of Duration on the Perception of Voicing.* - J.A.S.A. 27, pp. 761-764.

[5] LISKER (L.), 1957, *Closure Duration and the Intervocalic Voiced-Voiceless Distinction in Enalish.* - Language 33, pp.42-49.

[6] SATO (T.), 1958, *On the Differences in Time Structures of Voiced and Unvoiced Stop Consonants.* - J.A.S.J. 14, p. 117.

[7] LISKER (L.) & ABRAMSON (A.S.), 1964, *A Cross-Language Study of Voicing in Initial Stops : Acoustical Measurements.* - Word 20, pp. 384-422.

[8] LISKER (L.) & ABRAMSON (A.S.), 1967a, *The Voicing Dimension : Some Experiments in Comparative Phonetics.* - Proc. 6th Int. Congr. Phonetic. Sci., Prague, pp. 563-567.

[9] FINTOFT (K.) & SELNES (O.), 1971, *Duration as a Cue for the Perception of Voicing and Length.* - Proc. 7th Int. Congr. Acoustics, Budapest, Paper 20 C3.

[10] WAJSKOP (M.), 1971, *Identification des occlusives intervocaliques en français.* - R.A. 5, pp. 102-123, Institut de Phonétique de Bruxelles.

[11] SWEERTS (J.) & WAJSKOP (M.), 1972/73, *Les Indices du voisement dans les consonnes occlusives orales.* - R.A. 7/1, pp. 35-49 Institut de Phonétique de Bruxelles.

[12] CALAQUE (E.), 1973, *A propos de la réalisation de /z/ français par des locuteurs espagnols. Un problème de voisement.* - Bulletin de l'Institut de Phonétique de Grenoble 2, pp. 163-179.

[13] KLATT (D.H.), 1973, *Voice-Onset Time, Frication and Aspiration in Word-Initial Consonants Clusters.*-QPR-MIT 109, pp.124-136.

[14] WAJSKOP (M.) & SWEERTS (J.), 1973, *Voicing Cues in Oral Stop Consonants.* - Journal of Phonetics 1, pp. 121-130

[15] SERNICLAES (W.), 1973, *La Simultanéité des indices dans la perception du voisement des occlusives.* - C.R. 4° Journées d'Etudes du Groupe Communication Parlée du GALF, Bruxelles, pp. 359-367.

[16] SERNICLAES (W.) & BEJSTER (P.), 1974, *Influence du contexte vocalique sur la perception du voisement des occlusives.* - C.R. 5° Journées d'Etudes du Groupe Communication Parlée du GALF, Orsay.1, pp. 10-18.

[17] LIBERMAN (A.M.), DELATTRE (P.C.) & COOPER (F.S.), 1952, *The Role of Selected Stimulus Variables in the Perception of the Unvoiced Stop Consonants.* - Amer. J. Psychol. 65, pp. 497-517.

[18] LIBERMAN (A.M.), DELATTRE (P.C.) & COOPER (F.S.), 1958, *Some Cues for the Distinction between Voiced and Voiceless Stops in Initial Position.* - Language and Speech 1, pp. 153-167.

[19] DELATTRE (P.C.), LIBERMAN (A.M.) & COOPER (F.S.), 1955, *Acoustic Loci and Transitional Cues for Consonants.* - J.A.S.A. 27, pp. 769-773.

[20] LIBERMAN (A.M.), 1957, *Some Results of Research in Speech Perception.*- J.A.S.A. 29, pp. 117-123.

[21] DELATTRE (P.), 1958, *Les Indices acoustiques de la parole.* - Phonetica 2, pp. 108-118, 226-251.

[22] DELATTRE (P.), 1965, *Comparing the Phonetic Features of English, French, German and Spanish.* - Julius Groos Verlag, Heidelberg.

[23] DELATTRE (P.), 1967, *Des Indices acoustiques aux traits pertinents.* - Proc. 6th Int. Congr. Phonet. Sci., Prague, pp. 35-47.

[24] FUJIMURA (O.), 1961, *Some Synthesis Experiments on Stop Consonants in Initial Position.* - M.I.T. QPR 61, p. 153.

[25] FUJIMURA (O.), 1970, *Remarks on Stop Consonants - Synthesis Experiments and Acoustic Cues.* - Ann. Bull. Res. Inst. Logoped. Phoniat. 4, Univ. Tokyo, pp. 75-88.

[26] ABRAMSON (A.S.) & LISKER (L.), 1965, *Voice Onset Time in Stop Consonants : Acoustic Analysis and Synthesis.* - 5e Congrès International d'Acoustique, Liège A 51.

[27] LISKER (L.) & ABRAMSON (A.S.), 1967b, *Some Effects of Contexts on Voice Onset Time in English Stops.* - Language and Speech 10, pp. 1-28.

[28] SLIS (I.H.) & COHEN (A.), 1969a, *On the Complex Regulating the Voiced-Voiceless Distinction. I.* - Language and Speech 12, pp. 80-102.

[29] SLIS (I.H.) & COHEN (A.), 1969b, *On the Complex Regulating the Voiced-Voiceless Distinction. II.* - Language and Speech 12, pp. 137-155.

[30] HAGGARD (M.), AMBLER (S.), & CALLOW (M.), 1970, *Pitch as a Voicing Cue.* - J.A.S.A. 47, 613-617.

[31] LARREUR (D.), & BOË (L.J.), 1973, *Etude de l'influence des variations de la fréquence laryngienne sur l'intelligibilité et la qualité des consonnes sonores générées par vocodeur.* - Bulletin de l'Institut de Phonétique de Grenoble 2, pp. 103-126.

[32] STEVENS (K.N.) & KLATT (D.H.), 1971, *The Role of Formant Transitions in the Voiced-Voiceless Distinction for Stops.* - QPR MIT 101, pp. 188-197.

[33] STEVENS (K.N.) & KLATT (D.H.), 1974, *The Role of Formant Transitions in the Voiced-Voiceless Distinction for Stops.* - J.A.S.A. 55, pp. 653-659.

[34] GOLD (B.), 1962, *Computer Program for Pitch Extraction.* - J.A.S.A. 34, pp. 916-921.

[35] GOLD (B.) & RABINER (L.), 1962, *Parallel Processing Techniques for Estimating Pitch Periods of Speech in this Time Domain.*- J.A.S.A. 46, 442-448.

[36] DAVID (E.E.) Jr., 1957, *Signal Theory in Speech Transmission.* - IRE Trans. on Circuit Theory, CT-3, pp. 232-244.

[37] BOULOGNE (M.), 1972, *Détecteur de mélodie à autocorrélation.* - Rapport D.E.A. Ecole Nationale Supérieure d'Electricité et Radioélectricité, Grenoble.

[38] MARKEL (J.D.), 1973, *Application of Digital Inverse Filter for Automatic Formant and F_o Analysis.* - IEEE Trans. Audio. Elect. AU 21, pp. 154-160.

[39] ROSS (M.J.), SHAFFER (H.L.), COHEN (A.), FREUDBERG (R.) & MANLEY (H.J.), 1974, *Average Magnitude Difference Function Pitch Extractor.* - IEEE Acoust. Speech and Signal Process. ASSP-22, pp. 353-362.

[40] SERIGNAT (J.F.), 1974, *Contribution aux recherches sur la communication parlée. Travaux sur le vocoder à autocorrélation. Etude et simulation d'un vocoder à prédiction linéaire.*- Thèse de Docteur-Ingénieur - Grenoble.

[41] NOLL (A.M.), 1964, *Short Time "Cepstrum" Pitch Detection.* - J.A.S.A. 36, 1030 (A).

[42] OPPENHEIM (A.V.), 1969, *Speech Analysis-Synthesis System Based on Homomorphic Filtering.*- J.A.S.A. 45, pp. 459-462.

[43] ATAL (B.S.) & HANAUER (S.L.), 1971, *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave.*-J.A.S.A. 50, pp. 637-655.

[44] MAKHOUL (J.), 1973, *Spectral Analysis of Speech by Linear Prediction.*- IEEE Trans. Audio. Elect. AU 21, pp. 140-148.

[45] MARKEL (J.D.), 1972, *The SIFT Algorithm for Fundamental Frequency Estimation.*- IEEE Trans. Audio. Elect. AU 20, pp. 367-377.

[46] SUGIMOTO (T.), 1962, *On the Pitch Sensation and the Fundamental Frequency Excitation of Speech Sound.*- Elect. Com. Lab. Techn. J. 2.

[47] MAÏSSIS (A.), 1973, *Le Traitement de l'information acoustique, étape fondamentale de la reconnaissance de la parole.*- Thèse d'Etat, Paris VI.

[48] DOURS (D.) & FACCA (R.), 1974, *Méthode de segmentation et d'analyse par traitement direct du signal vocal. Application à la classification et la reconnaissance des voyelles et des consonnes.*- Thèses 3e Cycle et Docteur-Ingénieur, Univer. P. Sabatier, Toulouse.

[49] 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.

[50] DUDLEY (H.), 1939, *The Vocoder.*- Bell Labs. Record 17, pp. 122-126.

[51] WIREN (J.), & STUBBS (H.), 1956, *Electronic Binary Selection System for Phoneme Classification.*- J.A.S.A. 28, pp. 1082-1091.

[52] GRUENZ (O.O.) & SCHOTT (L.O.), 1949, *Extraction and Portrayal of Pitch of Speech Sounds.*- J.A.S.A. 21, pp. 487-495.

[53] MUNSON (W.A.) & MONTGOMERY (H.C.), 1950, *A Speech Analyzer and Synthetizer.*- J.A.S.A. 22, p. 678 (A).

[54] WEISS (M.R.) & HARRIS (C.M.), 1963, *Computer Technique for High-Speed Extraction of Speech Parameters.*- J.A.S.A. 35, pp. 207-214.

[55] CHRISTIANSEN (H.M.) & SCWEIZER (L.), SETHY (A.) & HOFFENREICH (F.), 1966, *New Correlation Vocoder.*- J.A.S.A. 40, pp. 614-620.

[56] CHARRAS (J.P.), *Le Vocodeur à canaux.*- Rapport D.E.A., Ecole Nationale Supérieure d'Electricité et Radioélectricité, Grenoble.

[57] SAKAI (T.), 1961, *The Phonetic Typewriter : its Fundamentals and Mechanism.*- Studia Phonologica 1, pp. 140-152.

[58] HIERONYMUS (J.L.), 1974, *Pitch Syncrhonous Segmentation.*- IEEE Symp. Speech Recognition T. 10, pp. 131-133.

[59] COMER (D.J.), 1968, *The Use of Waveform Asymetry to Identify Voiced Sounds.*- IEEE Trans. Audio. Elect. AU 16, pp. 500-506.

[60] PIROGOV (A.A.), 1974, *Vokodernaja Telefonija. Metody i Problemy.*- Ed. Svgaz' Moscou.

[61] LILJENCRANTS (J.), 1962, *A Few Experiments of Voiced-Voiceless Identification and Time Segmentation of Speech.*- SCS Stockolm, Paper C 8.

[62] 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. 104-113.

[63] BOË (L.J.) & RAKOTOFIRINGA (H.), 1972 a, *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 Grenoble 1, pp.1-9.

[64] BOË (L.J.) & RAKOTOFIRINGA (H.), 1972 b, *A Statistical Analysis : Laryngeal Frequency, its Relationship to Intensity Level; Duration.*- To be published in Language and Speech.

[65] ZURCHER (J.), 1972/73 , *Dispositif de détection de mesure du fondamental de la parole humaine.*- Analyse et synthèse de la Parole 1, pp. 7-15, Lannion.

[66] MILLER (G.A.) & NICELY (P.E.), 1955, *An Analysis of Perceptual Confusions Among Some English Consonants.*- J.A.S.A. 27, pp. 338-352. Trad. franç. in MEHLER (J.) & NOIZET (G.), 1974, Textes pour une psycholinguistique.- Mouton, Paris-La Haye, pp. 175-207.

[67] VOIERS (W.D.), 1968, *The Present State of Digital Vocoder Technique : a Diagnostic Evaluation.*- IEEE Trans. Audio. and Elect. AU 16. pp. 275-279.

[68] SMITH (P.S.), 1969 , *Perception of Vocoder Speech Processed by Pattern Matching.*- J.A.S.A. 46, pp. 1562-1571.

[69] PECKELS (J.P.) & ROSSI (M.), 1971, *Le Test de diagnostic par paires minimales. Adaptation au français du* Diagnostic Rhyme Test de W.D. VOIERS.- C.R. 2° Journées d'Etudes sur la Parole, Aix, Bd, Be, Bf, Bg, Bh. Groupe Communication Parlée, GALF.

[70] LANCE (G.N.) & WILLIAMS (W.T.), 1966/67, *A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems.*- The Computer Journal 9, pp. 373-380.

[71] JOHNSON (S.C.), 1967, *Hierarchical Clustering Schemes.*- Psychometrika 32, 241-254.

[72] STREVENS (P.), 1960, *Spectra of Fricative Noise in Human Speech.*- Language and Speech 3, pp. 32-49.

[73] VOIERS (W.D.), 1967, *Performance Evaluation of Speech Processing Devices III : Diagnostic Evaluation of Speech Intelligibility.*- AF Cambridge Research Labs., Final Rept.

[74] Mac KINNEY (N.P.), 1965, *Laryngeal Frequency Analysis for Linguistic.*- Research Communic. Sci. Lab. Univ. Michigan.

[75] JAKOBSON (R.), FANT (C.G.M.) & HALLE (M.), 1952, *Preliminaries to Speech Analysis.*- The M.I.T. Press, Cambridge.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

"Traitement indépendant ou interaction dans le processus d'intégration perceptive des indices de voisement ?".

W. SERNICLAES

RESUME

La multiplicité des indices qui interviennent dans la perception d'un trait phonétique pose le problème des rapports entre attributs du signal acoustique et catégorisation perceptive. Les résultats d'une expérience d'identification d'occlusives intervocaliques apportent des arguments en faveur de l'hypothèse d'un traitement indépendant de 2 indices de voisement - la durée de la voyelle précédant la consonne et le délai d'établissement du voisement de la voyelle finale (VOT).

SUMMARY

Multiple cues in voicing perception give rise to the problem of the relationship between acoustic features and perceptual categorization. Results of an identification experiment support an hypothesis concerning independent treatment of 2 voicing cues - preceding vowel duration and Voice Onset Time.

"Traitement indépendant ou interaction dans le processus d'intégration perceptive des indices de voisement?"
W. SERNICLAES

Introduction .

La plupart des travaux récents sur la perception des traits consonantiques ont pour objectif de préciser la fonction des "détecteurs" phonétiques (EIMAS et CORBIT, 1973). Le problème posé est celui du degré de spécialisation de ces mécanismes d'analyse. Si certains auteurs (par ex. COOPER, 1974) ont conclu à l'intervention d'un "analyseur de trait" qui effectuerait un traitement global de l'ensemble des indices acoustiques, d'autres (par ex. BAILEY, 1974) ont été amenés à supposer l'existence de plusieurs détecteurs d'indices pour un même trait.

Il nous semble que les résultats obtenus dans ces différents travaux ne sont pas nécessairement contradictoires étant donné l'ambiguïté de la notion d'"indice". De manière générale, on peut considérer qu'un indice est un élément arbitraire du signal acoustique dont l'efficacité perceptive a été démontrée. En tant que tels, les indices isolés n'ont pas de signification perceptive, l'information phonétique qu'ils transportent ne se manifestant que lorsqu'ils sont combinés avec d'autres indices (WAJSKOP et SWEERTS, 1973).

Ces constatations pourraient nous amener à conclure que l'identification d'un trait phonétique - de voisement par exemple - dépend d'un schème perceptif global qui ferait intervenir l'ensemble des indices pertinents (durée de la voyelle antéconsonantique, durée de la consonne, présence de vibrations laryngées, etc...). A l'encontre de cette hypothèse, certains résultats montrent que les effets perceptifs de différents indices ne sont pas nécessairement corrélés. A l'aide d'un test direct au niveau perceptif, nous avons montré (BEECKMANS et SERNICLAES, 1975) que les deux "indices" constitutifs de la dimension V.O.T.* (LISKER et ABRAMSON, 1964) ont des effets indépendants sur l'identification du trait de voisement des occlusives en séquence CV. Si un indice isolé n'apporte pas d'information autonome pour l'identification du trait, les effets perceptifs des indices ne dépendent donc pas nécessairement de l'ensemble des autres indices.

Le mode de structuration des indices établi par des expériences d'identification pourrait servir de point de départ pour préciser la notion de détecteur phonétique, pour déterminer le niveau de traitement où ils opèrent et spécifier leur fonction par rapport aux paramètres du signal. En effet, si l'on peut imaginer que deux indices indépendants sont analysés par des mécanismes séparés, il est peu vraisemblable que des indices corrélés, qui n'ont pas de signification perceptive autonome, soient traités par deux détecteurs différents.

Le travail que nous présentons a pour objectif de définir le mode de structuration perceptive de 2 indices de voisement de la consonne intervocalique : la durée de la voyelle initiale et le délai d'établissement du voisement de la voyelle finale (V.O.T. "positif"). Notre hypothèse est que ces deux

* la durée du segment de friction ou d'aspiration compris entre l'explosion consonantique et le début du voisement de la voyelle suivante constitue la branche positive de la dimension VOT. Le VOT est dit négatif lorsque des vibrations laryngées précèdent la détente consonantique.

indices sont analysés de manière indépendante car ils gardent une signification perceptive lorsqu'ils se présentent séparément en séquence VC -pour la durée de la voyelle préconsonantique-, ou en séquence CV- pour le délai d'établissement du voisement. Une autre interprétation voudrait que le mode de traitement perceptif de ces indices dépende de la séquence considérée. La restructuration du processus d'analyse pourrait se traduire par un traitement global des deux indices lorsqu'ils s'insèrent dans un cadre intervocalique.

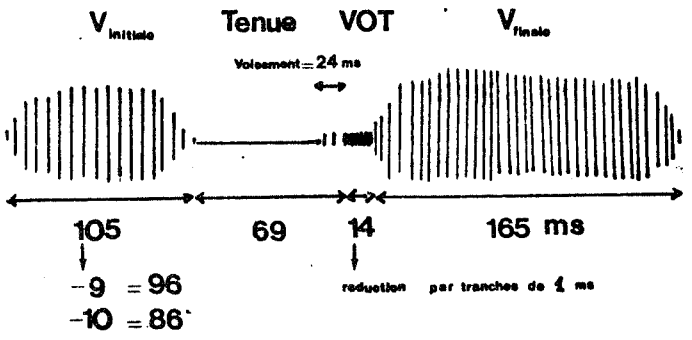
Procédure Expérimentale.

Trente séquences ambiguës entre /utu/ et /udu/ ont été constituées en manipulant la durée de la voyelle initiale et le délai d'établissement du voisement. Les valeurs pour chaque paramètre ont été déterminées à partir des données d'identification recueillies lors d'une expérience exploratoire.

Stimuli.

Les stimuli VCV ont été construits à l'aide d'un programme ordinateur "-ENFI"- réalisé par P. JOSPA. "ENFI" permet d'assembler des portions variables de signaux préalablement digitalisés et mémorisés sur les bandes LINC d'un PDP-12. La partie CV des stimuli se compose d'une séquence /tu/ prononcée par un locuteur francophone que l'on a fait précéder de 2 périodes de "pré-voisement" extraites d'une séquence /du/ prononcée par le même locuteur. Les résultats d'une expérience antérieure (BEECKMANS et SERNICLAES, 1975) montrent que pour le stimulus ainsi constitué, la modification du délai d'établissement du voisement (VOT "positif") a un effet optimal sur les réponses d'identification des sujets. La voyelle initiale a été sélectionnée dans un ensemble de séquences /utu/ prononcées par le même locuteur. Une voyelle /u/ extraite d'une séquence /utu/, et dont la hauteur du fondamental et l'intensité sont très proches des valeurs correspondantes de la séquence /tu/ utilisées, a été choisie. Les caractéristiques temporelles de la séquence obtenue par l'adjonction des segments /u/ et /tu/ sont présentées dans la Figure I.

Figure I.



A partir de cette configuration de base on a dérivé 2 séquences supplémentaires en excisant successivement 2 périodes dans la partie stable de la voyelle initiale. Trois valeurs de durées ont ainsi été obtenues pour cette voyelle : 105, 96 et 86 ms, qui correspondent respectivement aux valeurs de rapport temporel tenue/voyelle initiale : .65, .72 et .80 (voir figure 1). Pour chacune de ces 3 séquences 10 valeurs de délai d'établissement du voisement ont été choisies en réduisant la durée du segment d'aspiration de 14 à 5 ms par tranches de 1 ms (voir figure 1). Les 30 stimuli (3 durées de voyelle initiale x 10 VOT) ont servi à la constitution de 3 séries expérimentales dans lesquelles chacun apparaît 42 fois au total.

Sujets.

Trois locuteurs francophones ont passé l'expérience en 3 séances réparties dans l'intervalle d'une semaine. Leur tâche consistait à identifier le stimulus en choisissant l'une des deux réponses /utu/ ou /udu/.

3. Résultats et discussion.

Aspects généraux.

L'examen des résultats a été centré sur l'analyse des effets simultanés de la durée de la voyelle initiale et du VOT. Au préalable, certains aspects de l'évolution du taux de réponses sourdes (/utu/) doivent cependant être relevés.

Pour l'ensemble des sujets, on remarque que l'augmentation du taux de réponses sourdes en fonction de l'allongement du VOT est extrêmement rapide (voir figure 2). Dans 2 cas (sujets JVG et MS) le degré de précision est proche de celui qui a été obtenu récemment pour des occlusives initiales en anglais (DRAPER et HAGGARD, 1974). C'est essentiellement dans la zone comprise entre 8 et 12 ms environ que le VOT apparaît comme extrêmement puissant; de part et d'autre de cette zone l'évolution des réponses est plus lente et plus irrégulière. Cette zone correspondrait à un maximum de sensibilité dans l'analyse de cet indice.

En effet, l'étude du même indice dans d'autres contextes (/tu/) a abouti à ce même intervalle de plus grande précision, ce qui montre bien que c'est le traitement de l'indice qui est en cause et non son intégration avec d'autres indices. Enfin cette zone privilégiée est constante à travers les sujets et les contextes et elle est compatible avec les données de production, la frontière acoustique du VOT * en français se situant vers 10 ms (BEECKMANS et SERNICLAES, 1975).

Analyse des effets cumulés des 2 indices.

Le traitement des résultats a été circonscrit au stimuli pour lesquels on a observé une progression régulière des taux d'identification, c'est à dire ceux correspondant à l'intervalle de VOT compris entre 8 et 12 ms. L'analyse statistique a consisté en l'ajustement de courbes cumulatives normales à ces données et ce en fonction de 2 modèles différents. Dans un premier temps chacun des 9 groupes (3 sujets x 3 durées de voyelle initiale) de 5 points retenus pour l'analyse (entre 8 et 12 ms) ont fait l'objet d'un ajustement séparé à l'aide de la méthode des probits (FINNEY, 1952; p. 22).

* il s'agit bien entendu de mesures du VOT "positif", indépendantes de la durée du pré-voisement.

Figure II.2

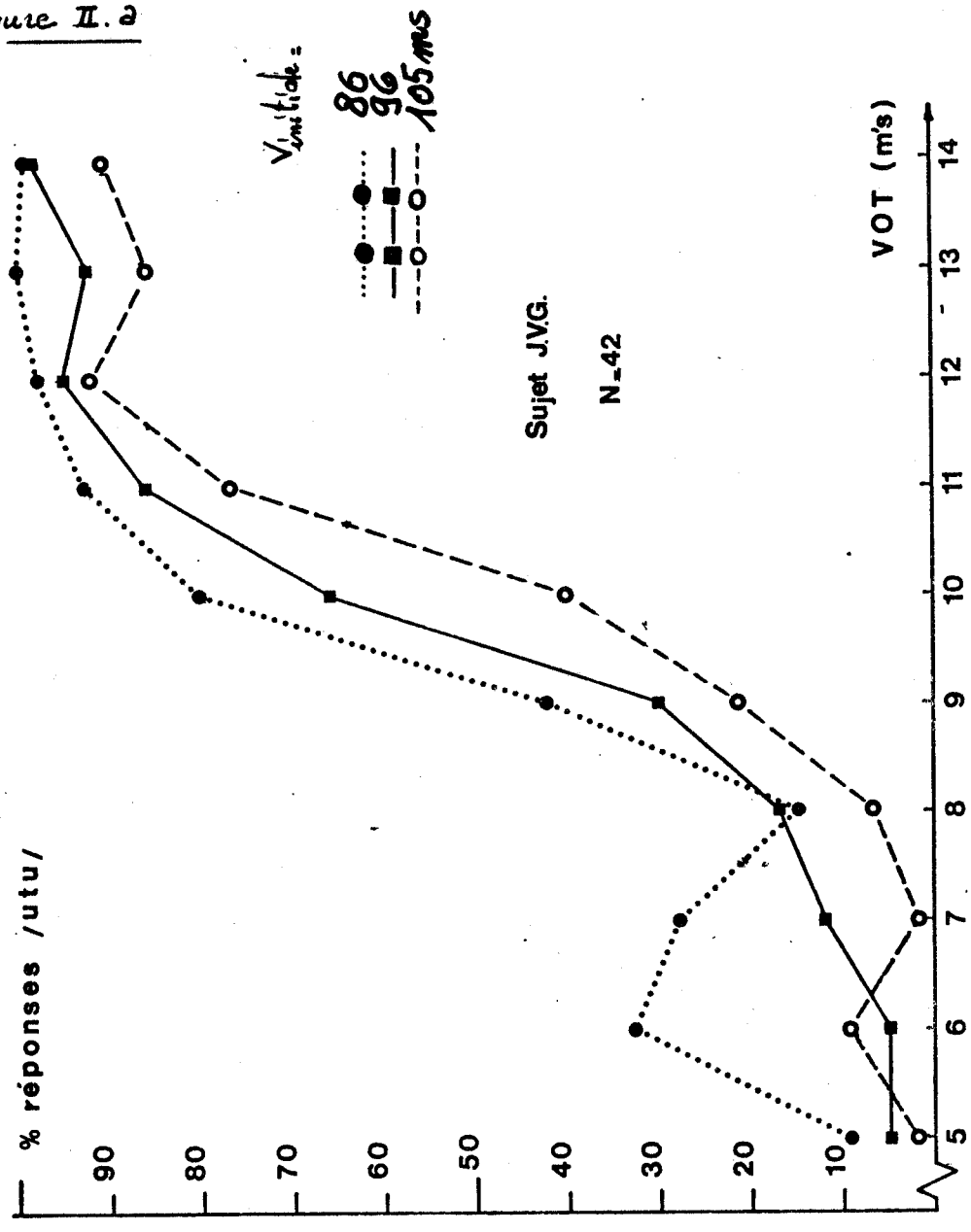

$V_{initiale}$ =
86
96
105 ms
Sujet J.V.G.
N=42
% réponses /utu/
VOT (m's)
90
80
70
60
50
40
30
20
10
5
6
7
8
9
10
11
12
13
14

Figure II.b

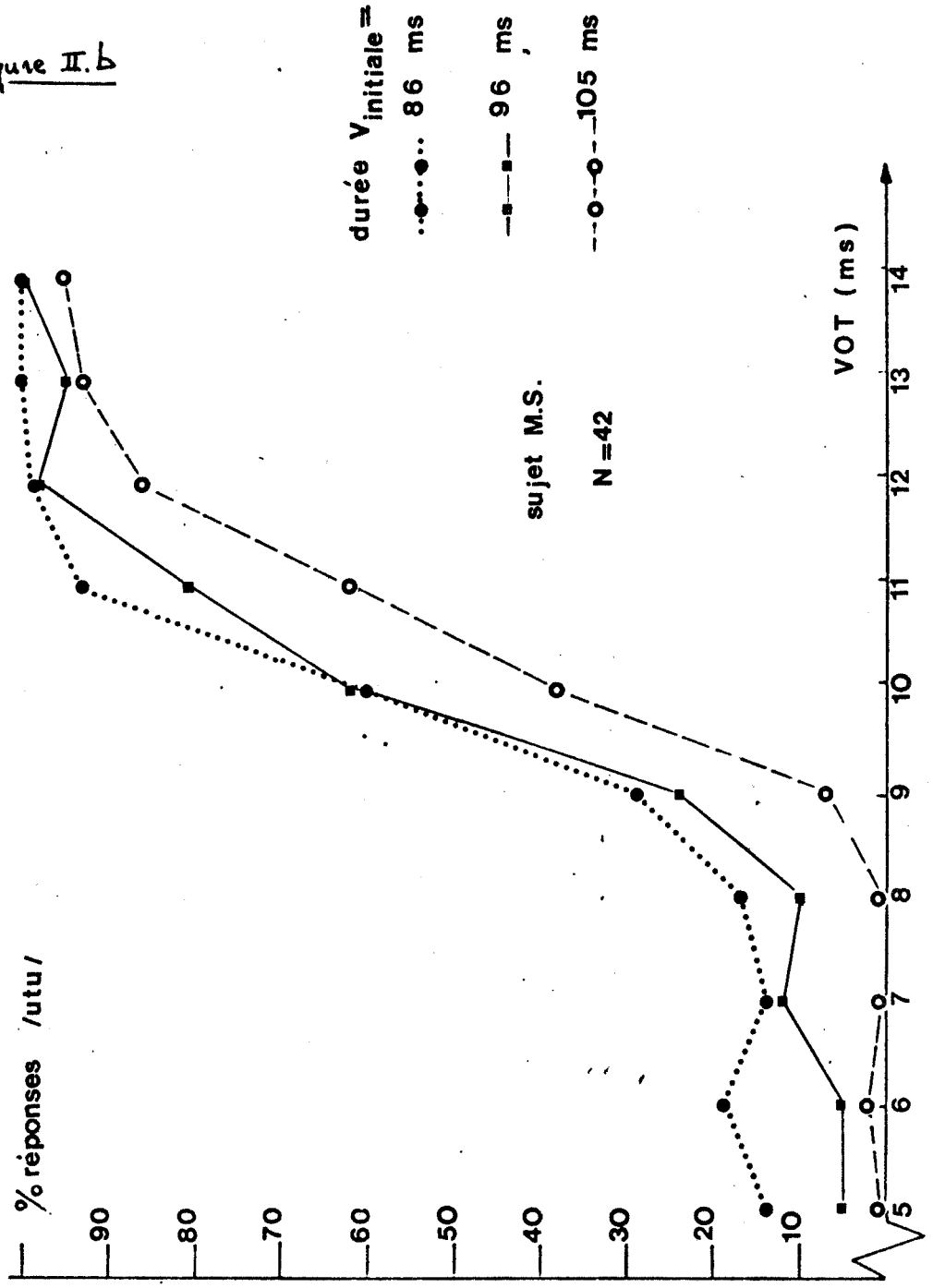
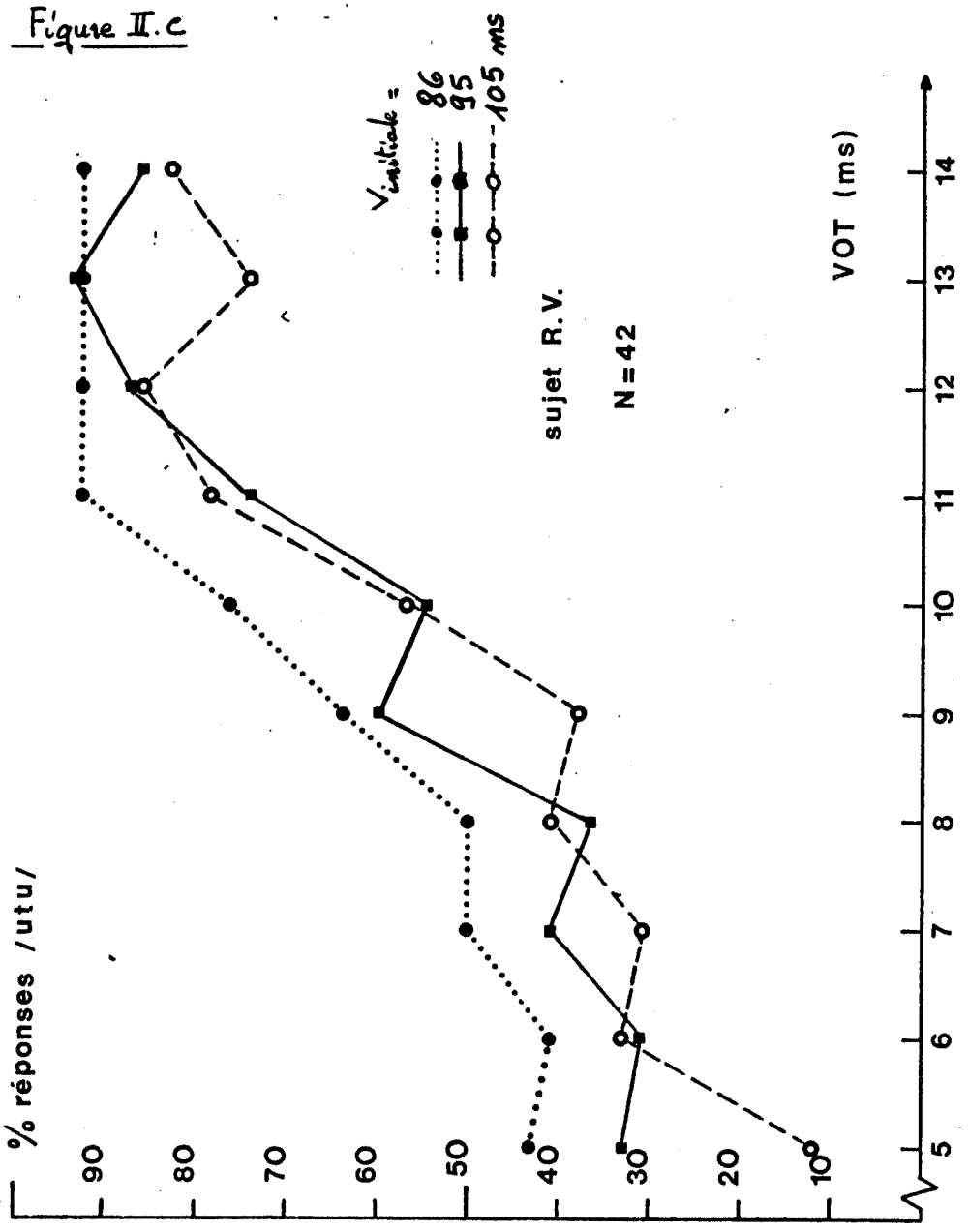

durée $V_{initiale}$ =
86 ms
96 ms
105 ms
sujet M.S.
N =42
% réponses /utu/
VOT (ms)
90
80
70
60
50
40
30
20
10
5
6
7
8
9
10
11
12
13
14

Figure II.c


% réponses /utu/
90
80
70
60
50
40
30
20
10
5
6
7
8
9
10
11
12
13
14
VOT (ms)
sujet R.V.
N=42
Vinitiale =
86
95
105 ms

Le modèle se présente sous la forme :

(1) probit $p = 5 + \frac{1}{\sigma}(x - \mu)$

où x est le V.O.T. et l'écart-type et μ la moyenne de la courbe cumulative. Pour chaque sujet et chaque durée de voyelle initiale des estimations séparées de σ et μ ont été calculées.

Ensuite, l'ensemble des 15 données expérimentales retenues pour chaque sujet a fait l'objet d'un ajustement global par des ogives parallèles (FINNEY, 1952; p. 104) en utilisant le modèle additif :

(2) probit $p = \alpha + \frac{1}{\sigma_1}(x_1 - \mu_1) + \frac{1}{\sigma_2}(x_2 - \mu_2)$

où x_1 est le V.O.T. et x_2 la durée de la voyelle initiale.
Dans ce modèle l'égalité des pentes des fonctions qui seront ajustées aux durées intervient en tant que contrainte initiale. L'écart dans la qualité des ajustements obtenus pour chacun des modèles nous permettra de tester l'additivité des effets introduits par les 2 indices, et ce pour chacun des 3 sujets.

Les estimations des écart-types obtenus en procédant à des ajustements séparés pour les différentes durées de la voyelle initiale (modèle (1)) sont présentés dans le tableau 1. Bien que la procédure d'ajustement utilisée n'impose pas de contraintes, on remarquera que pour chacun des sujets les ogives sont sensiblement parallèles.

Tableau 1. Ecart-types.

		Sujets		
		J.V.G.	M.S.	R.V.
Durée $V_{initiale}$ (ms)	105	1.14 ± 0.16	1.20 ± 0.13	2.71 ± 0.49
	96	1.46 ± 0.19	1.27 ± 0.15	3.16 ± 0.69
	86	1.23 ± 0.14	1.25 ± 0.15	2.71 ± 0.53

La valeur de l'ajustement global réalisé par le modèle additif (modèle (2)) ainsi que la pertinence de l'hypothèse de parallélisme qui le sous-tend peuvent être appréciées à l'aide de tests χ^2. Une valeur de χ^2 significative entraîne le rejet du modèle ou l'inadéquation de l'une des hypothèses sur lesquelles il est fondé. La valeur du χ^2 employé pour tester le parallélisme correspond à l'accroissement d'erreur qui résulte de l'utilisation du modèle additif (plus contraignant).

Les valeurs de χ^2 obtenues pour chacun des 3 sujets (voir Tableau 2) sont toutes largement inférieures aux seuils de signification usuellement choisis ($P < .05$). Les coefficients d'hétérogénéité obtenus pour les tests de parallélisme (χ^2 moyens- voir Tableau 2) qui sont soit inférieurs, soit proches de l'unité, indiquent que les valeurs calculées du χ^2 sont proches des valeurs attendues. Le modèle et en particulier l'hypothèse d'additivité peuvent donc être raisonnablement acceptés.

Tableau 2.

Sujet JVG

	dl	χ^2	χ^2_{moyen}
Parallélisme des régressions	3	1.53	0.51
Hétérogénéité résiduelle	9	3.35	0.37
Total	12	4.88	0.41

Sujet MS

	dl	χ^2	χ^2_{moyen}
Parallélisme des régressions	3	2.23	0.74
Hétérogénéité résiduelle	5	6.05	1.01
Total	8	8.27	1.03

Sujet RV

	dl	χ^2	χ^2_{moyen}
Parallélisme des régressions	3	2.02	0.67
Hétérogénéité résiduelle	9	6.05	0.67
Total	12	8.07	0.67

Si la réduction de la durée de la voyelle se traduit par un déplacement significatif du point d'équiprobabilité vers des VOT plus longs, par contre elle n'apporte pas de modification sensible et systématique de la pente des fonctions d'identification.

En résumé, on peut supposer l'intervention, à un premier niveau, d'une analyse autonome du V.O.T. qui montre un maximum de précision de part et d'autre de la frontière acoustique de cet indice, et, à un second niveau, une intégration additive des informations apportées par cet indice et par la durée de la première voyelle. Les différences individuelles se manifesteraient par les pondérations relatives accordées à chaque indice : pour les sujets M.S. et J.V.G.,on remarque que la pente des courbes d'identifications est beaucoup plus raide que celles relatives au sujet R.V.. Les deux premiers sujets accorderaient beaucoup plus d'importance à l'indice V.O.T. On peut donc rendre compte du processus de structuration des indices étudiés, en supposant l'existence de facteurs de pondérations d'indices propres à chaque sujet, mais invariants à travers les différents contextes phonétiques (SERNICLAES, 1974).

4. Conclusions.

L'hypothèse d'un traitement perceptif indépendant de deux indices de voisement de l'occlusive intervocalique du français
- la durée de la voyelle initiale et le délai d'établissement du voisement (V.O.T. "positif") de la voyelle finale - peut être soutenue en fonction de deux éléments d'appréciation différents. D'une part l'additivité des effets de ces indices sur les réponses d'identification peut raisonnablement être acceptée. D'autre part, la zone correspondant au maximum de sensibilité dans l'analyse du V.O.T. est identique à celle qui a été obtenue pour l'occlusive en position initiale.

Bibliographie.

- BAILEY, P.J. (1974) "Perceptual adaptation for acoustical features in speech". Speech Perception 2 , 29-34 (Department of Psychology, The Queen's University of Belfast).
- BEECKMANS, R. & SERNICLAES, W. (1975) "Positive and negative V.O.T. two independent cues in voicing perception". Proceedings of the 8th Int. Cong. of Phonetic Sciences (à paraître).
- COOPER, W.E. (1974) "Selective adaptation for acoustic cues of voicing in initial stops". Journal of Phonetics 2, 303-313
- DRAPER, G. & HAGGARD, M. (1974) "Facts and Artefacts in feature independence". Preprints of the SCS, Speech Transmission Laboratory 67-75, Stockholm 1974.
- EIMAS, P.D. & CORBIT, J.D. (1973) " Selective adaptation of Linguistic feature detectors" Cog. Psychol. 4, 99-109.
- FINNEY, D.J. (1952) Probit Analysis. University Press, Cambridge 1952.
- LISKER, L. & ABRAMSON, A.S. (1964) " A cross language study of voicing in initial stops" Word 20, 384-422.
- SERNICLAES, W. (1974) "Perceptual processing of acoustic correlates of the voicing feature". Preprints of the SCS, Speech Transmission Laboratory, 87-94, Stockholm 1974.
- WAJSKOP, M. & SWEERTS, J. (1973) "Voicing cues in oral stop consonants" Journal of Phonetics 1, 121-130.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

Perception des sons filtrés.

A. Landercy - R. Renard.

Université de Mons.

RESUME

Des expériences de reconnaissance des voyelles naturelles filtrées montrent qu'un champ fréquentiel très étroit permet une identification égale, voire supérieure à celle obtenue en C.D. Une seule zone formantique suffit à caractériser [i] [u] [a] [œ], les autres phonèmes nécessitent un filtrage plus complexe. L'examen des confusions permet de tirer des enseignements pour la reconnaissance automatique et pour la phonétique corrective.

SUMMARY

Experiments on filtered natural vowel recognition show that a very short frequency band allows an equal or superior identification than the direct channel. Only one formant band is sufficient to characterize [i] [u] [a] [œ], the other phonemes need a more complex filtering. The study of confusions enables us to form rules for the automatic recognition and for the corrective phonetics.

Perception des sons filtrés.

A. Landercy - R. Renard.
Université de Mons.

1. Introduction.

Afin d'essayer de déterminer le champ fréquentiel nécessaire et suffisant à la reconnaissance des phonèmes, nous avons réalisé une série d'expériences de perception et de reconnaissance de voyelles naturelles filtrées. Notre intention était de vérifier s'il existe une ou plusieurs bandes fréquentielles caractéristiques d'un phonème, c'est-à-dire suffisantes pour que le score de reconnaissance de ce phonème filtré dans cette ou ces bandes fût égal ou supérieur à celui obtenu en canal direct (C.D.).

Les résultats obtenus montrent qu'il est possible de déterminer des zones fréquentielles minimales extrêmement étroites dans lesquelles les phonèmes sont aussi bien et mieux identifiés qu'en C.D. (Landercy - Renard, 1974; idem, 1975). Nous les présentons très brièvement et discuterons essentiellement dans cet article des tendances générales des erreurs lorsque le filtrage élimine la zone fréquentielle correspondant au premier formant ainsi que des possibilités d'exploitation des résultats obtenus en ce qui concerne la phonétique corrective ou la reconnaissance automatique des phonèmes.

2. Schéma expérimental.

Nous avons centré cette première recherche sur les voyelles orales françaises. Quatre locuteurs francophones, deux hommes et deux femmes ont fourni le matériel vocal de cette expérience. Les voyelles, prononcées dans des mots clés, ont été extraites de leur contexte, amputées des transitions formantiques et normalisées en durée à 80ms à l'aide de segmentateurs électroniques et de leur organe de commande (Landercy, Sylin, Wajskop, 1969). Les fronts de montée et de descente en forme d'arc exponentiel de constante de temps 20ms reproduisent le mieux possible l'attaque vocale naturelle lors de l'émission d'une voyelle isolée (Landercy, 1971).

Pour chaque locuteur, nous avons réalisé quatre enregistrements magnétiques comportant chacun, en ordre aléatoire, les voyelles naturelles et filtrées. Les tests se présentent de la manière suivante : un chiffre annonce le numéro du stimulus; après un blanc de deux secondes, apparaissent deux stimuli identiques séparés entre eux d'une seconde; un temps mort de trois secondes permet alors au sujet de consigner sa réponse; se présente ensuite le chiffre qui annonce le stimulus suivant.

Chaque test a été présenté au moyen de casques TDH 39 à 30 étudiants francophones de l'Université de Mons à qui il était demandé d'identifier les voyelles. Les tests étaient à choix forcé et les sujets rémunérés. Le choix des

filtres pour chaque test s'est effectué comme suit :

Test o. Chaque voyelle a été filtrée, à l'aide de l'appareil Suvag Lingua (Renard, 1967) dans divers filtres d'une octave de largeur en pente raide (32dB/octave). Le choix des octaves à employer pour les diverses voyelles a été basé sur les "optimales" des voyelles françaises déterminées à Zagreb (Guberina et coll., 1965). Chaque voyelle a été filtrée dans cette octave optimale, dans les octaves précédentes et suivantes et dans les intermédiaires. Il s'agit donc d'un filtrage en bande continue et chaque voyelle était présentée six fois.

Dans les tests 1 et 2, chaque voyelle a été filtrée en bande discontinue à l'aide de deux filtres Krohn Hite montés en parallèle avec un gain identique : le modèle 3202 dont la pente est de 24dB/octave et le modèle 3342 dont la pente est de 48dB/octave. Le choix des fréquences de coupure de ces filtres, utilisés en passe-bas, passe-haut ou passe-bande) a été déterminé par les zones fréquentielles caractéristiques des formants des voyelles françaises.

Test 1. Les voyelles étaient filtrées dans des bandes fréquentielles étroites correspondant au premier (24dB/octave) et au second formants (48dB/octave) ($F_1 + F_2$), d'une part, et au premier (48dB/octave) et aux second et troisième (24dB/octave) formants, d'autre part [$F_1 + (F_2 + F_3)$]. Les voyelles étaient donc présentées trois fois.

Test 2. Les voyelles étaient filtrées trois fois et donc présentées quatre fois :

- en canal direct;
- en passe-bas avec une fréquence de coupure supérieure à celle du premier formant et en passe-haut avec une fréquence de coupure juste inférieure à celle du deuxième formant;
- en passe-bas avec une fréquence de coupure supérieure à celle du premier formant et en passe-haut avec une fréquence de coupure juste inférieure à celle du troisième formant;
- en passe-bas avec une fréquence de coupure supérieure à celle du premier formant et en passe-haut avec une fréquence de coupure juste inférieure à celle du quatrième formant.

Excepté pour le canal direct, la bande passante vers les basses fréquences avait été limitée à 250Hz afin d'atténuer l'intensité de la fondamentale de la voix.

Dans le troisième test enfin, toujours avec les mêmes filtres, nous avons supprimé le premier formant, et chaque voyelle a été présentée trois fois :

- en canal direct;
- filtrée dans une bande fréquentielle continue correspondant aux second et troisième formants;
- filtrée en passe-haut avec une fréquence de coupure juste inférieure au second formant.

3. Présentation des principaux résultats (1).

Lorsqu'elles sont perçues à travers un filtrage continu (tests 0 et 3) seules les voyelles [i] [a] [u] et l'archiphonème [Œ] obtiennent un score de réponses correctes, égal ou supérieur à celui obtenu en C.D.

Pour [i], dès que les zones fréquentielles supérieures à 2000Hz sont présentes, le score de reconnaissance dépasse 90% et égale le pourcentage obtenu en C.D.

Pour [u], dès que les fréquences graves sont présentes (entre 200 et 600Hz), le score de reconnaissance égale celui obtenu en C.D.

Pour [a], les filtres d'octave 800-1600Hz et 1200-2400Hz présentent un score de reconnaissance supérieur à celui obtenu en C.D.

Pour [œ] et [ø], un filtre passe-bande 1500-2500Hz ou un filtre passe-haut à partir de 1500Hz présente un score de reconnaissance [œ/ø] supérieur à 95%.

Pour toutes les autres voyelles, un filtrage continu a toujours donné un score de réponses correctes nettement inférieur à celui obtenu en C.D.

Lorsqu'elles sont filtrées de manière discontinue (tests 1 et 2), le score de reconnaissance des autres voyelles atteint toujours, pour certains filtres, celui obtenu en C.D. Passons rapidement en revue les résultats voyelle par voyelle.

---

(1) Les résultats complets de ces expériences seront publiés dans les n° 32 et 33 de la "Revue de Phonétique Appliquée". Mlle S. Dubois, logopède, attachée au Service de Phonétique de l'Université de Mons, s'est chargée de la présentation des tests, de la collecte des résultats et de leur dépouillement.

Les voyelles [ø et œ] obtiennent un score de reconnaissance intrinsèque égal au C.D. lorsqu'elles sont filtrées dans deux bandes étroites correspondant à leurs premier et second formants respectifs (autour de 400 et 1600Hz pour le [ø] et 500 et 1500Hz pour le [œ]. Si l'on élargit la bande du second formant jusqu'à ce qu'elle inclue le 3ème (2500Hz), le score de reconnaissance augmente et est significativement supérieur à celui obtenu en canal direct.

La voyelle [y], lorsqu'elle est filtrée dans deux bandes étroites correspondant à ses premier et second formants (de 250 à 350Hz et de 1700 à 1900Hz) obtient un score de reconnaissance de 95%, significativement supérieur à celui obtenu en C.D. (79%). L'élargissement de la bande du second formant n'a pas d'influence sur le score de reconnaissance.

Les voyelles [e] et [ɛ] nécessitent un filtrage où les premier et troisième formants doivent être présents pour obtenir un score de reconnaissance qui n'est pas significativement différent de celui obtenu en C.D. La présence ou l'absence du second formant ne semble pas jouer de rôle pour la reconnaissance de ces voyelles.

L'archiphonème [*O*] est reconnu comme tel dès que deux bandes étroites donnent indication des valeurs des deux premiers formants; la différenciation [o]/[ɔ] est nettement améliorée lorsqu'on élargit la bande fréquentielle du second formant.

De manière générale, les résultats que nous avons obtenus confirment - si besoin en était - le caractère sélectif de notre perception phonologique. Tout se passe comme si la perception des phonèmes de la langue maternelle était diversément fondée selon qu'il s'agit de phonèmes "primaires" [ɪ] [a] [u] et l'archiphonème central [Œ] ou "secondaires" [y] [e] [ɛ] [o] [ɔ] [ø/œ]. Les premiers seraient perçus sur la base d'une seule de leurs zones spectrales. Les seconds nécessitent un filtrage plus précis, cernant mieux les zones fréquentielles et supposeraient une structuration plus complexifiée, située à un niveau plus hiérarchisé.

4. Etude des confusions.

L'examen des confusions obtenues pour certains filtres permet de tirer certains enseignements. De manière générale :
a) n'importe quelle voyelle filtrée dans les fréquences graves, inférieures à 400Hz, tend à être confondue avec [u]. Ainsi, dans l'octave 150-300Hz : (résultats sur 120)
[o] est identifié 7 fois et confondu 81 fois avec [u];
[y] est identifié 2 fois et confondu 99 fois avec [u].

Dans l'octave 200-400Hz : (résultats sur 120)
[o] est identifié 27 fois et confondu 32 fois avec [u];
[ɔ] est identifié 20 fois et confondu 25 fois avec [u];
[ø] est identifié 24 fois et confondu 48 fois avec [u];
[y] est identifié 10 fois et confondu 95 fois avec [u].

b) n'importe quelle voyelle filtrée dans les fréquences aiguës, supérieures à 2000Hz; tend à être confondue avec [i]. Ainsi, dans l'octave 2400-4800Hz : (résultats sur 120)
[e] est identifié 49 fois et confondu 49 fois avec [i];
[ɛ] est identifié 5 fois et confondu 41 fois avec [i].

Dans l'octave 3200-6400Hz : (résultats sur 120)
[e] est identifié 35 fois et confondu 76 fois avec [i];
[ɛ] est identifié 11 fois et confondu 58 fois avec [i].

Etudions à présent de manière qualitative, la tendance générale des erreurs lorsque le filtrage a éliminé le premier formant (fig. 1). Cette tendance est notée sous forme vectorielle.

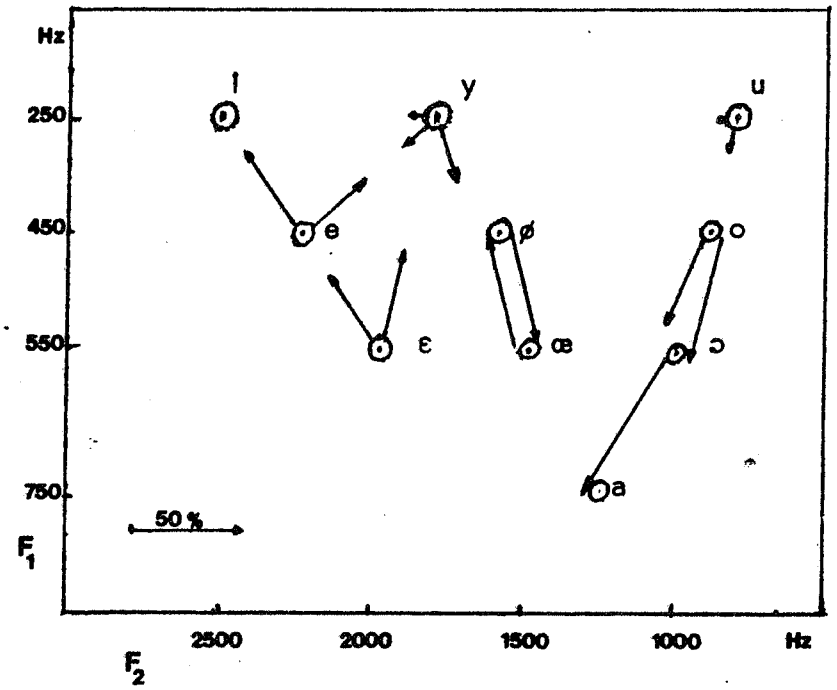


Fig. 1. Tendance générale des erreurs.

[i] et [a] sont toujours bien identifiés, de même que l'archiphonème [Œ], les réponses se répartissant également entre [ø] et [œ] quelque soit celui des deux qui est présenté. La différenciation [ø]/[œ] s'effectue donc essentiellement au niveau du premier formant.

Lorsque le filtrage ne présente que les fréquences aiguës, [y] est principalement confondu avec [œ], ou dans une moindre mesure avec [e] ou [i].

La nature de la confusion est déterminée par l'adéquation entre les fréquences présentées (fig. 2) et la structure formantique des voyelles de transfert.

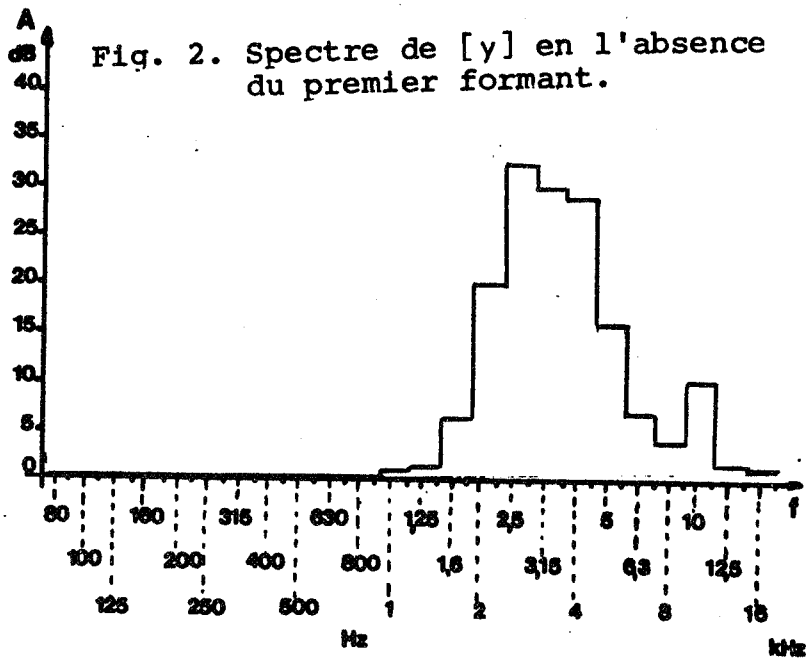


Fig. 2. Spectre de [y] en l'absence du premier formant.

[o] et [ɔ], par contre, ont nettement tendance à s'ouvrir vers [a]. La répartition spectrale de ces voyelles filtrées explique aisément ce phénomène. En effet, la densité importante des fréquences entre 800-2500Hz est typique de la répartition de [a] (fig. 3).

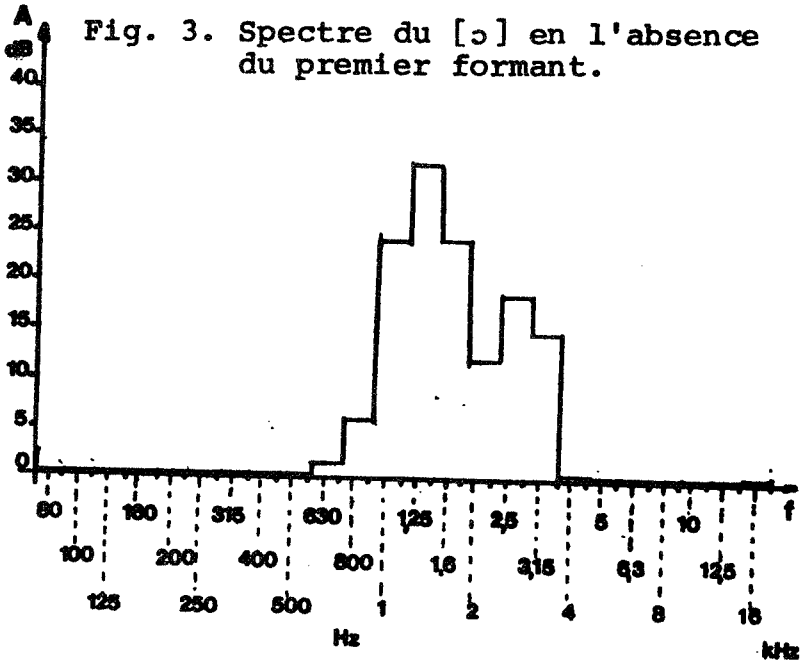


Fig. 3. Spectre du [ɔ] en l'absence du premier formant.

5. Applications.

Nous ne ferons que suggérer le genre d'applications que l'on peut tirer de ces diverses constatations soit en phonétique corrective, soit en reconnaissance automatique de la parole en prenant quelques exemples.

Si nous présentons [ɪ] filtré en passe-bas jusqu'à 2500Hz, il est perçu à 80% comme [y]. Si nous présentons [y] filtré en passe-haut à partir de 3500Hz (avec bande à 250Hz), il est perçu à 88% comme [ɪ]. Si nous présentons [ɪ] ou [y] filtré en passe-bas à 1000Hz, il est perçu comme [u]. En d'autres termes, dès qu'une voyelle présente, hormis la fondamentale, un maximum d'énergie en-dessous de 300Hz :

- s'il n'y a pas d'énergie au-dessus de 1000Hz, c'est un [u];
- s'il n'y a pas d'énergie en-dessous de 2000Hz, c'est un [ɪ];
- s'il y a de l'énergie entre 1500 et 2000Hz, c'est un [y].

D'autre part, si un son présente une concentration d'énergie importante entre 2000 et 3000Hz :

- s'il possède un maximum d'énergie en-dessous de 300Hz, c'est un [ɪ];
- s'il possède un maximum d'énergie au-dessus de 450Hz, c'est un [ɛ];
- s'il possède un maximum d'énergie entre 300 et 450Hz, c'est un [e].

Ces règles n'envisagent que la détection d'un maximum relatif et de la présence d'énergie dans une gamme assez large de fréquences et pourraient peut-être utilement être vérifiées en reconnaissance automatique de la parole.

Prenons à présent un exemple en phonétique corrective.

En italien de Sicile, le phonème [y] n'existe pas. Lorsque l'influence de la graphie n'intervient pas, l'erreur commise par le Sicilien tend soit vers [œ] (également inconnu dans sa langue), soit vers [e], soit vers [ɪ] (Intravaia, 1975). Il y a donc similitude complète entre les fautes à la production des Siciliens et celles à la perception des francophones lorsque l'on a éliminé le premier formant. Selon que l'erreur se dirige vers [œ] ou vers [ɪ], ce sera respectivement sur les fréquences graves ou sur les fréquences aiguës qu'il faudra attirer l'attention du sujet.

Nous pourrions multiplier ces exemples. Nous avons seulement voulu attirer l'attention sur l'importance relative des divers formants vocaliques au niveau perceptif. En fait, à ce niveau, nos expériences semblent montrer qu'une détermination précise des deux ou trois premiers formants ou de leur rapport entre eux n'est pas indispensable à la caractérisation des voyelles françaises. Celle-ci est obtenue à partir de deux plages fréquentielles, l'une étroite, et l'autre plus ou moins large selon la complexité du phonème vocalique.

x

x x

BIBLIOGRAPHIE.

GUBERINA, P., GOSPODNETIC, J., POZOJEVIC, M., SKARIC, I. et VULETIC, B., Correction de la prononciation des élèves qui apprennent le français, Revue de Phonétique Appliquée, 1, 1965, 81-94.

INTRAVAIA, P., Système de fautes des Siciliens qui apprennent le français, mémoire de licence, Université de Mons, 1975 (sous presse).

LANDERCY, A., Temporal segmentation - Influence of the envelope function on perception, Journal of Speech and Hearing Research, 14, 1971, 47-57.

LANDERCY, A., RENARD, R., Perception des voyelles françaises filtrées, Revue de Phonétique Appliquée, 32, 1974, 11-32.

LANDERCY, A., RENARD, R., Champ fréquentiel et reconnaissance de voyelles françaises, Revue de Phonétique Appliquée, 33, 1975 (sous presse).

LANDERCY, A., SYLIN, G., WAJSKOP, M., Etude et réalisation d'un segmentateur électronique et de son organe de commande, Revue d'Acoustique, 5, 1969, 31-36.

RENARD, R., L'appareil Suvaglingua, instrument de recherche et de correction phonétique, Revue de Phonétique Appliquée, 4, 1967, 59-67.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

PROSODIE ET INTELLIGIBILITE DE LA PAROLE

POUR DES AUDITEURS ETRANGERS

D. ROSTOLLAND et C. PARANT

Laboratoire de Physiologie du Travail du C.N.A.M. et du C.N.R.S. - Paris

RESUME

On a étudié l'effet de l'effort vocal sur l'intelligibilité de la parole, pour des auditeurs de quatre nationalités. Des listes de mots dissyllabiques, enregistrées en voix parlée et en voix criée, sont reproduites en présence de bruit. Certains éléments prosodiques de la voix criée (accentuation, durée des syllabes) modifient l'intelligibilité. Si la parole est perturbée à la fois dans sa production (facteurs prosodiques, voix criée), dans sa transmission (bruit) et dans sa réception (facteurs sémantiques, voix étrangère), on observe une perte d'intelligibilité qui peut atteindre 100 %. Cette situation se produit dans la réalité industrielle, pour des travailleurs étrangers.

SUMMARY

Prosody and Speech Intelligibility for Foreign Listeners

The effects of vocal effort on the speech intelligibility, for listeners of four nationalities, has been studied. The lists of disyllabic words, recorded with a spoken and shouted voice, are reproduced in the presence of noise. Intelligibility may be modified by certain prosodical elements of the shouted voice (accentuation, duration of the syllables). One can observe a loss of intelligibility which can reach 100 %, when speech is disturbed during production (prosodical factors, shouted voice), transmission (masking noise) and reception (semantic factors, foreign voice). For foreign workers, this situation is an industrial reality.

PROSODIE ET INTELLIGIBILITE DE LA PAROLE POUR DES AUDITEURS ETRANGERS

D. ROSTOLLAND et C. PARANT
(Laboratoire de Physiologie du Travail du C.N.A.M. et du C.N.R.S. Paris)

INTRODUCTION

Le thème de cette étude a pour origine la situation concrète des travailleurs étrangers dans l'industrie. Dans de nombreux secteurs de la vie industrielle, la présence d'un bruit de fond intense gêne la compréhension des messages verbaux, et cela est particulièrement sensible pour un travailleur étranger. C'est d'ailleurs un fait d'observation courante que la compréhension d'une langue étrangère devient subitement très difficile en présence de bruit. Tout se passe comme si l'auditeur devenait momentanément sourd, faute de pouvoir décoder la parole assez vite et perdant ainsi le "fil de la conversation".

Une étude préliminaire a montré que la présence d'un bruit, même peu intense, pouvait entraîner une diminution massive de l'intelligibilité d'une langue étrangère (ROSTOLLAND et PARANT, 1973). Nous reprenons ici, de façon systématique, cette étude étant donné l'importance des communications verbales dans les ateliers et les chantiers et dans l'industrie et les transports en général. L'intelligibilité des messages industriels a un effet direct sur la sécurité et l'efficacité du travail.

De nombreux travaux ont montré l'influence de divers types de bruit sur l'intelligibilité de la parole. Dans la plupart de ces travaux, il s'agit de voix parlée normale ou forte entre interlocuteurs de même langue maternelle. Plusieurs recherches ont été consacrées à l'influence de l'étendue du vocabulaire sur l'intelligibilité et quelques études concernant l'effet de l'effort vocal (PICKETT 1956 et 1958, PICKETT et al. 1958, WEBSTER et al. 1962, KRYTER 1962 et 1970, WEBSTER 1968, ROSTOLLAND et al. 1973). Certains aspects concrets, jusqu'ici sous-estimés, sont liés à la dégradation de l'information verbale, lorsque plusieurs perturbations interviennent simultanément dans la chaîne de communication parlée. En milieu bruyant l'efficacité de la communication se trouve compromise, même pour un travailleur français. Que reste-t-il alors d'intelligible pour un travailleur étranger ?

Si la distance entre les interlocuteurs augmente, le locuteur passe progressivement de la voix parlée à la voix forte ou criée, ce qui introduit nécessairement des déformations du signal et parfois des modifications de la structure syntaxique. Si de plus, l'auditeur est dans une ambiance bruyante, les chances de compréhension deviennent faibles puisque le signal est en partie masqué. Si, enfin, l'auditeur est de nationalité différente (a une mauvaise connaissance de la langue, a un vocabulaire moins étendu que le locuteur) cela entraîne une nouvelle diminution de la redondance. Dans la communication parlée habituelle, on peut se comprendre, à cause de la redondance, même si la parole est perturbée dans sa production (facteurs ectosémantiques, voix criée) ou bien dans sa transmission (canal,bruit de masque) ou bien dans sa réception (facteurs sémantiques, voix étrangère). Dans de nombreux cas réels de la vie industrielle, ces trois effets s'ajoutent et les contraintes articulatoires acoustiques, linguistiques, entraînent la perte de l'excès d'information nécessaire au décodage.

Il paraît utile de mieux connaître l'influence globale de ces facteurs sur l'intelligibilité du langage, de façon à disposer d'une estimation plus exacte des possibilités de communication, et à pouvoir proposer sur des bases plus réalistes, des améliorations de la communication verbale en milieu bruyant.

METHODE

Le matériel verbal est constitué de listes de mots français. Ces mots sont de deux syllabes seulement car ce sont les plus courants dans cette langue. La liste I, la plus difficile, comprend 300 mots issus des listes françaises R5B (FOURNIER 1951). La liste II est constituée de 200 mots plus concrets, issus du français de base (Français Fondamental -C.R.E.D.I.F. 1959). Chaque mot est précédé d'un article, ce qui permet de mieux "placer" la voix lors de l'enregistrement, et contribue à réguler le niveau d'émission physiologique. Les deux listes sont enregistrées, en voix parlée et en voix criée, avec un locuteur français masculin d'origine parisienne, mais sans accent particulier. Chaque enregistrement de parole est précédé d'une bande étalon de bruit blanc servant à mesurer le niveau d'intensité sonore de la voix. La reproduction des quatre enregistrements est effectuée en champ libre, le haut-parleur étant placé à 1 m. du sujet. Le niveau d'intensité sonore, mesuré à l'emplacement des oreilles, est fixé à 100dB.

Les sujets ont été choisis parmi des étudiants, après contrôle audiométrique (perte auditive moyenne inférieure à 5 dB). L'âge des 16 sujets (5 français et 11 étrangers) était compris entre 20 et 30 ans. Le tableau I indique le pourcentage de mots connus dans les deux listes, avant l'expérience. Pendant le test d'intelligibilité, les sujets n'ont rien à écrire mais ont pour consigne de simplement répéter ce qu'ils entendent. Les réponses des sujets étrangers ont été enregistrées afin de réduire, par un contrôle ultérieur, les erreurs provenant de la prononciation. L'intervalle de temps entre chaque mot est de 4 sec. ce qui permet parfois, à un auditeur français, de reconstituer "au vol" un mot à moitié masqué par le bruit.

Sujets	*Liste I*	*Liste II*
B	*75 %*	*85 %*
C	*46 %*	*76 %*
P	*88 %*	*98 %*

Tableau I - Pourcentages des mots connus dans les deux listes pour les sujets Brésiliens, Chiliens, Portugais. (Sujets Français : 100 %).

Le bruit de masque provient d'un générateur de bruit "blanc", filtré de manière à atténuer les fréquences élevées (bruit "rose"). Le spectre acoustique n'est pas rigoureusement uniforme à cause des résonances de l'enceinte et surtout de la cabine. Néanmoins ce spectre est à large bande, recouvrant largement le "spectre moyen" de la voix, et se rapproche de certains bruits rencontrés dans l'industrie ou les transports (machines à écriquer, à meuler, rotatives, souffleries, air comprimé, réacteurs d'avions, etc ...). La source de bruit est placée à 1 m. du sujet et on règle le niveau d'intensité sonore grâce à un atténuateur étalonné (111dB) de façon à obtenir si possible toute la courbe d'intelligibilité. Le niveau le plus élevé, qui peut dépasser 110 dB C sans distorsions appréciables, n'est utilisé que dans la situation limite suivante : liste II retransmise en voix parlée, pour un auditeur français et lorsque l'on désire des valeurs d'intelligibilité qui tendent vers 0%. L'autre situation limite

est la suivante ; liste I retransmise en voix criée pour un auditeur étranger et lorsque l'on désire des valeurs d'intelligibilité qui tendent vers 100 %. Nous verrons d'ailleurs qu'en général cela est impossible à réaliser, même si le niveau de bruit est négligeable.

Nous avons réalisé en laboratoire 4 situations expérimentales bien différenciées, soit par le type de voix, soit par le type de liste. Pour chaque situation, on effectue 16 épreuves d'intelligibilité. Par épreuve, une vingtaine de mesures sont nécessaires pour obtenir des coefficients de corrélation largement significatifs ($p < 0,01$). Pour une mesure, c'est-à-dire pour déterminer un point de la courbe d'intelligibilité, on utilise 20 à 30 mots. L'expérience a lieu dans une cabine audiométrique de 18 m^3 dont le temps de réverbération est de 0,5 sec. (Muraphone).

RESULTATS ET DISCUSSION

INTELLIGIBILITE DANS LE SILENCE

Le tableau II rassemble, pour les 16 sujets regroupés par nationalité, les pourcentages d'intelligibilité dans le silence en fonction du type de voix et du type de liste. Dans cette étude, on appelle "intelligibilité" le nombre de mots correctement répétés par les auditeurs français et étrangers. Il s'agit, dans une certaine mesure, pour les auditeurs étrangers en particulier, d'un test d'articulation (KRYTER 1970).

	Voix parlée		*Voix criée*	
Sujets	*Liste I*	*Liste II*	*Liste I*	*Liste II*
F	*100 %*	*100 %*	*92 %*	*97 %*
B	*94,5 %*	*98,5 %*	*74 %*	*85,5 %*
C	*86 %*	*93,5 %*	*60,5 %*	*82,5 %*
P	*96,5 %*	*95 %*	*84 %*	*94,5 %*

Tableau II - Intelligibilité dans le silence pour les sujets Français, Brésiliens, Chiliens, Portugais, en fonction du type de voix et du type de liste. Le niveau de reproduction sonore en champ libre est de 100 dB.

Pour les sujets français, l'intelligibilité est excellente dans tous les cas, sauf en voix criée avec la liste la plus difficile (perte d'intelligibilité de 8 %).

Les Brésiliens ont peu de difficultés à répondre quel que soit le type de liste, s'il s'agit de voix parlée (perte inférieure à 5 %). La voix criée entraîne une perte de 26 % avec la liste I et de 15 % avec la liste II.

Les chiliens ont une moins bonne connaissance de la langue et on observe, déjà avec la voix parlée, une baisse d'intelligibilité d'autant plus importante que la liste est difficile : perte de 14 % avec la liste I et seulement de 6 % avec la liste II. Avec la voix criée, la baisse est nettement plus accentuée : perte de 40 % avec la liste I et de 18 % avec la liste II.

Les Portugais ont à peine plus de difficultés à répondre que les Français, s'il s'agit de voix parlée avec les deux listes, ou encore de voix criée avec la liste la plus facile. Par contre, on note 16 % de perte en voix criée avec la liste la plus difficile.

D'une façon générale la voix parlée est mieux comprise que la voix criée et la liste II que la liste I. L'effet se fait sentir d'autant plus que la redondance initiale du message est déjà diminuée par les distorsions de la voix criée. En d'autres termes, le changement de type

de voix et les effets prosodiques qui en découlent, entraînent un glissement du niveau phonétique vers un niveau phonologique inhabituel, propre à la voix criée, où le décodage est plus difficile, faute d'indices et d'apprentissage. Les Français eux-mêmes peuvent être considérés comme des auditeurs étrangers s'il s'agit de voix criée.

Si l'on compare maintenant le pourcentage de mots connus, avant l'expérience, et les valeurs de l'intelligibilité dans le silence en voix parlée (tableaux I et II), on observe que :

Les sujets Brésiliens ne connaissent que 75 % des mots de la liste I et 85 % pour la liste II, mais on enregistre dans le silence les intelligibilités suivantes : 94,5 % et 98,5 %. Ainsi, ces sujets répètent correctement mais sans les comprendre, environ 20 % des mots de la liste I et 14 % ceux de la liste II.

Avec les sujets Chiliens qui connaissent respectivement 46 % et 76 % des mots, les intelligibilités dans le silence sont : 86 % et 93,5 %. Ces sujets répètent donc correctement, mais sans les comprendre, environ 40 % des mots de la liste I et 18 % de ceux de la liste II.

Enfin, les sujets Portugais connaissent 83 % et 98 % des mots et en répètent 96,5 % et 95 %. Dans ce cas, les sujets répètent, sans les comprendre, 8,5 % des mots de la liste I, tandis que avec la liste II ils ne répètent même pas tous ceux qu'ils connaissent, n'étant pas assez sûrs de leur prononciation.

Ainsi, une des difficultés que l'on rencontre lorsque l'on veut évaluer l'intelligibilité d'une langue étrangère, provient du fait qu'il y a superposition d'un test de type logatome avec un test classique d'audiométrie vocale. En réalité, on fait apparaître une relation entre les niveaux d'intégration extrême, acoustiques et sémantiques, mais l'on ignore ce qui se passe aux niveaux intermédiaires, phonétiques, phonologiques et lexicaux-syntaxiques. L'application pratique des résultats obtenus doit donc se faire avec prudence.

INTELLIGIBILITE DANS LE BRUIT

Pour les 4 situations expérimentales considérées, et avec les 16 sujets, on a déterminé les droites de régression de l'intelligibilité en fonction du bruit de masque. En fait, la fonction n'est plus linéaire au voisinage du maximum d'intelligibilité, et on utilise les valeurs dans le silence, afin de tracer la courbe, par extrapolation, lorsque l'intensité du bruit tend vers zéro. Les familles de courbes permettent de connaître, en fonction de divers paramètres (voix, listes, sujets) les variations de niveau de bruit correspondant à un pourcentage d'intelligibilité donné et les variations d'intelligibilité pour un bruit donné (fig.1).

Le tableau III indique le niveau de bruit pour lequel l'intelligibilité est de 50 % suivant les situations expérimentales.

	Voix parlée				Voix criée			
Sujets	*Liste I*		*Liste II*		*Liste I*		*Liste II*	
F	*96,0*	*1,6*	*99,5*	*3,0*	*84,5*	*2,1*	*92,5*	*3,0*
B	*91,0*	*5,5*	*98,0*	*2,3*	*79,5*	*6,9*	*90,5*	*3,2*
C	*94,5*	*1,7*	*95,5*	*4,6*	*77,5*	*3,6*	*90,5*	*1,8*
P	*95,5*	*6,3*	*100,0*	*3,0*	*86,5*	*3,4*	*93,0*	*3,7*

Tableau III. Les nombres représentent les valeurs moyennes des niveaux de bruit de masque en dB (avec l'écart-type) masquant la moitié des mots, en fonction du type de voix et du type de liste ainsi que de la nationalité des auditeurs. Le niveau de la parole est de 100 dB.

On constate que le niveau de bruit "tolérable" maximum est d'environ 100 dB, c'est-à-dire le même que celui de la parole. Dans ces conditions, les auditeurs Français comprennent la moitié des mots de la liste II en voix parlée. Cette situation étant prise pour référence, de combien de décibels faut-il diminuer le bruit pour conserver 50 % d'intelligibilité dans les autres situations ? Les résultats sont donnés par le tableau IV et la figure 2.

	Voix parlée		*Voix criée*	
Sujets	*Liste I*	*Liste II*	*Liste I*	*Liste II*
F	*3,5*	*0*	*15*	*7*
B	*8,5*	*1,5*	*20*	*9*
C	*5*	*4*	*22*	*9*
P	*4*	*0*	*13*	*6,5*

Tableau IV. Les nombres représentent la diminution de bruit nécessaire pour retrouver une intelligibilité de 50 %. Le niveau de bruit initial est de 100 dB.

On voit ainsi de quelle manière l'influence du type de voix et du type de liste se trouve renforcée par le changement de nationalité. On constate que le rapport signal-sur-bruit doit être de 22 dB dans le cas le plus défavorable (liste I en voix criée pour un auditeur Chilien).

Cherchons maintenant, à l'aide des courbes de la figure 1 (et des courbes analogues pour les sujets Chiliens et Portugais), les variations d'intelligibilité pour un niveau de bruit donné. La perte d'intelligibilité centrée à 50 %, due au changement de la liste est rarement négligeable (voix parlée pour un auditeur Chilien : 8 %) et elle peut atteindre 62 % en voix criée pour un auditeur Brésilien. La perte d'intelligibilité centrée à 50 % due au changement de voix est plus importante (liste II pour un auditeur Chilien : 32 %), et elle peut atteindre 76 % avec la liste I pour un auditeur français. Le tableau V ci-dessous donne les pertes moyennes d'intelligibilité de chaque situation expérimentale, le niveau de bruit étant maintenu constant.

	Changement de voix		*Changement de liste*	
Sujets	*Liste I*	*Liste II*	*Voix parlée*	*voix criée*
F	*76 %*	*60 %*	*32 %*	*54 %*
B	*52 %*	*60 %*	*38 %*	*62 %*
C	*72 %*	*32 %*	*8 %*	*60 %*
P	*60 %*	*70 %*	*40 %*	*46 %*

Tableau V - Pertes d'intelligibilité dues au changement de voix (avec la liste I ou II) et pertes dues au changement de liste (avec la voix parlée ou criée).

Les pertes d'intelligibilité centrées à 50 %, pour les sujets étrangers par rapport aux sujets français sont indiquées par le tableau VI. Dans 5 cas sur 6 on vérifie l'importance de la diminution d'intelligibilité et une dégradation d'autant plus grande que la liste est difficile. On observe une perte maximum de 40 % avec les auditeurs Chiliens écoutant la liste I en voix criée. C'est également dans cette situation que les auditeurs Portugais sont le plus défavorisés par rapport aux Français. Par contre, c'est avec la liste I en voix parlée que les auditeurs Brésiliens perdent relativement plus d'information, soit 30 %.

	Voix parlée		Voix criée	
Sujets	Liste I	Liste II	Liste I	Liste II
B	30 %	16 %	24 %	22 %
C	12 %	34 %	40 %	16 %
P	4 %	0 %	12 %	- 2 %

Tableau VI. - Pertes d'intelligibilité dues au changement de sujets (les sujets français servant de référence) en fonction du type de voix et du type de liste.

Nous voyons que les pertes dues aux divers facteurs pris séparément ne sont pas négligeables, mais laissent en général une certaine probabilité de compréhension du message. Cherchons enfin à connaître les situations dans lesquelles la communication est impossible, la probabilité étant nulle. Si, par exemple, l'intelligibilité est de 50 % pour les sujets Français écoutant la liste II en voix parlée, elle est nulle dans tous les cas suivants :
Sujets Français ou étrangers, écoutant la liste I ou II en voix criée.

Lorsque l'intelligibilité est nulle dans la situation la plus difficile pour un auditeur étranger (liste I en voix criée) on remarque qu'elle est de 100 % pour un Français dans la situation la plus facile (liste II en voix parlée). Si on ne considère que la liste I, lorsque l'intelligibilité est nulle pour un Brésilien écoutant de la voix criée, elle est de 91 % pour un Français écoutant de la voix parlée. Avec les Chiliens, dans les mêmes conditions, les Français ont une intelligibilité de 97 % et avec les Portugais on trouve seulement 55 %.

CONCLUSION

Bien que cette étude ne soit qu'une première approche du problème de l'intelligibilité de la parole pour des auditeurs étrangers, nous avons pu évaluer les influences de trois facteurs intervenant à l'émission (prosodie), à la transmission (bruit) et à la réception du message (auditeur étranger). Il en existent beaucoup d'autres, par exemple les contraintes du message lui-même (PICKETT 1969), le temps de réverbération du local d'écoute, la vigilance et la fatigue auditive.

Malgré le niveau d'instruction élevé des auditeurs (étudiants) et le fait qu'ils répètent certains mots sans les comprendre, la baisse d'intelligibilité avec les listes abstraites en voix criée est très importante. Quelles sont les conséquences pratiques prévisibles ? Il est nécessaire de poursuivre cette étude avec comme auditeurs, non plus une population d'étudiants mais des ouvriers Français et étrangers (Arabes, Portugais, Espagnols, Italiens). En réalité l'analyse des situations de travail montre qu'il existe parfois un décalage important entre les qualifications accordées aux travailleurs et les contraintes de la situation de travail dont la simplicité peut n'être qu'apparente.

Un problème pratique se pose, au niveau de la réponse des sujets, s'il s'agit d'ouvriers étrangers qui ne parlent pas assez bien le Français pour répondre verbalement. Faut-il dans ces conditions effectuer le test d'intelligibilité à partir, non plus de séries de mots mais de séries d'images représentant des objets concrets ou des situations simples ? Doit-on utiliser un "vocabulaire" adapté à chaque milieu de travail ?

Un nouveau test d'intelligibilité (IRVINE 1974) semble intéressant avec des auditeurs étrangers : il s'agit non plus de faire répéter des listes de mots, mais de répondre à des questions formulées de telle manière que le contexte aide la compréhension et que le mot à répondre soit utilisé dans la question. Il est en effet plus important de savoir si le message a été compris, si l'on a bien compris ce que le locuteur voulait dire, plutôt que de savoir si le sujet est capable de répéter des mots sans s'écarter trop du cadre phonologique de la langue considérée. Des recherches sont en cours en laboratoire et en atelier sur ces question.

BIBLIOGRAPHIE

ALLEN G.D. (1971) Acoustic Level and Vocal Effort as Cues for the Loudness of Speech. JASA 49-6-2-1831-1841

BOE L.J. et LARREUR D. (1974) Les caractéristiques intrinsèques de la fréquence laryngienne : production, réalisation et perception. 5ème journée d'étude. G.A.L.F. Orsay. 19-28

BRANDT J.F. and RUDER K.F. (1969) Vocal Loudness and Effort in Continuous Speech. J.A.S.A. 46.6.2. 1543-1548

CHERRY C. and WILEY R. (1967) Speech Communication in Very Noisy Environments. Nature. Vol. 214 June 10. p. 1164

C.R.E.D.I.F. (1959) Le français fondamental 1er et 2ème degré. Ed. I.P.N. Paris

DREYFUS-GRAF (1973) Tests d'intelligibilité de la parole codée. Symposium : Intelligibilité de la parole. Liège. p. 209-221

DUTHEIL R., FORET J., PARANT C., ROSTOLLAND D. (1973) Influence de divers types de distorsion sur la compréhension des messages verbaux en milieu bruyant. Rapport C.N.A.M./Physiologie n° 36. Paris. 63 pages

FOURCIN A.J. (1974) Speech Perception in the Absence of Speech Productive Ahility. IV. 1.20. in Speech and Hearing. University College. London.

FOURNIER J.E. (1951) Audiométrie vocale. Ed. Maloine. Paris

FRY D.B. (1966) Mode de perception des sons du langage in Phonétique et Phonation. 191-206. Ed. Masson. Paris

GARDE E. (1965) La Voix. N° 627. Presses Universitaires de France. Paris

GIOLAS T.G. and COOKER H.S. (1970) The predictability of wordo in sentences. The Jour. of Aud. Research. 10-4. 328-334

GUIRAUD P. (1972) La Sémantique. N° 655. Presses Universitaires de France. Paris

GUIRAUD P. (1973) La Sémiologie. N° 1421. Presses Universitaires de France. Paris

IRVINE D.H. (1974) A new Type of Speech Intelligibility Test. Ergonomics. 17-6. 783-788

KRYTER K.D. (1962) Methods for the Calculation and Use of the Articulation Index. J.A.S.A. 34.11. 1689-1697

KRYTER K.D. and WILLIAMS C.E. (1966) Masking of Speech by Aircraft Noise. J.A.S.A. 39.1. 138-150

KRYTER K.D. (1970) The Effects of Noise on Man. Academic Press. New York and London.

LAFON J.C. (1963) Intelligibilité du message phonétique. in Communications et Langages. 187-198. Ed. Gauthier-Villars. Paris

LARREUR D. et BOE L.J. (1974) Synthèse paramétrique de l'intonation de la phrase énonciative en français. 5ème journée d'étude. G.A.L.F. Orsay.63-71

LEHMANN R. (1962) Etude psychophysique de l'intelligibilité du langage. Thèse. Faculté des Sciences de Paris. 43 p.

LEIPP E. (1968) Le problème de l'intelligibilité de la parole. G.A.M. Faculté des Sciences. Paris

LEIPP E. (1968) Structure physique et contenu sémantique de la parole. Revue d'Acoustique. G.A.L.F. N° 3-4. p. 259-266

MALMBERG B. (1966) Stabilité et instabilité des structures phonologiques. p. 207-232. in Phonétique et Phonation. Ed. Masson. Paris

MALMBERG B. (1969) Phonétique Française. Ed. Hermods Malmö - Suède

METTAS O. (1966) Les facteurs ecto-sémantiques du discours et leur caractérisation par synthèse. in Phonétique et Phonation, p. 177-188. Ed. Masson - Paris

MEYER-EPPLER W. (1963) Problèmes informationnels de la communication parlée. in Communications et Langages. 51-65. Ed. Gauthier-Villars. Paris

MILLER G.A. and HEISE G.A. (1951) The Intelligibility of Speech as a Function of the Context of the Test Materials. J. Exp. Psychol. 41. 329-335

MILLER J.D. (1974) Effects of Noise on People. J.A.S.A. 56.3. 729-764

MOLES A. (1963) Les bases de la théorie de l'information et leur application aux langages. in Communications et Langages 15.33. Ed. Gauthier-Villars. Paris

MOLES A. (1966) Méthode cybernétique et structures linguistiques. in Phonétique et Phonation. 233-251. Ed. Masson. Paris.

PIAGET J. (1974) Le Structuralisme. N° 1311. Presses Universitaires de France. Paris

PICKETT J.M. (1956) Effects of Vocal Force on the Intelligibility of Speech Sounds. J.A.S.A. 28.5. 902.905

PICKETT J.M. (1958) Limits of Direct Speech Communication in Noise. J.A.S.A. 30.4. 278-281

PICKETT J.M. and POLLACK I. (1958) Intelligibility at High Voice Levels and The Use of a Megaphone. J.A.S.A. 30.12. 1100.1104

PICKETT J.M. (1969) Message Constraints, A. Neglected Factor in Predicting Industrial Speech Communication. A.S.H.A. Reports N° 4. 121-128

PIMONOW L. (1968) La reconnaissance du langage en fonction du débit informationnel instantané. Revue d'Acoustique. G.A.L.F. N° 3-4. p. 193-205

POLLACK I., PICKETT J.M. (1958) Masking of Speech by Noise at High Sound Levels. J.A.S.A. 30.2. 127-130

POLLACK I.,RUBENSTEIN H. and DECKER L. (1959) Intelligibility of Known and Unknown. Message Sets. J.A.S.A. 31.3. 273-279

POLLACK I. (1964) Message Probability and Message Reception. J.A.S.A. 36.5 937-945

ROSTOLLAND D. and PARANT C. (1973) Distorsion and Intelligibility of Shouted Voice. Symposium : Speech Intelligibility. Liège. p. 293-304

ROSTOLLAND D. and PARANT C. (1974) Physical Analysis of Shouted Voice. The 8th International Congress on Acoustics. London. 14 p.

VALLANCIEN B. (1963) Aspect informationnel de l'intelligibilité. in Communications et Langages. 199-203. Ed. Gauthier-Villars. Paris

VIVES R. et coll. (1974) Reconnaissance de grands dictionnaires prononcés par plusieurs locuteurs. 5ème journée d'étude. G.A.L.F. Orsay. 125-131

WAJSKOP M. (1967) Identification de voyelles en fonction de leur durée. in Proc. of the 6th Int. Cong. of Phonetics Sciences. Prague.

WEBSTER J.C. and KLUMPP R.G. (1962) Effects of Ambient Noise and Nearby Talkers on a Face-to-Face Communication Task. J.A.S.A. 34.7. 936-941

WEBSTER J.C. (1965) Speech Communications as Limited by Ambient Noise. J.A.S.A. 37.4. 692-699

WEBSTER J.C. (1968) Effects of Noise on Speech Intelligibility. Proceedings of the National Conference on Noise as a Public Health Hazard. A.S.H.A. N°4. WASHINGTON D.C.

WILLIAMS C. and HECKER M. (1968) Relation between Intelligibility Scores for Four Test Methods and Three Types of Speech Distortion. J.A.S.A. 44.4. 1002-1006

WISNER A. (1965) Effets des bruits sur l'homme au travail. Synopsis 7.14. 17.24

WISNER A. (1971) Signaux sonores en milieu bruyant. Complexité de la situation réelle et expériences de laboratoire. 17ème Congrès de Psychologie. Liège.

FIGURE 1

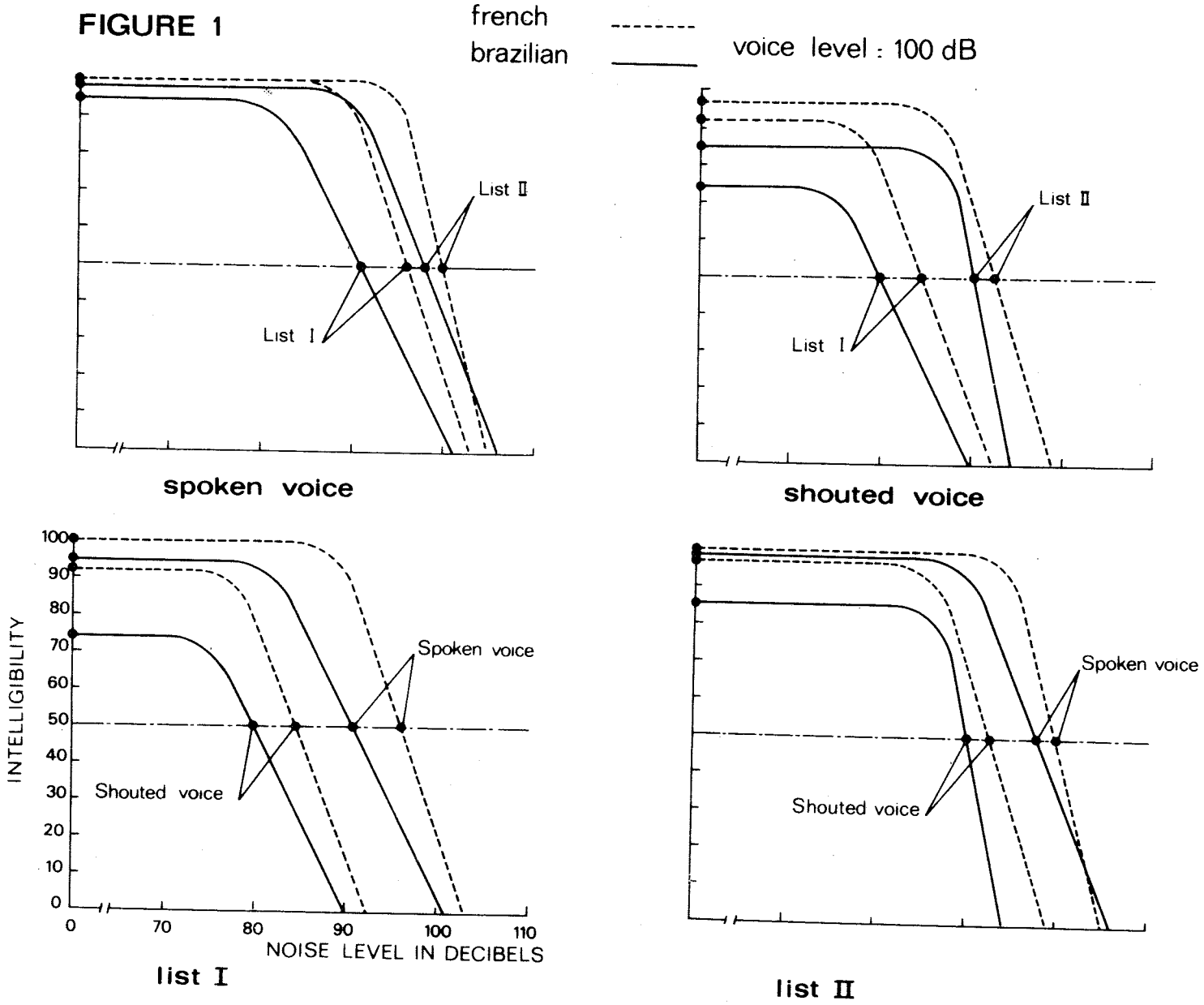
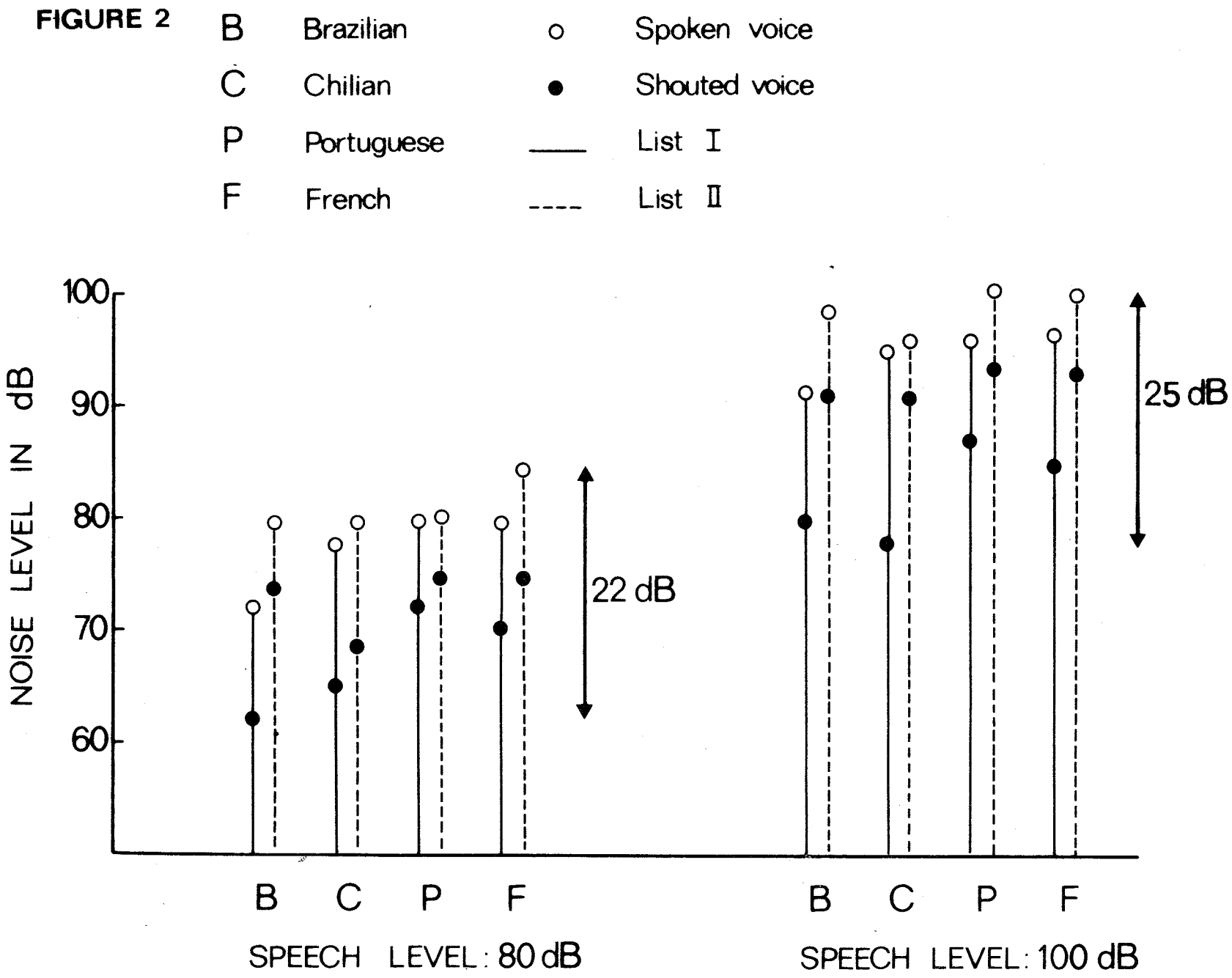

french
brazilian
voice level : 100 dB
List II
List I
spoken voice
List II
List I
shouted voice
INTELLIGIBILITY
100
90
80
70
60
50
40
30
20
10
0
Spoken voice
Shouted voice
0
70
80
90
100
110
NOISE LEVEL IN DECIBELS
list I
Spoken voice
Shouted voice
list II

FIGURE 2


B Brazilian
C Chilian
P Portuguese
F French
Spoken voice
Shouted voice
List I
List II
NOISE LEVEL IN dB
100
90
80
70
60
22 dB
25 dB
B C P F
B C P F
SPEECH LEVEL: 80 dB
SPEECH LEVEL: 100 dB

THEME 3A

AIDE AUX HANDICAPES

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UN SYSTEME DESTINE A LA REEDUCATION DES DEFICIENTS AUDITIFS PROFONDS : LE SYSTEME P.A.R.M.E.

P. LORAND, R. BESSON Département E.T.A., Centre National d'Etudes des Télécommunications à Lannion.

Docteur R. MAZEAS , Centre de Traitement de l'Ouïe et de la Parole à Fougères.

RESUME

Dans la première partie de cette communication, les auteurs décrivent le principe des appareils "P.A.R.M.E." étudiés et construits au C.N.E.T. d'après l'idée de M. PIMONOW. On donne les caractéristiques techniques essentielles. Dans la seconde partie sont exposés les résultats obtenus avec P.A.R.M.E. dans les classes d'enfants sourds profonds, au C.T.O.P. de Fougères, depuis 1970.

SUMMARY

In the first part of this communication, the authors describe the principle of the apparatus called "P.A.R.M.E.", which was studied and manufactured in CNET, on the idea of L. PIMONOW. The main technical characteristics are given. In the second part, the results obtained with P.A.R.M.E. since 1970, with deep deaf pupils, at C.T.O.P of Fougères, are exposed.

UN SYSTEME DESTINE A LA REEDUCATION DES DEFICIENTS AUDITIFS PROFONDS LE SYSTEME "P.A.R.M.E."

par P. LORAND, R. BESSON Centre National d'Etudes des Télécommunications à Lannion

et Dr R. MAZEAS, Centre de Traitement de l'Ouïe et de la Parole à Fougères.

I - HISTORIQUE

En 1969, M. L. PIMONOW, Directeur à l'Ecole Pratique des Hautes Etudes, Conseiller Scientifique au CNET, obtint un contrat de la D.G.R.S.T. permettant le financement de l'étude et de la réalisation d'un appareil destiné à la rééducation des classes d'enfants sourds, qui devait se terminer par l'étude d'une prothèse individuelle utilisant les mêmes principes.

A partir de septembre 1969, le Département Etudes et Techniques d'Acoustique du CNET construisit sept appareils de classe intitulés P.A.R.M.E. (Prototype d'Appareil pour la Rééducation des Mal - Entendants). Présentés dans un coffret de 40 cm × 30 cm × 15 cm environ, ils étaient réalisés sous forme de cartes enfichables permettant aisément le dépannage et les modifications.

Chacun des sept appareils est conçu pour desservir six enfants dans la classe. Le professeur utilise un microphone et chaque élève dispose d'un écouteur de prothèse à hautes performances et peut effectuer son auto-contrôle grâce à un microphone individuel. Ces appareils furent confiés en particulier au C.T.O.P. de Fougères, à l'Institut National des jeunes sourds et au Centre d'études pédagogiques pour la rééducation des déficients auditifs (rue Godot de Mauroy), à Paris.

A la rentrée scolaire de 1971, ils furent tous réunis à Fougères pour des raisons évidentes de facilité d'entretien et pour pouvoir effectuer une expérimentation sérieuse sur des élèves plus nombreux. Il apparut en outre que la manière dont les éducateurs utilisent les appareils joue un rôle primordial dans les résultats : à ce point de vue, les "conseils" des techniciens du CNET étaient plus faciles à dispenser à un seul Centre. Comme les méthodes pédagogiques actuelles nécessitent une grande liberté de mouvement des enfants, divers essais furent tentés : liaisons par boucles magnétiques et réception par boîtiers amplificateurs individuels, etc.....

A la fin de 1972, trois prototypes de prothèse individuelle portable furent réalisés, dont les dimensions sont de 16cm × 11cm × 4cm. Un nouvel appareil de classe fut aussi étudié, tenant compte des modifications imposées par l'expérimentation des appareils précédents, et en cherchant un prix de revient minimum.

.../...

Tous ces appareils sont actuellement encore en service au C.T.O.P. de Fougères, intégrés dans des installations permettant des exploitations diverses, en particulier les liaisons par émetteurs-récepteurs à modulation de fréquence.

II - DESCRIPTION TECHNIQUE

Les appareils de classe, comme la prothèse P.A.R.M.E., sont destinés à aider l'audition des sourds très profonds : pour ceux-ci, les résidus d'audition sont en général concentrés dans la partie basse du spectre audible (en dessous de 1 000 Hz). Cependant des essais audiométriques ont montré que des fréquences plus élevées peuvent être entendues pourvu qu'elles soit présentées avec un niveau très élevé.

P.A.R.M.E. est donc caractérisé par trois points principaux :

1/ - Une transposition partielle des fréquences au-dessus de 1 000 Hz qui sont présentées sous forme de deux bandes de bruit en dessous de 1 000 Hz. Un signal non traité, amplifié, est juxtaposé à ces voies transposées.

2/ - Un niveau de sortie sur l'écouteur très élevé (pouvant atteindre 140 dB/ 2.10^{-5} Pa).

3/ - Un système de compression sélectif muni de constantes de temps telles qu'on obtient à travers l'appareil une augmentation apparente du niveau de sortie des plosives par rapport au niveau des sons permanents.

La figure ci-après donne le schéma-bloc des appareils.

Caractéristiques essentielles

1/ - Compresseur sélectif

. Filtre de bande 300/5 000 Hz dans la boucle de réaction.

. Taux de compression du compresseur dans la bande en régime permanent : 30 dB.

. Constantes de temps - établissement : 10 millisecondes.
- retour au repos: 50 millisecondes.

2/ - Voies transposées

1ère voie : . Analyse dans la bande 1000 - 2000 Hz.

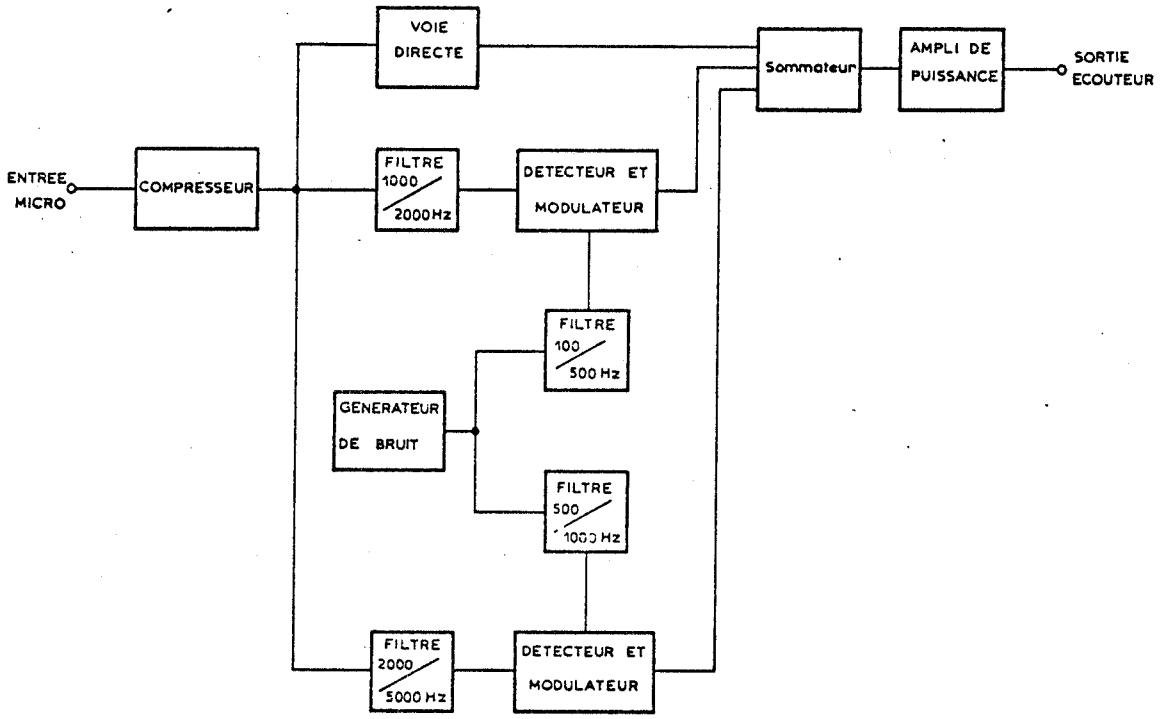
. Sortie : Modulation d'amplitude d'une bande de bruit blanc filtrée entre 100 et 500 Hz.

2ème voie : . Analyse : 2000/5000 Hz (peut dépasser 5000 Hz sans inconvénient).

. . Sortie : Modulation d'une bande de bruit blanc filtré entre 500/1000 Hz.

.../...

SCHEMA BLOC DE LA PROTHESE "PARME"



3/-Sommation et amplification de puissance

- Egalisation des niveaux de voies transposées et de la voie directe avant sommation et amplification de puissance.
- Sortie d'un signal à très fort niveau dans un écouteur de bonne qualité. Niveau nécessaire mesuré sur un coupleur de 2 cm 3 : 140 dB re. 2.10^{-5} Pa.

4/-Système anti-bruit

Un transistor à effet de champ court-circuite le générateur de bruit en l'absence de signal à l'entrée, afin de diminuer artificiellement le bruit de fond de l'appareil, en l'absence de signal. Ce dispositif n'est pas strictement nécessaire pour les sourds très profonds ; il pourrait être évité aussi avec des modulateurs de meilleure qualité.

5/-Filtres

Pente souhaitable 24 dB à l'octave ; 18 dB peuvent être suffisants. Pour des raisons de consommation, on a utilisé des filtres passifs miniaturisés sur les prototypes de prothèse.

.../...

6/ - Alimentations

Pour les appareils de classe : à partir du secteur 220 V alternatif.

Pour la prothèse portable :

+ 3,6 volts : 3 accumulateurs AGLO - 1,22 volt - 0,6 Ampère-Heure
- 3,6 volts : 3 accumulateurs AGLO - 1,22 volt - 0,6 Ampère-Heure.

Consommation à vide 10 mA, consommation pour un signal de parole continu et un niveau de sortie moyen de 130 dB : 30 mA. L'autonomie de fonctionnement est de 24 heures sur un signal de parole continu, sans recharge des accumulateurs.

7/ - Microphone

. Bande passante souhaitable 100/10 000 Hz = $\pm$ 5 dB.

. Impédance et efficacité en fonction de l'étage d'entrée. On a retenu le modèle LEM DO 35.

8/ - Ecouteur

. Possibilité d'obtenir 140 dB avec moins de 5 % de distorsion.

. Bande passante à 10 dB jusqu'à 2500 Hz. Très peu d'écouteurs donnent satisfaction sur ces points.

La construction de ces appareils "de classe" ne présente pas de difficultés et pourrait être entreprise industriellement par toute firme d'électronique intéressée ; mais la rééducation effectuée en classe devrait être complétée par le port permanent d'une prothèse individuelle subminiaturisée. Or la conception et la réalisation de tels appareils de dimensions suffisamment petites (8 cm × 5 cm × 2 cm environ) ne peuvent être entreprises que par un nombre très restreint de firmes, hautement spécialisées dans ce domaine. La complexité du montage et le nombre d'acheteurs potentiels relativement petit pour ces prothèses spéciales conduiront à des prix de revient élevés. Ces raisons expliquent qu'on n'ait pas réussi actuellement à faire fabriquer la prothèse PARME en série et qu'elle n'est donc pas disponible commercialement.

III - UTILISATION DE PARME ET RESULTATS OBTENUS

Le travail avec l'appareil P.A.R.M.E. a été réalisé dans deux groupes d'enfants sourds profonds : le premier, d'âge de 10 à 12 ans, le second, d'âge de 4 à 6 ans.

Les mesures auditives pratiquées en collaboration avec le C.N.E.T. ont montré des seuils d'audition exprimés ici en niveau par rapport à 2.10^{-5} Pa variant de 108 à 121 dB pour les fréquences de 80 Hz à 2000 Hz.

.../...

Le travail auditif s'est fait d'une part en collectif en classe, et d'autre part en individuel en cabine, à raison de 6 à 8 séances hebdomadaires d'un quart d'heure chacune.

Le travail d'éducation et d'adaptation auditive a été le même dans les deux groupes. Dans un premier temps, après familiarisation avec l'appareil et utilisation couplée de la labio-lecture, le travail a consisté à faire reconnaître à l'audition pure des mots et expressions opposées uniquement par leur nombre de syllabes ; par exemple : oui - papa - un bateau - il fait du vent -. Cet exercice, au départ assez difficile pour un sourd profond, à cause de son Δ t très long, a été réussi rapidement après quelques séances.

Le deuxième exercice s'attaque à la différenciation auditive des consonnes sourdes et sonores : toi - moi ; papa - maman ; un chapeau - un chameau, en faisant remarquer les différences à l'enfant et, au besoin, en prolongeant l'émission des phonèmes sonores.

La reconnaissance des divers phonèmes est ainsi poursuivie toujours par opposition deux à deux ; par exemple : bateau - râteau ; bateau - chapeau ; oui - si ; chut - si, ainsi que pour les voyelles : i - a ; i - è ; i - o ..... etc.

Il en a été de même pour les explosives sourdes et sonores : P - B ; T - D ; K - G.

Il va de soi qu'au départ de chaque exercice toutes les techniques classiques de reconnaissance sont utilisées, en particulier la labio-lecture, le toucher, etc....

Mais rapidement, en quelques séances, la reconnaissance des mots et expressions présentés se fait à l'audition seule

De cette façon, nous sommes arrivés à faire percevoir des oppositions phonétiques de plus en plus précises. Cependant, certaines confusions persistent assez longtemps : i - u, par exemple.

Il est nécessaire, bien entendu, de revenir souvent en arrière et surtout après les retours de vacances.

Parallèlement à l'éducation auditive, le travail a été mené sur le plan de l'orthophonie, du rythme, du timbre.

Les résultats que nous avons obtenus avec l'appareil P.A.R.M.E. ne sont bien entendu pas chiffrables. Mais les professeurs de classe (qui n'étaient pas responsables du travail individuel en cabine) et les parents sont unanimes à dire les progrès accomplis. Ceux-ci sont d'autant plus patents dans le groupe des enfants de 10 - 12 ans, qu'ils avaient auparavant été soumis aux amplificateurs classiques, linéaires, sans transposition, sans compression et au casque.

.../...

Nous avons classé les résultats obtenus sous cinq rubriques :

- D'abord l'intérêt pour les séances de travail en cabine. On sait la difficulté de ces séances et le degré d'attention réclamée par le professeur orthophoniste et, par là même, le peu d'empressement de l'enfant à s'y donner. Dans le cas du P.A.R.M.E., la plupart des enfants étaient heureux de venir aux exercices, considéraient ceux-ci comme un jeu et voulaient savoir qui avait le mieux réussi.

- Sur le plan de la reconnaissance auditive pure, sans l'aide de la lecture labiale, les résultats sont très positifs si on se réfère aux peu de restes audiométriques. Au bout de deux ans de travail, il apparaît qu'un système de référence et d'opposition phonétique s'est créé. Cet apprentissage ne permet pas la reconnaissance d'une conversation courante. Mais, dans un cadre précisé d'avance, comme : noms de fleurs - d'animaux - prénoms - expression sur un thème, on a pu noter jusqu'à 8 et 9 répétitions correctes sur 10, chez plus de la moitié des enfants.

- Sur le plan de l'orthophonie, c'est-à-dire de la qualité de l'articulation, les résultats sont encore meilleurs. Au bout de quelques séances, les phonèmes difficiles à acquérir habituellement chez le grand sourd : phonèmes à caractéristiques aiguës, en particulier les constrictives sourdes : f - s - ch, se précisent et s'améliorent. Il en est de même de la voyelle i, tellement difficile à poser, du fait de ses formants aigus. De la même façon, les oppositions sourdes-sonores se mettent en place rapidement.

- Sur le plan de l'auto-contrôle auditif, on note également un apprentissage rapide. La voix est mieux posée, plus régulière dans son timbre et dans son rythme. Les variations du fondamental laryngé restent dans les limites de la normale. La voix de fausset, quand elle existe, disparaît sous le contrôle du P.A.R.M.E. La seule répétition, avec contrôle auditif, suffit à faire corriger les oppositions sourdes-sonores. Le recours au toucher comme moyen de correction devient inutile. De même, dans le cas de phonèmes omis, la seule répétition par le professeur permet la correction.

- Enfin, sur le plan de l'application prothétique.
Comme nous n'avons pas à notre disposition de prothèses individuelles du même type que le P.A.R.M.E., nous avons dû nous orienter vers l'application de prothèses classiques de type III, certaines conventionnelles, d'autres de type contours d'oreille, parfois avec compression. Nous avonsété frappés par l'aisance avec laquelle les enfants ont accepté puis utilisé ces prothèses, comme si l'éducation au P.A.R.M.E. avait éveillé ou créé chez les grands sourds un système de référence utilisable même avec des appareils sans transposition.

Pour illustrer ces résultats, voici rassembléesles appréciations des professeurs au sujet du comportement d'un enfant sourd très profond à différentes étapes de sa scolarité (3 à 7 ans).

.../...

Octobre 1970 : Très appliqué en classe pour les différentes activités et les séances de parole. Ne réagit pas aux amplificateurs (linéaires).

Décembre 1970 : Ne s'intéresse pas à l'éducation auditive. Accorde toute son attention à la lecture labiale et aux exercices de parole.

1er trim. 1971 : Ne marque aucun intérêt pour les séances d'écoute. Très bonne lecture labiale.

2e trim. 1971 : Tout travail aux amplificateurs l'ennuie.

2e année 1972 : Entraînement auditif intensif avec de minimes résultats.

3e année 1973 : Début de travail au P.A.R.M.E. Petits résultats en reconnaissance auditive qui permettent d'envisager une prothèse individuelle.

4e année 1974 : 2e année de P.A.R.M.E. Malgré une surdité profonde, s'intéresse de plus en plus aux séances d'écoute. Réussit à discriminer une trentaine de mots et d'expressions courantes. Traitement à poursuivre.

En conclusion de cette brève note, nous pouvons dire :

Tout d'abord, que les appareils P.A.R.M.E. que nous avons eus entre les mains depuis quatre ans sont des appareils d'une solidité à toute épreuve, et qu'ils n'ont presque jamais été en panne. Pour qui connaît l'utilisation des prothèses chez les enfants sourds, cette remarque est d'un grand poids.

Le principe du P.A.R.M.E. semble être bien adapté à la surdité profonde. Jamais jusqu'à présent aucun type d'appareil ne nous a donné autant de satisfaction. Il est difficile de dire pour l'instant ce qui, dans les caractéristiques de l'appareil, est le plus important : le codage de la partie aiguë du spectre en bruit blanc filtré grave, la compression des plosives, la puissance de sortie, l'utilisation des écouteurs miniatures. D'autres études plus fouillées seront nécessaires pour cela.

Je voudrais cependant insister sur la notion de puissance de sortie. Nombreux ont été ceux qui ont crié gare quand on a parlé de 135 dB à la sortie de l'écouteur. Nous connaissons les surdités par traumatisme sonore, et aussi le problème du seuil de la douleur et celui du recrutement. Mais les mesures précises montraient chez ces enfants un seuil liminaire très élevé. D'où la necessité de passer ce cap pour pouvoir leur faire parvenir des informations auditives. En moyenne, l'appareil travaille à 15 dB au-dessus du seuil audiométrique de la bande 100 - 2000 Hz. Nous avons alors l'impression qu'enfin nous atteignons un niveau intéressant pour les enfants, et différent de ce

qu'ils ont perçu auparavant. Les réactions des enfants à cette puissance sont d'abord une réaction de surprise, d'étonnement, puis de contentement : "oh, c'est fort ! J'entends bien" ! Ils supportent généralement bien cette puissance élevée. Une seule fois, au début, nous avons noté une réaction de rejet. Un système de réglage individuel a été alors mis en place et toute manifestation de gêne a disparu chez ce sujet.

Quant aux audiogrammes successifs, ils ont été stables et n'ont pas varié pendant le cours de l'expérience, ni après. Il est vrai que nous n'avons jamais obtenu chez les enfants en étude, ni au départ ni ensuite, les fréquences aiguës de 3500 et 4000. Et ce, malgré l'utilisation d'un audiomètre mis également au point par le C.N.E.T. et qui nous permet de monter à 140 dB sur les fréquences à partir de 500 Hz.

Il faut bien admettre qu'à partir du moment où on a affaire à de grands sourds, il est nécessaire de dépasser leur seuil d'audition si on veut leur faire passer un message sonore. De plus, mettre 15, voire 20 dB au-dessus du seuil ne semble pas, à l'expérience que nous avons, être toxique pour les cellules sensorielles.

En définitive, le P.A.R.M.E. est pour nous un bon système d'éducation auditive de l'enfant sourd profond. Nous espérons grandement qu'il pourra être miniaturisé rapidement pour pouvoir être porté individuellement par les enfants qui auront été éduqués avec ce procédé. Car les résultats définitifs devraient encore être meilleurs.

-=-=-=-=-

6èmes JOURNEES D'ETUDE SUR LA PAROLE
Toulouse 28 au 30 Mai 1975

UN SYSTEME D'AIDE VISUELLE AUX SOURDS PROFONDS

M. LAGNEAU

RESUME : UN SYSTEME D'AIDE VISUELLE AUX SOURDS PROFONDS.

Ce système d'aide visuelle a permis de représenter le signal acoustique en deux étapes successives :

1) Par simulation numérique (une méthode de compression de données a permis de visualiser des mots par des trajectoires planes orientées).

2) Par une réalisation analogique de faible coût : un opérateur de réduction câblé permet de fournir en temps réel ces trajectoires sur oscilloscope.

SUMMARY : A SYSTEM OF VISUAL AIDS FOR THE SEVERELY DEAF.

This system of visual aids has lead to represent the acoustic signal successively via :

1) Numerical simulation : words have been displayed as oriented trajectories in the plane using a feature extraction method.

2) A low-cost analog circuitry : a hardwired data compression operator displays in real time these trajectories on a scope .

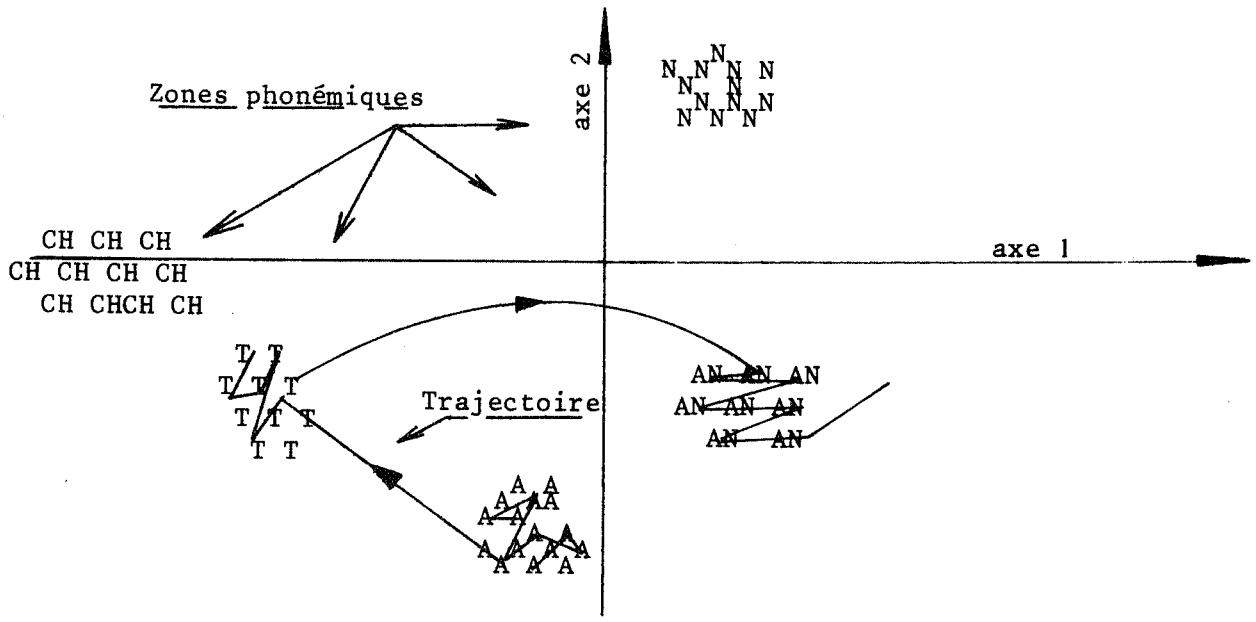
6ième Journée d'Etude sur la parole , G.A.L.F. Toulouse 28-30 Mai 1975

U N S Y S T E M E D ' A I D E
V I S U E L L E
A U X S O U R D S P R O F O N D S

M.LAGNEAU,Laboratoire d'Automatisme,E.N.S des Télécommunications,
46 rue Barrault,75 636 Paris Cedex 13.

I) I N T R O D U C T I O N

Une des applications interessantes des travaux sur la reconnaissance de la parole pourrait être orientée vers l'aide aux sourds,surtout dans l'acquisition du langage. Il apparaît donc essentiel de leur fournir un appareil qui ne leur demande pas un effort d'adaptation trop long ou trop important.
Le système étudié au Laboratoire de l'E.N.S.T est basé sur l'analyse spectrale et a pour but de donner une image de la parole,non au niveau du son,mais au niveau du mot : on représentera un mot par une trajectoire plane , ce qui suppose qu'à tout instant l'on puisse définir le signal acoustique à l'aide seulement de deux paramètres bien choisis/27/ , /28/
Soit par exemple le mot ' ATTENDS ' : dans un cas idéal,la visualisation de ce mot pourrait se présenter ainsi :



Un tel appareil est évidemment d'un grand intérêt pour la pédagogie , mais il se heurte à certains problèmes :

- peut on définir un signal acoustique aussi riche en informations à l'aide de deux paramètres à tout instant ?
- existe t-il une indépendance vis à vis du locuteur ? Il est indispensable dans un premier temps que l'on puisse fournir au sujet à rééduquer un modèle à reproduire , modèle fourni par le professeur , ou du moins la voix du professeur , or le but de la rééducation n'est pas pour l'élève une imitation inconditionnelle de la voix du professeur , mais de la découverte de la sienne propre. Le modèle fourni ne doit pas trop varier avec l'accent , le timbre,enfin tous les caractères personnels affectés à la voix.
- les trajectoires seront-elles bien acceptées et comprises facilement par l'élève ?

En effet , celui-ci doit sans effort être habitué à une structure géométrique la plus simplifiée possible. Numériquement , ce problème semble avoir donné des résultats encourageants /29/ , /30/.

Actuellement , un effort est orienté vers une solution peu coûteuse et entièrement analogique , et dont les principes reposent évidemment sur la simulation numérique faite auparavant.

II) LES DEUX ETAPES FONDAMENTALES DANS LA DESCRIPTION DU SYSTEME

II-1 Simulation numérique

L'expérimentation du système a pu être menée sur une échelle suffisante pour mettre au point les détails de la mise en pratique (hardware).

L'analyse spectrale a été réalisée en utilisant un vocoder simplifié à 20 canaux constitués par un banc de filtres passe-bandes du second ordre/29/.

Les données d'apprentissage ont été constituées par des voix de quatre enfants âgés de dix ans environ,prononçant un vocabulaire simple de cinquante mots,puis par des voix de deux personnes adultes prononçant cette dernière liste+ un vocabulaire constitué de quinze mots; cette dernière liste a été prononcée quatre fois afin de tester la stabilité des formes dans le temps.

L'enregistrement s'est fait au laboratoire d'acoustique de l'E.N.S.T dans d'excellentes conditions.

Voici la liste des quinze mots;en regard de chacun d'eux se situe le nombre d'échantillons à la période T = 10 ms.

1) LE ROND 78,78,74,77
2) LE CARRE 88,98,93,92
3) LE LOSANGE 107,105,108,122
4) LE VECTEUR 140,102,132,114
5) A DROITE 71,74,72,72

6)	A GAUCHE	87,87,91,95
7)	EN BAS	66,74,69,72
8)	AU DESSUS	83,86,84,90
9)	FAIRE	47,43,47,45
10)	UNE ROTATION	110,118,110,115
11)	UNE SYMETRIE	107,109,104,107
12)	AGRANDIR	101,97,93,92
13)	DIMINUER	89,97,92,88
14)	DEPLACER	92,104,98,89
15)	INCLINER	75,69,77,70

La méthode de réduction adoptée est l'analyse factorielle des correspondances /21/ , /27/ , /28/
Les figures 1 à 3 montrent l'allure des trajectoires obtenues avec et sans filtrage , visualisées sur console graphique numérique .
On peut remarquer que les trajectoires brutes sont compliquées,difficiles à retenir. Il convient donc de les simplifier.
Un certain nombre de méthodes ont été appliquées /30/ , la plupart cherchant à minimiser la distance entre la trajectoire et la trajectoire finale dont on fixe le nombre de points,mais sans compter la difficulté de définir une distance entre trajectoires qui tiennent compte des sens de parcours,des concavités,des rebroussements.
Les méthodes suivantes ont été essayées :

- trajectoires simplifiées en prenant 1 point sur 2 , 1 point sur 3 , ...
- trajectoires lissées par des enveloppes et prise en compte de l'enveloppe moyenne.
- trajectoires lissées par approximation à l'aide d'une norme.
- trajectoires simplifiées par suppression des redondances de points (normalisation d'une trajectoire par rapport au temps)
- trajectoires simplifiées après segmentation
- squelettisation

La figure 3 montre l'allure des trajectoires simplifiées
Numériquement , nous nous sommes donc contentés d'utiliser des méthodes très simples basées sur des moyennages et des détections de zones stables à l'aide de seuils;ces méthodes ont malheureusement échouées sur certains phonèmes tels que 'i' et 'u' qui en général sont trop courts pour être détectés;Les'l' et les 'r' dont les classes sont trop diffuses ne sont pas non plus détectés du fait que deux points consécutifs sont en général trop éloignés.

II- 2 STABILITE DE L'APPRENTISSAGE ET DE LA RECONNAISSANCE

On rappelle qu'il est possible de calculer les facteurs de variables nouvelles sans refaire une analyse /21/. Dans le cadre de cette étude , ce calcul permet de placer sur les graphiques tous les éléments nouveaux lors d'une reconnaissance; la place que l'on assigne à un vecteur échantillon i n'est autre que celle qu'il occuperait si on l'assignait au tableau soumis à l'analyse en lui donnant une masse évanescente (i.e. en conservant le profil de la ligne i , seul déterminant en analyse factorielle des correspondances).

L'utilisation pratique de la formule de projection $X_q(i) = \sum_{j=1}^{Card(j)} U_{jq} \frac{P_{ij}}{P_{i.}\sqrt{P_{.j}}}$

(voir au § II-3 pour l'explication des variables) demande que le système de vecteurs propres U_{jq} issu de la matrice de contingence associée aux données pour un même locuteur soit invariant,quelque soient ces données.

Une étude expérimentale a été réalisée en deux étapes :

- étude de la stabilité des $\{U_{jq}\}$ lors de l'apprentissage
- étude de la stabilité des formes lors de la reconnaissance

II-2-1 Stabilité de l'apprentissage

Cette étude a porté essentiellement sur la liste des quinze mots prononcés quatre fois par un même locuteur. On a extrait , pour ces quatre analyses ,quatre systèmes de vecteurs propres dont on a comparé les directions . Les résultats numériques ont été les suivants :

LISTE 1		LISTE 2		LISTE 3		LISTE 4	
V_1^1	V_2^1	V_1^2	V_2^2	V_1^3	V_2^3	V_1^4	V_2^4
•24971	-•70801	•26427	-•69002	•28780	-•71395	•29531	-•7160
•42084	-•23103	•37779	-•25592	•41745	-•18439	•36515	-•2020
•28834	•41413	•25103	•40328	•25459	•41641	•26807	•3561
•20074	•40863	•21287	•43156	•18856	•41255	•22112	•4463
•08613	•17616	•12990	•16741	•10184	•17681	•11548	•1852
•07342	•22234	•08769	•22820	•07196	•22893	•08857	•2512
-•09840	•09430	-•03445	•10741	-•06194	•09440	-•05577	•0960
-•12439	•05070	-•11730	•06821	-•13168	•04803	-•13322	•0523
-•19265	•03643	-•20215	•03117	-•20509	•01899	-•18707	•0190
-•21972	•02098	-•21932	•03420	-•23823	•01830	-•22293	•0141
-•29371	•02219	-•24724	•00880	-•29710	-•00193	-•25424	•0054
-•29987	-•04834	-•30665	-•04934	-•29773	-•05848	-•32456	-•0514
-•28710	-•06997	-•34194	-•06696	-•28681	-•07178	-•32788	-•0602
-•35285	-•05899	-•39268	-•04765	-•35360	-•06531	-•35992	-•0447
-•36963	-•04282	-•35266	-•04284	-•35043	-•06540	-•35038	-•0454

- Les vecteurs V_i ont 15 composantes (vecteurs propres de la matrice 15x15)
- La norme euclidienne de ces vecteurs vaut 1

On a calculé $Cos(\vec{V}_1^i , \vec{V}_1^j)$ pour tout $i \neq j$

$$Cos(\vec{V}_2^i , \vec{V}_2^j) \qquad \left\{ i=1,\ldots,4 \quad \text{et } j=1,\ldots,4 \right\}$$

Toutes les valeures calculées ont été comprises entre 0,98 et 1 , ce qui est assez satisfaisant pour conclure à l'invariance pratique des vecteurs propres pour l'analyse sur un même locuteur. De plus, les trajectoires correspondant aux mots identiques de chaque liste sont pratiquement identiques .

II-2-2 Stabilité de la reconnaissance /29/ , /30/

On considère maintenant une nouvelle liste de mots prononcée par le même locuteur.

1) ATTENDS	54	11) UNE MAISON	55
2) UN CAILLOU	63	12) UNE NOIX	49
3) LE CRAYON	67	13) GENTIL	58
4) LA POUPEE	65	14) LE FUTUR	69
5) L'ECOLE	55	15) CHOCOLAT	60
6) UN OISEAU	64	16) UNE CHAISE	61
7) DES GANTS	51	17) CONTENT	51
8) PEUREUX	40	18) FATIGUE	55
9) UN VELO	55	19) UNE VOITURE	68
10) UN AVION	67	20) UN CHAT	45

Il a été vérifié expérimentalement les deux points suivants :

a) Si l'on appelle $\vec{W}_i^k$ les vecteurs propres calculés dans cette analyse et $\vec{V}_i^k$ les vecteurs propres calculés précédemment , alors les calculs ont montré que $0,98 < Cos(\vec{W}_i^k , \vec{V}_i^k) < 1$

b) la projection des données relatives à la liste des quinze mots(liste prononcée quatre fois) dans les axes $\vec{W}_i^k$, (k=1,2 et i=1,...,15) fournit approximativement les mêmes facteurs , ce qui contribue à visualiser les mêmes trajectoires et à conclure que la reconnaissance est stable.

II - 3 SIMULATION ANALOGIQUE

II-3-1 Formalisation du problème

Le problème consiste à présenter en temps réel , sur écran cathodique , des trajectoires associées aux mots prononcés , devant un micro.
L'apprentissage a été fait au centre de calcul (en phase II-1) et une liste de 2 fois 15 paramètres est disponible à tout instant : ce sont les valeurs optimales des coefficients à afficher sur le bloc de réduction à l'aide de potentiomètres. Ce bloc de réduction est inspiré évidemment de la formule de projection des données supplémentaires :

$$X_q(i) = \sum_{j=1}^{Card(j)} U_{jq} \frac{P_{ij}}{P_{i.}\sqrt{P_{.j}}}$$

Il est bien évident que cette formule n'est pas directement applicable puisque $X_q(i)$ dépend de i , mais $P_{.j}$ peut être fixé comme valeur constante.
On donne les notations suivantes :

$K = \sum_{ij} T_{ij}$ $\qquad T_{ij}$ ($i^{ème}$ échantillon du $j^{ème}$ canal)

$P_{ij} = \frac{T_{ij}}{K}$ et $P_{i.} = \sum_{j} P_{ij}$ et $P_{.j} = \sum_{i} P_{ij}$ (i=1,..,Card(i),j=1,15)

peuvent être interprétés comme des lois marginales (ce sont des estimations de probabilités)
U_{jq} sont donc les paramètres de réductions

On prendra de plus $\frac{P_{ij}}{P_{i.}} = \frac{T_{ij}}{S}$, S étant fournie en sortie du vocoder grâce à un additionneur qui permet de réaliser la sommation de tous les canaux. Cette valeur est donc l'énergie totale de tous les canaux; elle permet en outre de réaliser une opération extrêmement importante : la normalisation en amplitude du signal acoustique.

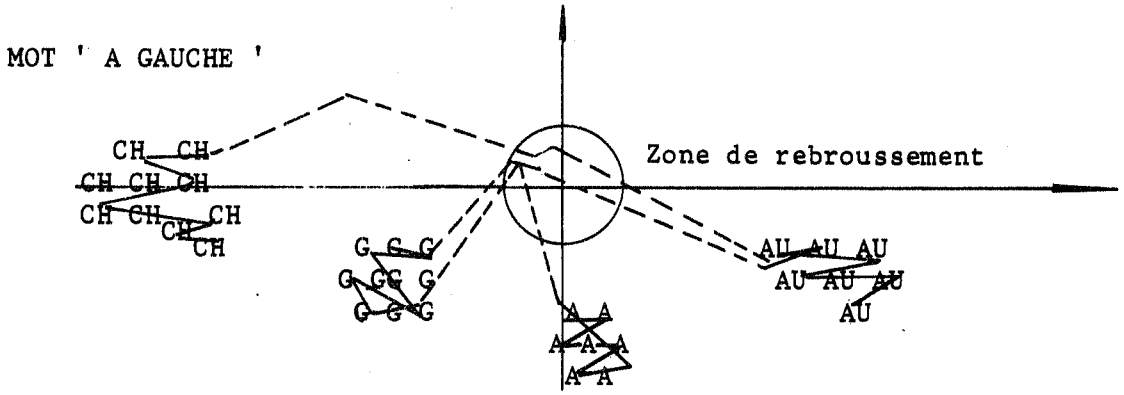
II-3-2 Visualisation et résultats

A tout instant sont disponibles $X_1(i)$ et $X_2(i)$ qui , appliqués aux bornes de l'oscilloscope permet d'observer le signal.
L'écran cathodique doit posséder une rémanence suffisante pour mémoriser (pendant 0,5 s environ) la trajectoire et permettre une perception globale suffisante.
Les essais réalisés en Février 1975 ont montré la mise en évidence très nette des trajectoires et les essais de stabilité (répétitions des mêmes mots) ont

été satisfaisants : les trajectoires sont pratiquement identiques pour un même locuteur.

Toutefois , un meilleur lissage de la géométrie des trajectoires a été jugé nécessaire ; pour cela un filtrage du second ordre est actuellement en cours de réalisation pour certaines sorties du vocoder (M. THAI).

D'autre part , un autre problème nous est apparu : les zones de transition dans chaque mot se traduisent par l'apparition de boucles parasites qui passent par les mêmes régions du plan sur l'écran cathodique . L'exemple ci dessous illustre ce cas :



Pour palier à cet inconvénient , nous mettons au point actuellement un circuit analogique dont le but est justement de supprimer cet effet : on se fixe une tension de référence V_o , voisine de celle que fournirait les zones de silence (qui,du fait des bruits de fonds et autres fluctuations , ne sont pas nuls) ; le signal acoustique, en sortie du vocoder , inférieur ou égal à ce seuil (opération réalisée par un comparateur) provoque un gel provisoire de ce signal pendant la période de transition et ainsi supprime les boucles parasites (ou du moins les atténue).

D'autre part , un effort de présentation est également en cours d'expérimentation : un certain nombre de signes distinctifs sont imposés au dessin de la trajectoire pour en faciliter la compréhension : on impose par exemple une brillance plus marquée sur les zones phonétiques afin de mieux distinguer les régions caractéristiques. A cet effet , un calque doté de toutes les régions phonétiques essentielles pourrait être superposé à l'écran.

III CONCLUSION

Ce dispositif devrait être d'une aide appréciable pour l'éducation , du fait que la représentation graphique est réalisée au niveau du mot (et non du son élémentaire) (procédé /27/ , /28/)

D'autre part , le plan d'observation est systématiquement adapté à l'optimum des données d'apprentissage.

Le coût de l'appareillage demeure très peu élevé puisque le recours à un calculateur n'est nécessaire que lors de l'apprentissage , et peut être réalisé une fois pour toutes pour une application donnée : tout apprentissage spécifique en vue d'une rééducation particulière demandera une base de données qui pourra être transmise au centre de calcul pour l'extraction des coefficients du plan de visualisation .

R E F E R E N C E S

/1/ R.K. POTTER, G.A. KOPP, H.C. GREEN : Visible speech
Van nostrand-Princeton - 1947

/2/ G.A. KOPP, H.G. KOPP : Visible speech for the deaf : An investigation to evaluate the usefulness of the visible speech cathode ray tube translator as a supplement to the oral method of teaching speech to deaf and severely deaf children
Speech and Hear - Report on Grant RD-526-Wayne State Univ.- DETROIT-1963

/3/ R.STARK, J. CULLEN, R. CHASE : preliminary work with the new Bell telephone visible speech translator
Amer. Ann. Deaf - Vol 113 - pp 205-214-1968

/4/ A. RISBERG : Visual aids for speech correction
Amer. Ann. Deaf. - Vol 113 - pp 173-194-1968

/5/ J.O'NEILL : Contributions of the visual componants of oral symbols speech comprehension
J. Speech Hear.disorders - Vol 19 - N°4-pp 429-439-1954

/6/ R. JAKOBSON,G. FANT,M. MALLE : Preliminaries to speech analysis
MIT Press- CAMBRIDGE- 1963

/7/ H. UPTON : Wearable eyeglass speech reading aid
Amer.Ann.Deaf.- Vol 113 - pp 222-229 -1968

/8/ H. TRAUNMÜLLER : A visual lipreading aid
Speech comm.Sem.- STOCKHOLM - 1974

/9/ J. M. PICKETT, A. CONSTAM : A visual speech trainer with simplified indication of vowuel spectrum
Amer.Ann. Deaf.- Vol 113 - pp 253-258 - 1968

/10/ I. B. THOMAS, R.C. SNELL : A rticulation training through visual speech patterns
Volta Rev.- Vol 72 - pp 310-318-1970

/11/ A. J. GOLDBERG : A visual feature indicator for the severely hard of hearing. IEEE Trans. on AU - Vol AU-20- N°1-pp 16-23-1972

/12/ R. A. NICKERSON, K. N. STEVENS : Teaching speech to the deaf : Can a computer help?
IEEE Trans. on AU- Vol AU-21-N°5- pp 445-455- 1973

/13/ R.G. CRICHTON, F.FALLSIDE : The developement of a deaf speech training aid using linear prediction analysis
Speech communication seminar- STOCKHOLM- 1974

/14/ J. L. FLANAGAN : Speech analysis, synthesis and perception
Springer-Verlag (2nd ed.)- NEW YORK- 1972

/15/ A. MAISSIS : Le traitement de l'information acoustique, étape fondamentale de la reconnaissance automatique de la parole
Thèse de docteur ès sciences- Univ.PARIS VI-1973

/19/ C. GUEGUEN, G. CARAYANNIS : Analyse de la parole par filtrage optimal de Kalman
Automatisme (Dunod ed) - Tome 18 - N°3 - 1973

/20/ L. LEBART, J. P. FENELON : Statistique et informatique appliquées
Dunod - PARIS - 1971

/21/J. P. BENZECRI et Col. : L'analyse de données - Tome 2 : L'analyse factorielle des correspondances
Dunod - PARIS - 1973

/22/ L. F. PAU : Méthodes statistiques de réduction et de reconnaissance des formes. Normalisation des paramètres phonétiques. Application à la reconnais sance de la parole.
Thèse de docteur-ingénieur- Université de PARIS-SUD-ORSAY-Mai 1972

/23/ W. PRONOVOST : Developments in visual displays of speech information
Volta Rev. - Vol 69 - pp 365-373-1967

/24/ J. M. PICKETT : Recent research on speech-analyzing aids for the deaf
IEEE Trans. on AU - Vol AU-16 - pp 227-234 - 1968

/25/ H.LEVITT : Speech processing aids for the deaf
IEEE trans. on AU - Vol AU-20 - pp 23-28-1972

/26/ J. M. PICKETT : Status of speech - analyzing communication aids for the deaf
IEEE Trans. on AU - Vol AU-20 - N°1 - pp 3-8 - 1972

/27/ L. F. PAU Statistical reduction and recognition of speech patterns,Conference on machine perception of patterns and pictures, Conf. publ. N°13,IEE Londres, Avril 1972 ,126-133; Brevet français 72-22958

/28/ C. GUEGUEN,L. F. PAU, A. MAISSIS : Communication homme - machine à support vocal , Echo des recherches , octobre 1972, 35-49

/29/ T. PIERRAT : Aide aux sourds - E.N.S.T PARIS - Octobre 1973 - Juin 1974

/30/ M. LAGNEAU : Application d'une méthode statistique à la reconnaissance d'un vocabulaire réduit - Mémoire Ingénieur C.N.A.M-16 Septembre 1974

FIGURE 1 : TRAJECTOIRE BRUTE
DU MOT ' ATTENDS ' /29/

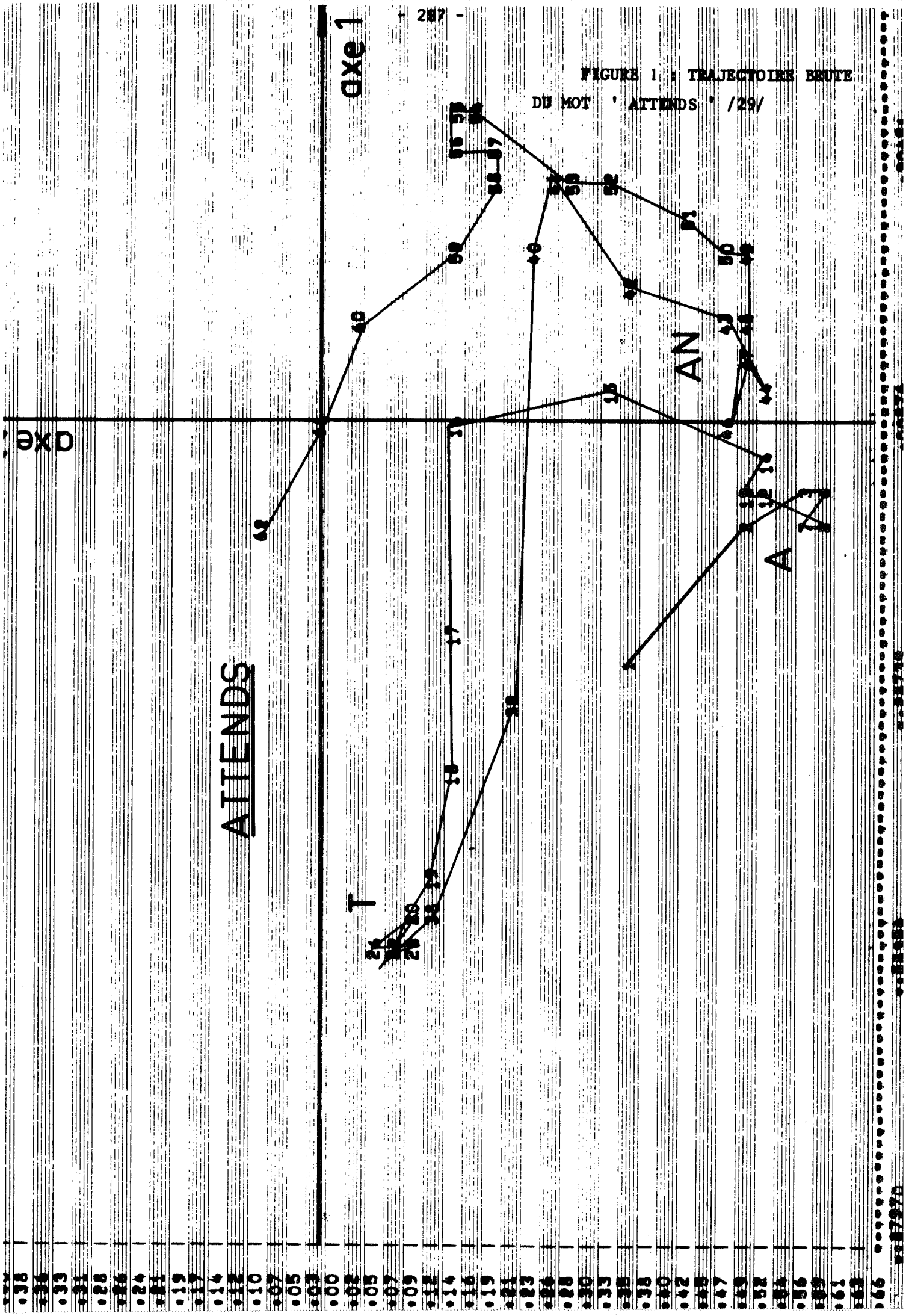
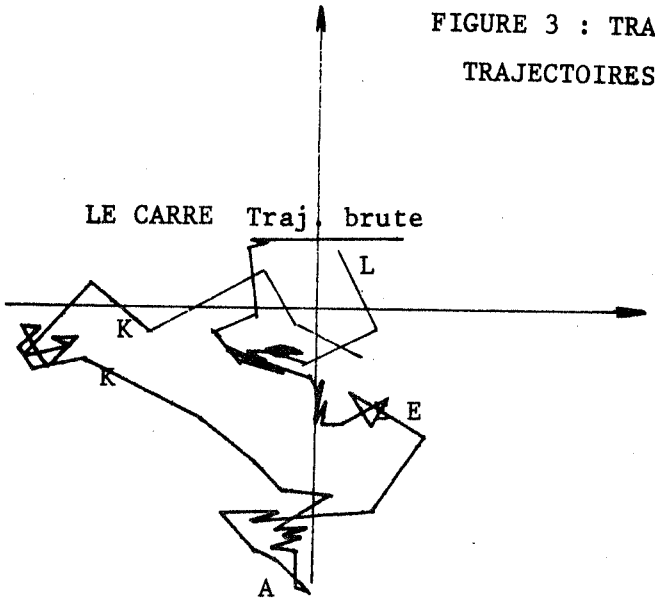
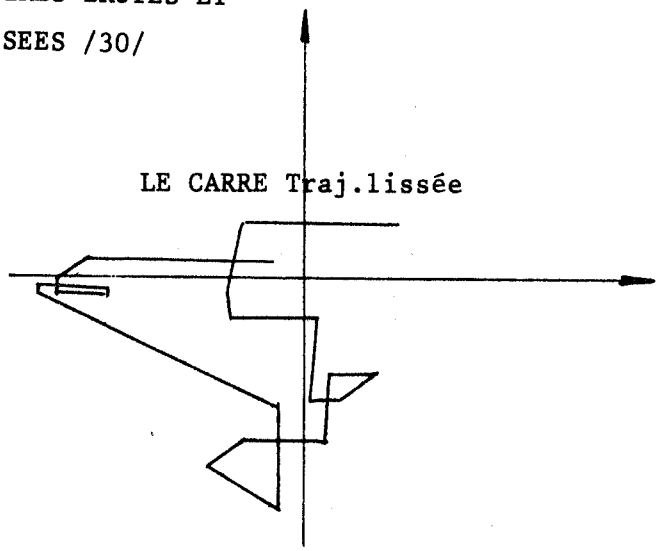
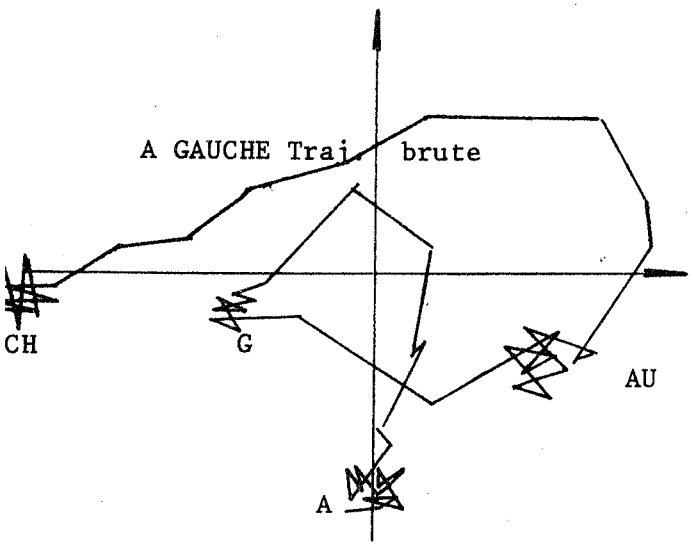


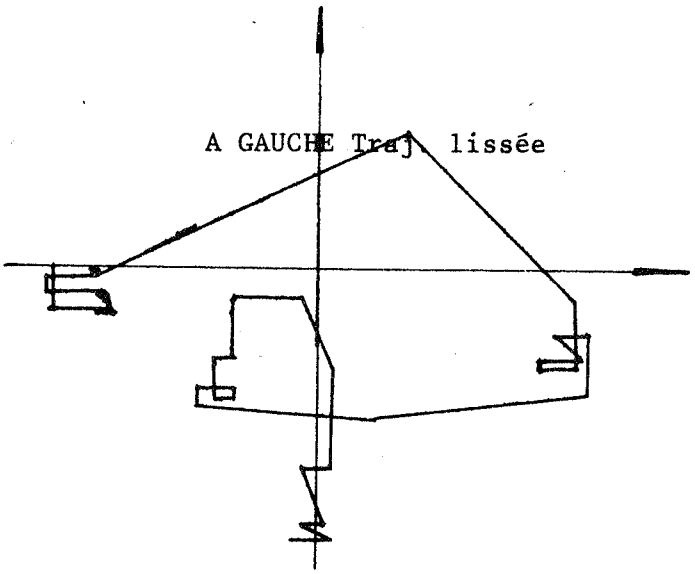
FIGURE 2 : TRAJECTOIRE BRUTE
DU MOT ' CHOCOLAT ' /29/

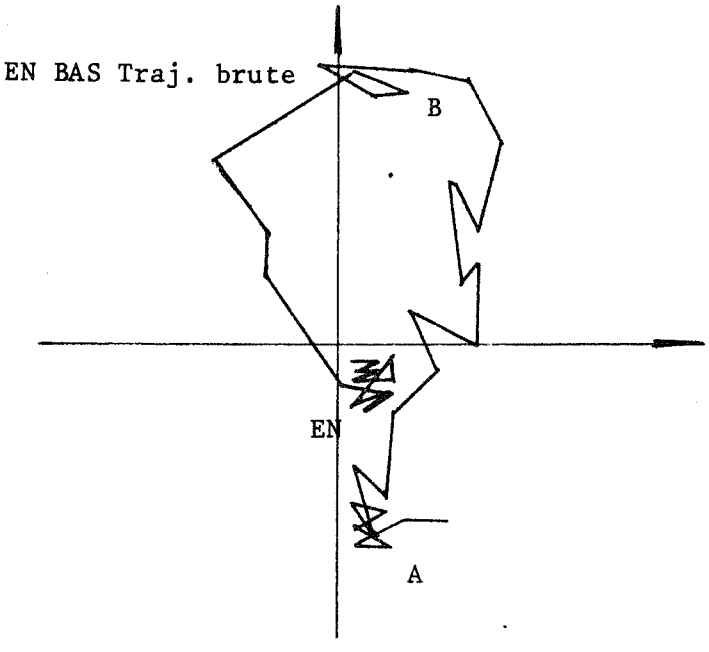
FIGURE 3 : TRAJECTOIRES BRUTES ET TRAJECTOIRES LISSEES /30/

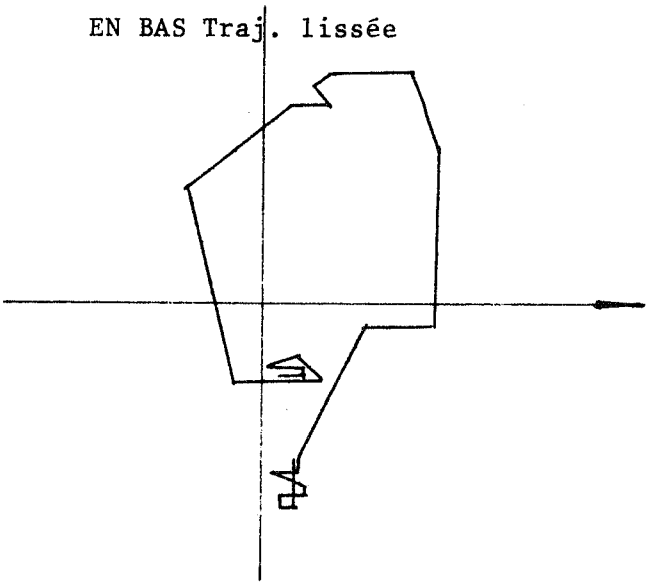












6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

SIRENE : un projet de système interactif pour la rééducation vocale des enfants non-entendants

Jean-Paul HATON, Marie-Christine HATON et Michel LAMOTTE
Laboratoire d'Electricité et d'Automatique
Université de Nancy I
C.O. 140 - 54 037 NANCY Cedex

RESUME

Le projet SIRENE, commencé depuis plus d'un an, consiste en la conception d'un système conversationnel sur mini-calculateur pour la rééducation vocale des déficients auditifs. Le système comprend essentiellement un analyseur spectral réglable, un détecteur de fondamental, une console de visualisation graphique et des programmes de visualisation, de conversation avec l'élève et de rééducation, utilisant les méthodes de la reconnaissance automatique de la parole. Les premiers résultats sont exposés et l'extension de ces recherches est discutée.

SUMMARY

The SIRENE project is oriented toward the design of an interactive speech training system for the deaf using a minicomputer. This system is essentially made up of a spectral analyzer, a pitch detector, a CRT display and programs of visualization, conversation and reeducation with the use of automatic speech recognition techniques. Preliminary results are given and the extension of this work is discussed.

SIRENE : un projet de système interactif pour la rééducation vocale des enfants non-entendants

Jean-Paul HATON, Marie-Christine HATON et Michel LAMOTTE
Laboratoire d'Electricité et d'Automatique
Université de Nancy I
C.O. 140 - 54 037 NANCY Cedex

1) INTRODUCTION.

L'enseignement assisté par ordinateur fait partie des applications intéressantes de l'informatique. Dans ce domaine, la rééducation vocale des déficients auditifs occupe une place encore modeste mais destinée sans doute à s'accroître. Nous présentons ici les études en cours et les premiers résultats obtenus dans le cadre du projet SIRENE destiné à concevoir, sur un minicalculateur, un système interactif pour la rééducation des enfants non-entendants, sourds de naissance ou du moins depuis leur plus jeune âge (*). Les conséquences de la surdité chez ces sujets sont très graves car ils acquièrent très difficilement une connaissance suffisante de leur langue maternelle. Etant, de plus, privés du contrôle des sons qu'ils émettent, ils n'arrivent à communiquer oralement que de façon très insuffisante et avec un retard important.

L'aide à apporter à ces enfants peut se situer au niveau de la perception d'un message parlé ou au niveau de l'émission de la parole. C'est à ce second aspect du problème (rééducation de la voix) que nous nous intéressons. L'idée de base consiste à utiliser les sens de la vision, ou du toucher, pour remplacer la contre-réaction auditive qui fait défaut dans le cas où l'ouïe résiduelle est inutilisable. Un tel système doit permettre au rééducateur de faire comprendre à l'élève les mouvements articulatoires qu'il doit effectuer et les fautes qu'il a commises. Dans une étape ultérieure, on peut envisager un système entièrement automatique permettant un entraînement intensif et autonome de l'élève, sans pour autant diminuer l'importance du rôle du rééducateur.

(*) Ce projet, qui a débuté en 1973, est financé en partie avec l'aide de l'Institut National de la Santé et de la Recherche Médicale (I.N.S.E.R.M.).

L'utilisation d'un ordinateur dans la conception d'un tel système présente, mis à part le prix de revient, de nombreux avantages. Elle permet d'introduire une modularité nécessaire si l'on veut disposer d'un appareillage évolutif et non pas figé. Ceci permet aussi d'avoir une grande souplesse et une grande facilité d'emploi, condition sine qua non de succès. De plus, seul l'ordinateur permet le stockage de données nécessaires pour effectuer des comparaisons, des études d'évolution dans le temps des sujets, ou des tests statistiques sur une population d'élèves. Enfin, l'utilisation de la reconnaissance automatique de la voix permet de juger objectivement la qualité du son émis par le sujet, sans risque d'accoutumance du maître à la voix de son élève.

Sans avoir mené de façon exhaustive une étude complète des défauts des voix de sourds, nous avons dressé, pour une première étape, une liste des fautes les plus courantes qui seraient susceptibles d'être corrigées : niveau sonore d'émission, hauteur de la voix, mélodie, rythme et respiration, confusion de phonèmes, prononciation de mots, etc...

Il est nécessaire d'envisager des algorithmes d'entraînement à la prononciation de mots ou de courtes phrases, car il est bien évident que la parole ne s'apprend pas seulement au niveau phonémique.

Un aspect surement important, et mal connu à cause du manque d'expérience, est l'aspect psychologique des relations entre l'élève et un système automatique. Bien du travail doit être fait dans ce domaine, si l'on ne veut pas avoir de gros déboires. De plus, le but du système ne doit pas être une copie pure et simple du maître par l'élève, mais plutôt une prise de conscience par l'enfant de sa propre voix et de toutes ses possibilités.

A côté de ces importants facteurs, la conception de notre système doit obéir à un certain nombre de contraintes de fonctionnement, principalement :

- nécessité de travailler en temps réel, pour que l'élève ait une réponse immédiate lors d'une émission,

- facilité d'emploi pour ne pas polariser l'élève (ou l'éducateur) sur cet aspect du système,

- conception d'images à présenter à l'élève suffisamment simples, mais pertinentes, pour ne pas risquer de dépasser sa capacité d'assimilation, avec une gradation depuis la forme la plus élémentaire (oui/non) jusqu'à une présentation plus élaborée,

- présentation attrayante de ces images, de façon à ne pas rebuter l'enfant et à obtenir sa participation effective. La notion de jeu avec le système est ici primordiale.

Historiquement, les chercheurs ont pensé très tôt à apporter des aides visuelles ou tactiles à la rééducation vocale des sourds. Depuis l'introduction du "Visible Speech Translator" [1] , de nombreux appareils ont été proposés, en particulier par l'équipe du Pr. FANT à Stockholm : N indicator, S indicator, visualisateur de spectres LUCIA, etc... [2] . Mais ces dispositifs, le plus souvent analogiques, n'envisageaient qu'un aspect limité du problème. Le seul système informatique proposé dans ce domaine est celui de STEVENS [3] . Ce système est conçu pour la langue anglaise et n'utilise pas l'apport de la reconnaissance automatique de la parole, qui peut être, à notre avis, très important.

En ce qui concerne la langue française, peu de systèmes ont été proposés jusqu'à présent, mais l'intérêt porté à ces recherches va sans doute augmenter rapidement.

2) DESCRIPTION DU SYSTEME ACTUEL.

Le système actuellement développé dans le cadre du projet SIRENE est centré sur un minicalculateur NOVA 2 de DATA GENERAL. Il comprend trois parties principales :

- une série de capteurs,
- des moyens de présentation de l'information à l'élève,
- des programmes de paramétrisation, de rééducation, de visualisation et de dialogue avec l'élève.

Le fait de concevoir un système informatique présente de nombreux avantages, comme on l'a déjà dit, tant du point de vue de l'utilisation que de l'évolution du système. En fait, l'état actuel du système est surement provisoire ; il est destiné à évoluer au fur et à mesure de l'avancement du projet et des tests que nous effectuons. La Fig. 1 montre schématiquement la disposition des différents éléments du système.

L'étage électronique de prétraitement du signal vocal comprend :

- un analyseur spectral à 32 filtres couvrant la gamme de fréquences 100Hz - 7000Hz, à fréquence d'échantillonnage réglable de 50Hz à 150Hz,
- un détecteur de mélodie "Mélographe" du CNET / ETA,
- un convertisseur A/D permettant de numériser de la parole à la fréquence voulue.

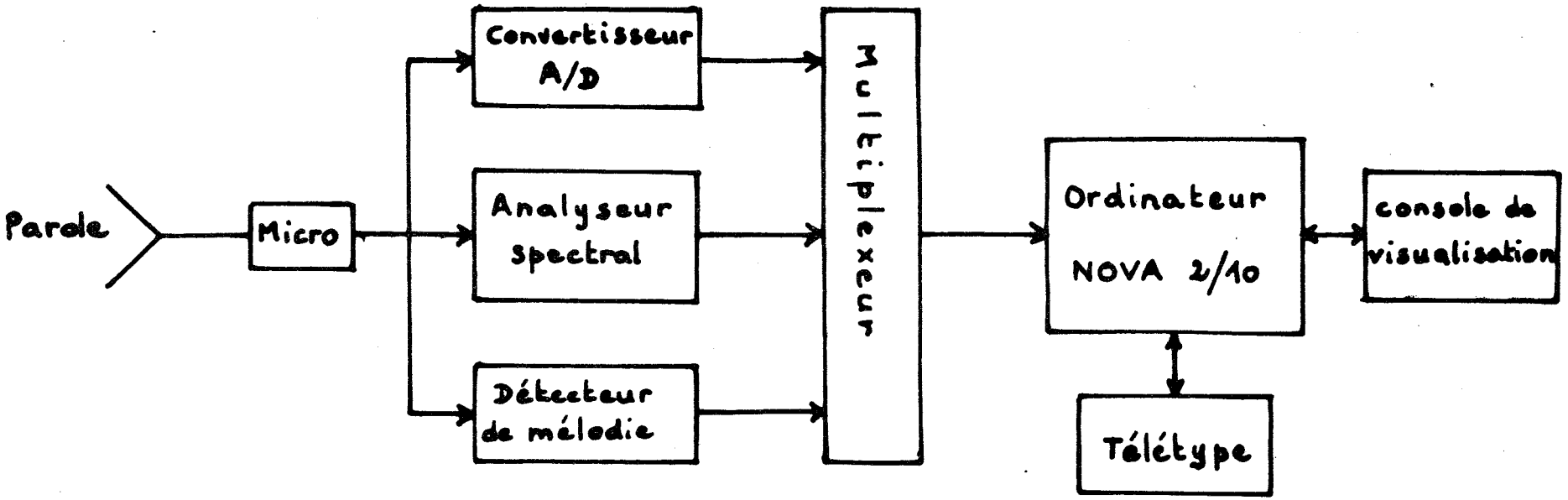


Schéma général du système

Fig. 1

Les données de ces différents appareils sont fournies en temps réel au calculateur.

Actuellement, les paramètres de la voix sont visualisés sur une console graphique TEKTRONIX 4012 qui sert de lien entre l'élève et le système. En particulier, la sélection des différents programmes se fait en conversationnel à partir du clavier de la console. D'autres moyens de présentation des paramètres seront éventuellement étudiés à l'avenir.

3) VISUALISATIONS ET PROCEDURES DE REEDUCATION.

Les aides que l'on peut apporter à l'enfant non-entendant pour améliorer son attitude communicative appartiennent à deux catégories. Les unes, aides à la perception, doivent permettre à l'enfant de mieux comprendre ce qui a été prononcé par son interlocuteur. Les autres, aides à la phonation, doivent, grâce à un programme de rééducation adaptée à l'enfant, lui permettre d'acquérir la maîtrise des différents organes qui interviennent dans la production de la parole. Pour que ces deux types d'aides soient complets et efficaces, il est nécessaire d'extraire et de présenter au sujet un certain nombre de paramètres pertinents de l'onde de parole ; en conséquence les aides à la perception et à la phonation peuvent se recouvrir plus ou moins. Cependant c'est à la deuxième catégorie que nous nous intéressons ici.

a) Intensité sonore et mélodie :

Le niveau sonore global, le fondamental (sa position et ses variations) sont des paramètres importants de l'émission vocale, souvent mal rendus par les sourds. Les deux phénomènes sont liés chez les sourds, où l'on constate souvent une montée du pitch lorsque le sujet veut augmenter l'intensité sonore émise. C'est pourquoi la rééducation du niveau sonore a lieu à 3 stades différents :

- au niveau élémentaire : on cherche simplement à donner conscience à l'enfant de l'intensité qu'il émet lors d'un son soutenu, avec seulement deux indications : trop fort, trop faible. Il s'agit d'un jeu dans lequel un mobile doit se déplacer entre 2 droites horizontales représentant les extrema moyens d'intensité. Le déplacement vertical est asservi à l'intensité sonore et le déplacement horizontal est une fonction linéaire du temps ;

- au niveau du mot : on visualise une forme globale qui rend compte de la variation d'intensité au cours de l'élocution (cf. par. 4-d) ;

- en liaison avec le pitch : on demande à l'élève d'augmenter la fréquence du fondamental tout en conservant un niveau sonore raisonnable. La fig. 2 schématise une configuration obtenue dans ce jeu. Le déplacement vertical du mobile est asservi à la hauteur de la voix et le chemin qui mène au but est limité par les fossés "trop faible" et "trop fort" (déplacement horizontal lié à l'intensité).

D'autres programmes sont destinés à la rééducation spécifique de la mélodie. Ils permettent à l'enfant de trouver la position fonctionnelle optimale de son fondamental (place de sa voix), d'apprendre à tenir un son avec un fondamental fixe et de commander les variations de mélodie (la fig. 3 montre un exemple de jeu possible, dans lequel le déplacement horizontal du spot est linéaire dans le temps et le déplacement vertical lié au fondamental).

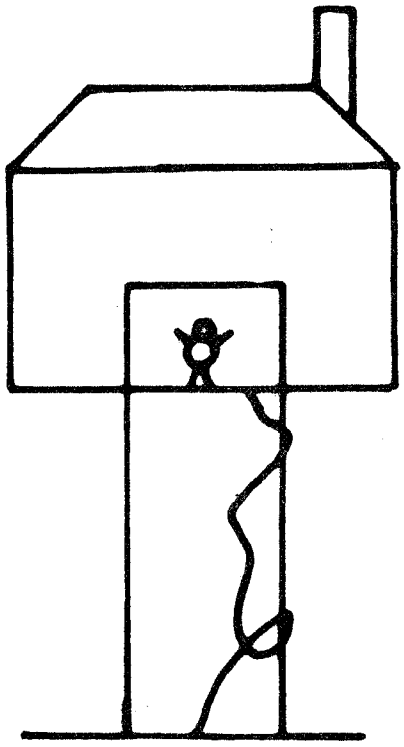


Fig. 2

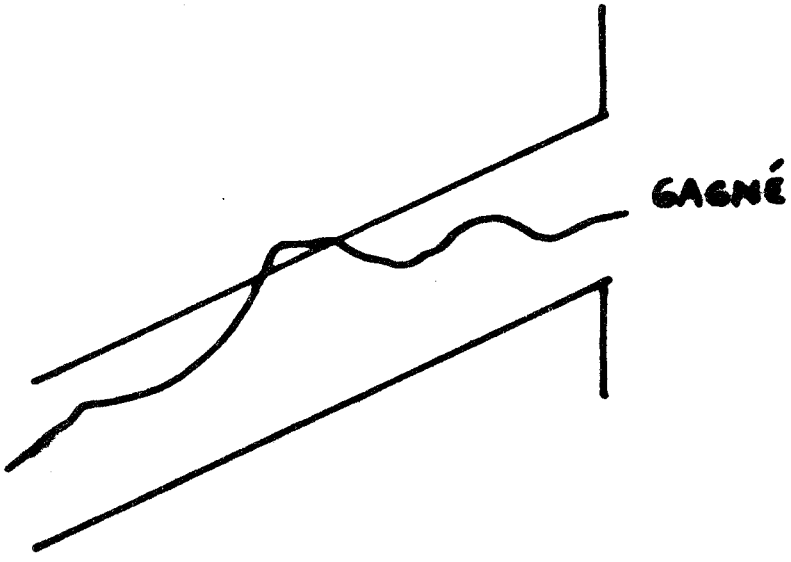


Fig. 3

b) Rythme :

Les erreurs de rythme et de pauses de respiration sont très courantes chez les sourds. L'entraînement à la respiration est prévu lors de la rééducation sur mots ou courtes phrases (on indique la position moyenne des pauses). L'entraînement au rythme utilise un programme de segmentation, à partir de la prononciation de groupes C-V. La segmentation peut être effectuée :

- en considérant les variations d'énergie du signal,
- en utilisant les indications d'une fonction de voisement, quand c'est possible,
- en visualisant les traces correspondantes dans un plan optimal obtenu par une méthode de Karhunen-Lœve [4] . Cette transformation fournit une bonne dispersion des différents phonèmes, elle est donc particulièrement intéressante pour l'étude du rythme, tant dans la syllabe que dans le mot (cf. fig. 4 pour la syllabe "cha").

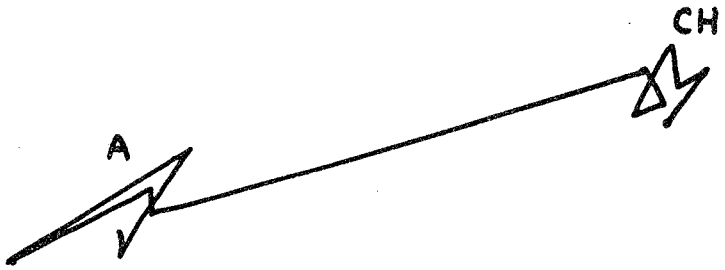


Fig. 4

c) Phonèmes :

La rééducation vocale des sourds passe par l'apprentissage des sons élémentaires du langage : voyelles, puis consonnes. Le sonagramme est certes utile mais le plus souvent trop chargé, difficile à interprêter et pas lié directement aux phénomènes de phonation. Un programme permet de simplifier les données de l'analyse spectrale de façon à visualiser en temps réel un spectre dépouillé avec seulement quelques niveaux de luminance. Ce spectre peut servir à entraîner l'élève à prononcer un son soutenu, mais sa complexité en diminue l'intérêt. L'analyse spectrale est par contre fondamentale dans l'interprétation du son émis par l'élève et dans sa comparaison automatique avec le son demandé.

L'entraînement à la prononciation des voyelles est assuré par une série de programmes de visualisations, de jeux et de conseils à l'élève :

- visualisation d'une forme attrayante obtenue à partir du résultat d'un filtrage inverse [5] donnant les trois premiers formants. La fig. 5 donne un exemple de telle forme ;

- visualisation simultanée d'une forme représentant la fonction d'aire du conduit vocal dans l'approximation d'un tube à sections variables (cf. fig. 5). Cette image tente de créer un lien grossier entre les positions articulatoires et la voyelle demandée (ouverture de la bouche, ...);

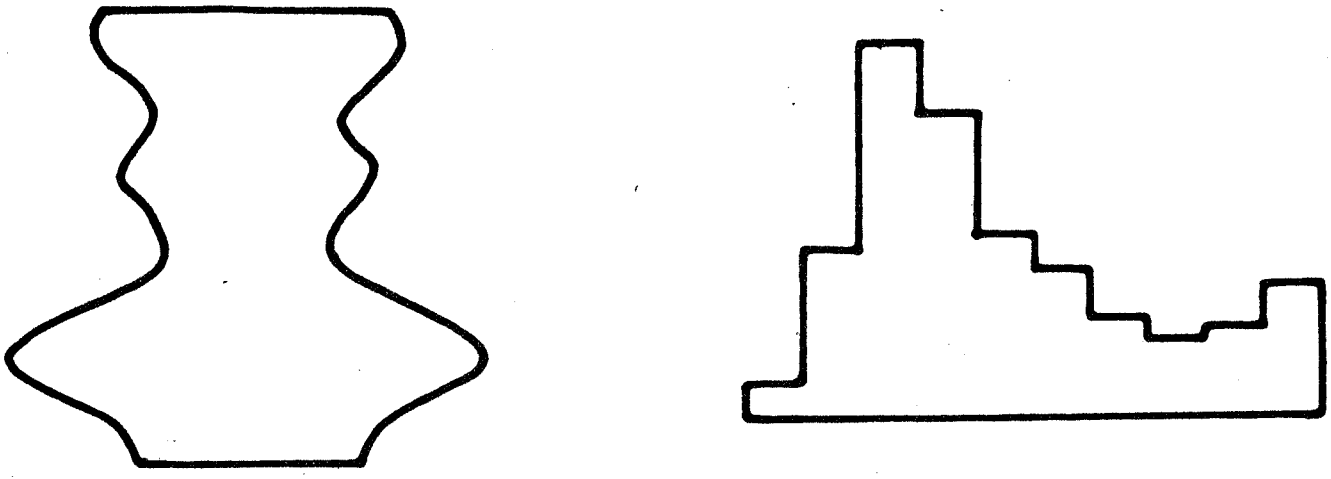


Fig. 5

- jeu de cible en visualisant par un point du plan F1-F2 le son émis par l'enfant, et en matérialisant une cible représentant la zone où doit se trouver le point. Le jeu est rendu plus attractif par l'introduction d'un score (essais répétitifs), permettant éventuellement une émulation entre les enfants.

Alternativement, un mauvais essai peut donner lieu à une interprétation par reconnaissance automatique et à l'émission d'un message destiné à corriger l'élève.

La visualisation peut être effectuée aussi dans des plans optimaux (cf. fig. 4). Ceci permet d'utiliser la procédure aussi bien pour des voyelles que pour des consonnes soutenues (fricatives en particulier). La rééducation des consonnes est effectuée par des jeux permettant de faire la distinction entre consonnes voisées et non voisées, entre /s/ et /ʃ/, etc...

d) Mots et phrases :

Cette nouvelle étape de la rééducation ne peut être envisagée qu'après un solide apprentissage aux niveaux inférieurs. L'apprentissage des mots est directement lié au rythme, indispensable à une bonne compréhension. La rééducation du rythme peut être effectuée par visualisation d'une forme globale donnant la variation d'intensité au cours de l'élocution du mot (fig. 6). Cette visualisation est intéressante car elle fournit simultanément des indications sur :

- le rythme du mot ou de la phrase et les pauses éventuelles de respiration,
- le voisement (indiqué par des hachures),
- le niveau sonore à tout instant.

La visualisation des traces dans des plans optimaux a également été proposée pour l'aide à la perception de mots [6]. Nous utilisons une telle visualisation (analogue à celle de la fig. 4 correspondant à une syllabe) pour la rééducation du rythme dans le mot plutôt que pour l'apprentissage du mot lui-même. En effet, la relation entre la forme visualisée et l'aspect acousto-phonétique du mot est très vague, par contre la segmentation du mot en unités élémentaires est bien apparente.

On peut également compléter l'apprentissage d'un vocabulaire restreint par un jeu spécifique à ce vocabulaire. La fig. 7 montre le déroulement d'un jeu de labyrinthe permettant d'utiliser le vocabulaire de quatre mots : GAUCHE, DROITE, AVANT, ARRIERE, qui commandent le mouvement du mobile. Ces jeux utilisent un programme de reconnaissance automatique pour interpréter la commande donnée par l'enfant.

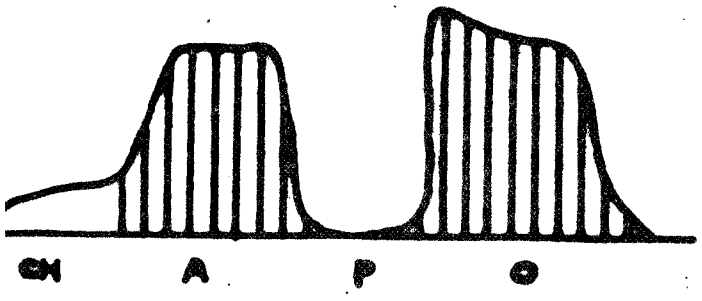


Fig. 6

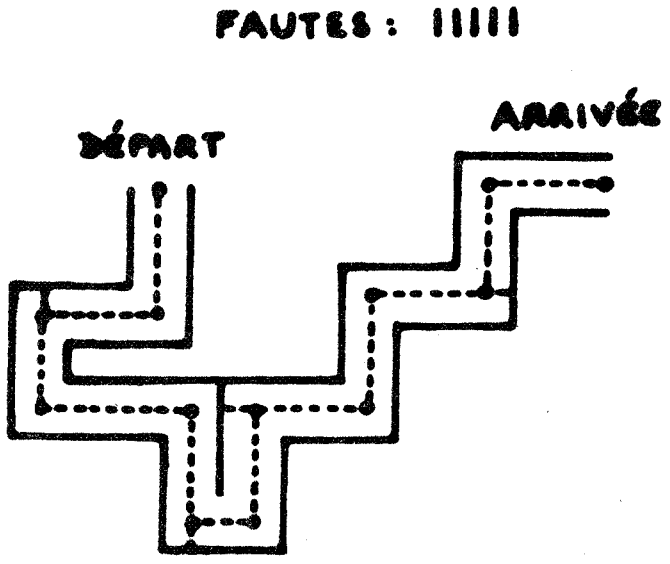


Fig. 7

4) DEVELOPPEMENT FUTUR.

Nous avons exposé brièvement les principes de base qui ont guidé notre approche du problème de la rééducation des enfants non-entendants. Ce travail ne fait que débuter, un certain nombre de programmes fonctionnent, d'autres sont en cours de mise au point. Il reste maintenant à tester systématiquement ces programmes sur un nombre aussi grand que possible d'enfants. Il s'agit là d'un travail de longue haleine qui permettra de modifier au mieux les modules du système, voire d'en remplacer certains. Ce travail sera mené de front avec une tentative de formalisation des déficiences caractéristiques des voix de sourds, en vue d'une automatisation plus aisée de la rééducation. L'étude sur l'analyse polynomiale de la mélodie s'inscrit déjà dans cette voie [5]. Par ailleurs, d'autres domaines tels que l'apprentissage d'une langue étrangère, pourraient utiliser des systèmes analogues.

Le rôle de la reconnaissance automatique de la parole doit être également précisé, de façon à tirer le meilleur parti de ses possibilités. Les techniques de reconnaissance peuvent servir à juger objectivement les sons émis par l'élève, comme nous l'avons déjà dit, et par suite à tenter de corriger les défauts, mais nous envisageons aussi des utilisations différentes. En particulier, elles peuvent servir à définir des tests de progrès, très utiles pour étudier l'évolution d'un élève, et leur rôle est important dans l'optique d'un système à auto-interprêtation du comportement de l'élève en vue d'un entraînement sans moniteur.

BIBLIOGRAPHIE.

[1] R.K. POTTER et al., Visible Speech, D.Van Nostrand Co., Inc., New-York, 1947.

[2] cf. de nombreux "Quarterly Progress Report" Speech Transmission Laboratory, KTH, Stockholm.

[3] R.A. NICKERSON, K.N. STEVENS "Teaching speech to the deaf : Can a computer help?" IEEE Tr. A.U., 21, pp.445-445, Oct.1973.

[4] J.P. HATON, M. LAMOTTE "Selection of features by information compression in speech recognition" Proc. Int. Conf. on Machine perception of patterns and pictures -Teddington-G.B.-April 1972, pp. 134-140.

[5] M.C. HATON "Analyse par filtrage inverse de la parole" L.E.A. Groupe Communication Homme-Machine, rapport interne, Janvier 1975.

[6] C.J. GUEGUEN et A.H. MAISSIS "Un système d'aide aux sourds profonds" Xèmes Assises Nationales de la prothèse auditive, Paris, 5-8 Oct. 1972.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UTILISATION D'UN DIVISEUR DE FREQUENCES AUDIBLES EN EDUCATION ORTHOPHONIQUE

LAMOTTE M. et VIGNERON C.

RESUME

On décrit le fonctionnement d'un diviseur de fréquences audibles utilisé en orthophonie pour l'éducation des sourds présentant des restes auditifs dans les basses fréquences. Discussion des contraintes que doit présenter le signal vocal transformé : temps réel, respect des positions relatives des traits caractéristiques de la parole.

SUMMARY

Description of a frequency divisor device for learning deaf children wich have some perception in lower frequences. Some points concerning features of transformed speech signal : real-time and keeping of complete speech pattern.

UTILISATION D'UN DIVISEUR DE FREQUENCES AUDIBLES EN EDUCATION ORTHOPHONIQUE

LAMOTTE M. et VIGNERON C.

Un bon nombre de surdités profondes se caractérisent par des restes auditifs dans les basses fréquences. L'atténuation dans les hautes fréquences (à partir de 1000 à 2000Hz) est telle que même une forte amplification ne peut récupérer cette zone. C'est pourquoi divers procédés ont été proposés pour ramener les informations contenues dans le haut du spectre dans la gamme des fréquences basses audibles : translation du spectre, détection de plusieurs domaines fréquentiels et codage dans les basses fréquences.

Le dispositif que nous décrivons utilise un principe différent : la compression de fréquences.

PRINCIPE.

On peut considérer la parole comme une forme globale représentée graphiquement par un sonagramme. Si, par une division de fréquence sur la totalité du spectre, ce sonagramme est "comprimé vers les basses fréquences, l'allure de la forme n'est pas altérée : l'échelle des temps est respectée et toutes les fréquences sont divisées par un facteur constant.

Il en résulte que les positions relatives des différents traits caractéristiques de la parole (formants, bruits d'occlusion, etc...) restent inchangées.

En se basant sur ce fait, on peut espérer apporter, par cette compression spectrale, une aide appréciable aux malentendants qui possèdent des restes auditifs dans les graves. En effet, les relations qui lient les points d'articulation du canal vocal aux traits acoustiques d'un sonagramme sont les mêmes pour la représentation non comprimée et comprimée.

Le principe de l'opération est le suivant :

Dans un premier temps, une mémoire enregistre une fraction du signal vocal échantillonné au rythme normal de la parole (opération d'écriture). Ensuite, cette mémoire délivre le même signal mais à un rythme plus lent (opération de lecture). En fait, deux mémoires qui fonctionnent alternativement l'une en

écriture, l'autre en lecture, fournissent un signal continu.

Plus précisément, le signal vocal analogique est échantillonné et quantifié. Chaque bit des échantillons numériques est entré simultanément dans deux registres à décalage à entrée et sortie série. Le premier registre (registre d'écriture) est commandé à la fréquence h, le second (registre de lecture) à la fréquence h/k où k > 1 est le coefficient de division qui peut être entier ou fractionnaire. La sortie du registre de lecture se fait donc à un rythme plus lent et, après conversion numérique-analogique, constitue le signal utile étalé dans le temps par rapport au signal initial.

Si n est le nombre de bits de chaque registre, le second registre se vide en $\frac{nk}{h}$ secondes, soit un cycle de fonctionnement. Pendant ce temps, le premier registre accepte en entrée nk échantillons. Les n(k-1) premiers sont perdus ; seuls les n derniers restent en mémoire. A la fin de ce cycle, les commandes d'entrées (h et h/k) et les sorties de registres sont inversées : la sortie du premier registre, à rythme lent, donne le signal utile, tandis que l'entrée du second accepte les informations à rythme normal.

En résumé, le système prélève dans le signal initial des "bouffées" qui sont étirées dans le temps de façon à reconstituer un signal continu.

ALLURE DU SIGNAL DE SORTIE.

Nous venons de voir que sur les nk échantillons du signal d'entrée, seuls les n derniers servent à construire le signal de sortie qui peut présenter diverses allures.

Si, par exemple, une sinusoïde a une fréquence telle que les échantillons d'une période remplissent complètement la mémoire. Avec un facteur de division k = 2, on enregistre, durant la sortie à rythme lent des informations, deux périodes du signal d'entrée, dont seule la dernière est conservée. On constate de plus que le signal de sortie a l'allure d'une sinusoïde de période double. Pour ce genre de signal très redondant, il n'y a donc pas de perte d'information et dans l'utilisation de l'appareil on cherchera à se rapprocher de ce cas idéal (Fig. 1).

Si au contraire la fréquence du signal sinusoïdal est telle que le nombre des échantillons d'une période est supérieur ou inférieur à la capacité de la mémoire, la sortie se compose de morceaux de sinusoïde présentant une discontinuité à leurs raccords (Fig. 2). On compense ce défaut en réglant la fréquence d'échantillonnage sur celle du signal d'entrée.

Dans un signal vocal, le fondamental varie d'une personne à l'autre et à tout instant pour une même personne. On règle alors la fréquence d'échantillonnage pour enregistrer environ une période du fondamental et s'adapter ainsi au locuteur. Une amélioration importante consisterait à asservir la fréquence d'échantillonnage à la mélodie.

CONCLUSION.

Cet appareil expérimental a été conçu pour transmettre aux sourds qui possèdent des restes auditifs dans les graves, les informations essentielles contenues dans la parole tout en conservant à cette dernière un caractère aussi naturel que possible. En effet, les traits caractéristiques du signal vocal gardent leur intensité et leur position en valeur relative et, de ce fait, le sourd a plus de facilité pour faire correspondre la position de ses organes phonateurs et ce qu'il entend.

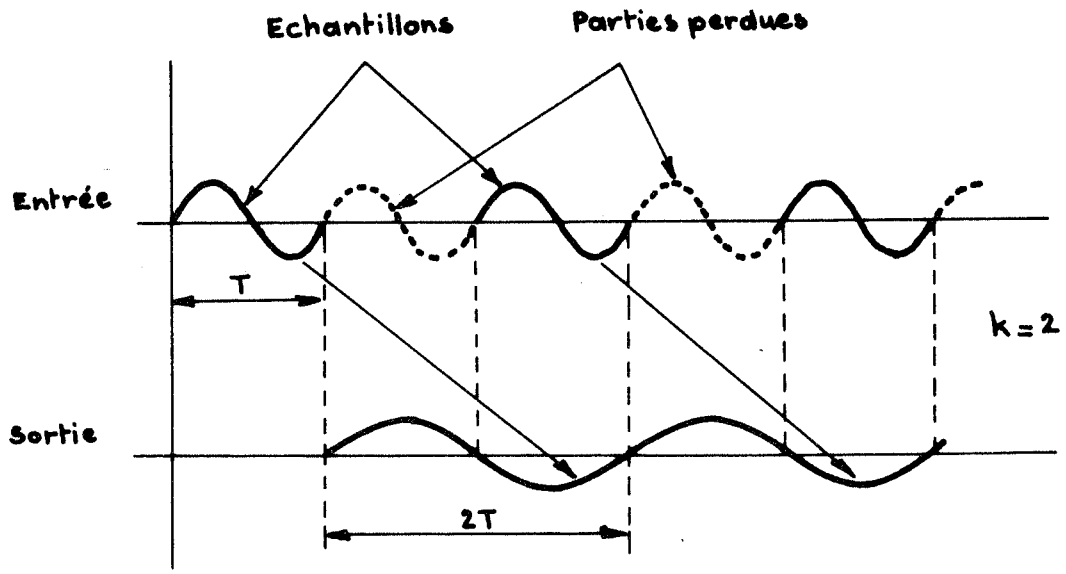


Fig 1

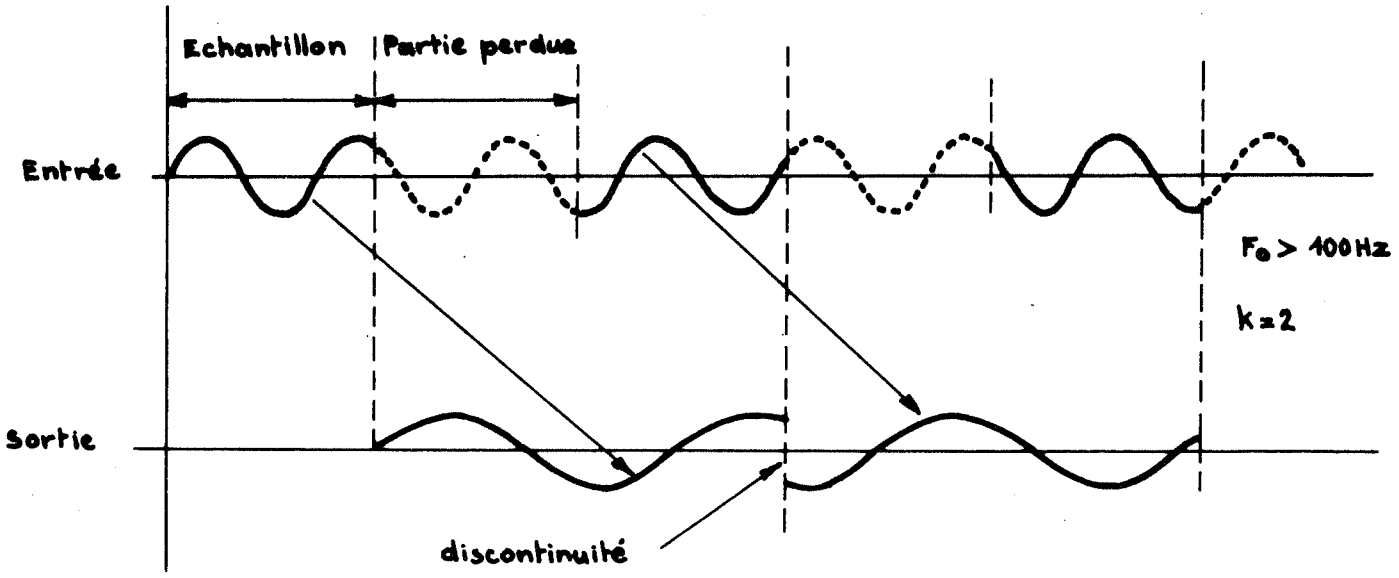

Echantillon
Partie perdue
Entrée
Sortie
$F_e > 100$ Hz
k = 2
discontinuité

Fig 2

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

ESSAI DE CARACTERISATION DES VOIX D'ENFANTS SOURDS PAR ANALYSE POLYNOMIALE DE LA MELODIE

Marie-Christine HATON - Jean-Paul HATON
Laboratoire d'Electricité et d'Automatique
Université de Nancy I
C.O. 140 - 54 037 NANCY Cedex

RESUME

Cet article présente une méthode d'analyse du contour mélodique permettant la caractérisation des voix par les coefficients des polynômes de Tchebycheff dans un développement polynomial. La méthode est appliquée à l'analyse d'un contour de phrase et à l'étude de l'évolution de la fréquence fondamentale lors de l'élocution d'une voyelle soutenue par un enfant sourd à différentes étapes de sa rééducation.

SUMMARY

This paper describes a method for the analysis of pitch contours using a Tchebycheff polynomial development. The method is to be used in a computer-aided speech training system for deaf children and makes it possible to classify the voices of the deaf. Some results are given, concerning the analysis of a sentence and the study of the evolution of pitch variations for a deaf child in reeducation.

ESSAI DE CARACTERISATION DES VOIX D'ENFANTS SOURDS PAR ANALYSE POLYNOMIALE DE LA MELODIE

Marie-Christine HATON - Jean-Paul HATON
Université de Nancy 1

1. INTRODUCTION

Toute tentative de rééducation vocale d'un enfant sourd doit nécessairement être précédée d'une analyse acoustique assez fine, permettant de déterminer en quoi les paramètres de la voix des sourds diffèrent des paramètres de la parole normale. Dans le cadre du système de rééducation des enfants non-entendants en cours de réalisation au Laboratoire (projet SIRENE), nous présentons ici une tentative de caractérisation et de classification des voix, fondée sur l'étude des variations temporelles du contour mélodique. La méthode utilise une analyse polynomiale dans une base de Tchebycheff, décrite au par. 3. Le détecteur de mélodie utilisé est le Mélographe CNET/ETA.

2. CARACTERISTIQUES DE LA VOIX DE MAL-ENTENDANTS

On retrouve souvent chez les enfants mal-entendants des défauts qui affectent l'intelligibilité et la qualité de la voix. Parmi les derniers, ceux qui concernent la mélodie sont de grande importance.

En effet, alors que les formes spectrales subissent l'influence du système de transmission que constituent le conduit vocal et le conduit nasal, et qu'elles dépendent de l'intensité, de l'impédance de radiation aux lèvres, de la distance du locuteur au microphone, etc..., la mélodie au contraire renseigne directement sur les organes vocaux et la façon dont ils sont utilisés.

La difficulté de positionner les cordes vocales et de commander la musculature au niveau du larynx font que le contour mélodique des enfants mal-entendants présente certaines caractéristiques. On a notamment observé les traits suivants :

- la mélodie peut être anormalement élevée (larynx élevé et trop contracté, on parle de "voix de fausset") ou au

contraire plus grave que la normale (larynx bas ou manque de tonicité qui entraîne une détente excessive) ;

- les variations de la fréquence fondamentale sont désordonnées et soumises surtout aux efforts fournis par l'enfant dans la production des sons ; alors que dans la parole normale, l'intonation dépend du type de phrases prononcées (affirmation, interrogative, ...), de la cadence de la phrase et de facteurs psychologiques.

Les erreurs de rythme, la fréquence anormalement grande des pauses de respiration, la contraction des sons individuels ou au contraire leur étirement, le manque d'expression sont autant de facteurs qui exercent un effet direct sur la forme du contour mélodique.

C'est à ces variations de la fréquence fondamentale au cours d'une séquence de parole voisée que nous nous intéressons ici.

3. METHODE D'ANALYSE

Analyser un contour (c'est-à-dire une courbe traduisant les variations dans le temps d'une grandeur, ici la mélodie) présente un certain nombre de difficultés dues aux différences importantes entre locuteurs, ainsi qu'aux différences pour une même phrase prononcée plusieurs fois par un même locuteur.

Pour un paramètre donné de l'analyse, au contraire, il est aisé de déterminer les différentes causes de variabilité. On retient ici comme paramètres les coefficients des polynômes de Tchebycheff dans une approximation polynomiale du contour étudié.

On se limite à une courte portion du contour sélectionnée par une fenêtre temporelle rectangulaire. Le contour est balayé par le déplacement de cette fenêtre dans le temps. Divers essais nous ont conduits à retenir une largeur de 50ms au minimum, 100ms au maximum. Le déplacement est pris égal à une demi-largeur de fenêtre.

On obtient de cette façon un ensemble de coefficients variant dans le temps avec autant de valeurs que de positions de la fenêtre temporelle.

Pour une fenêtre donnée, on prélève 2N+1 valeurs $f(x_k)$ du contour numérisé correspondant aux abscisses régulièrement espacées :

$$x_k = \frac{k}{N} \quad , \; k \text{ entier variant de } -N \text{ à } +N.$$

Il est possible de trouver un polynôme de degré 2N tel que, pour toute valeur de k :

$$P_{2N}(x_k) = f(x_k).$$

On réalise ainsi une interpolation par le polynôme :

$$P_{2N}(x) = \sum_{i=0}^{2N} a_i x^i.$$

Les coefficients a_i peuvent donc caractériser une portion du contour. Cependant, si l'on veut considérer une approximation de la fonction discrète $f(x_k)$, au sens d'un certain critère, le fait de négliger un monôme affecte tous les coefficients des autres monômes.

Pour éviter cela, on cherche à écrire le polynôme $P_{2N}(x)$ sous la forme :

$$P_{2N}(x) = \sum_{i=0}^{2N} b_i T_i(x)$$

où les $T_i(x)$ sont les polynômes de Tchebycheff.

Ces polynômes sont définis sur l'intervalle $[-1,+1]$ par la relation de récurrence :

$$T_{n+1}(x) = 2x\ T_n(x) - T_{n-1}(x)$$

et les conditions initiales :

$$T_o(x) = 1$$
$$T_1(x) = x.$$

Ils sont orthogonaux (mais non normés) avec la fonction de poids $w(x) = \dfrac{1}{\sqrt{1-x^2}}$.

Si l'on se limite aux m+1 premiers termes dans le développement de $P_{2N}(x)$, on a une approximation de la fonction au sens des moindres carrés, c'est-à-dire que la

quantité

$$\sum_{k=-N}^{+N} \varepsilon^2(x_k) \text{ est minimale,}$$

où $$\varepsilon(x) = f(x) - \sum_{i=0}^{m} b_i \, T_i(x).$$

Du fait de l'orthoganilité, les coefficients b_i de l'approximation ne sont pas affectés par la suppression des derniers polynômes, ce qui permet de ne conserver que les paramètres significatifs pour chaque position de la fenêtre.

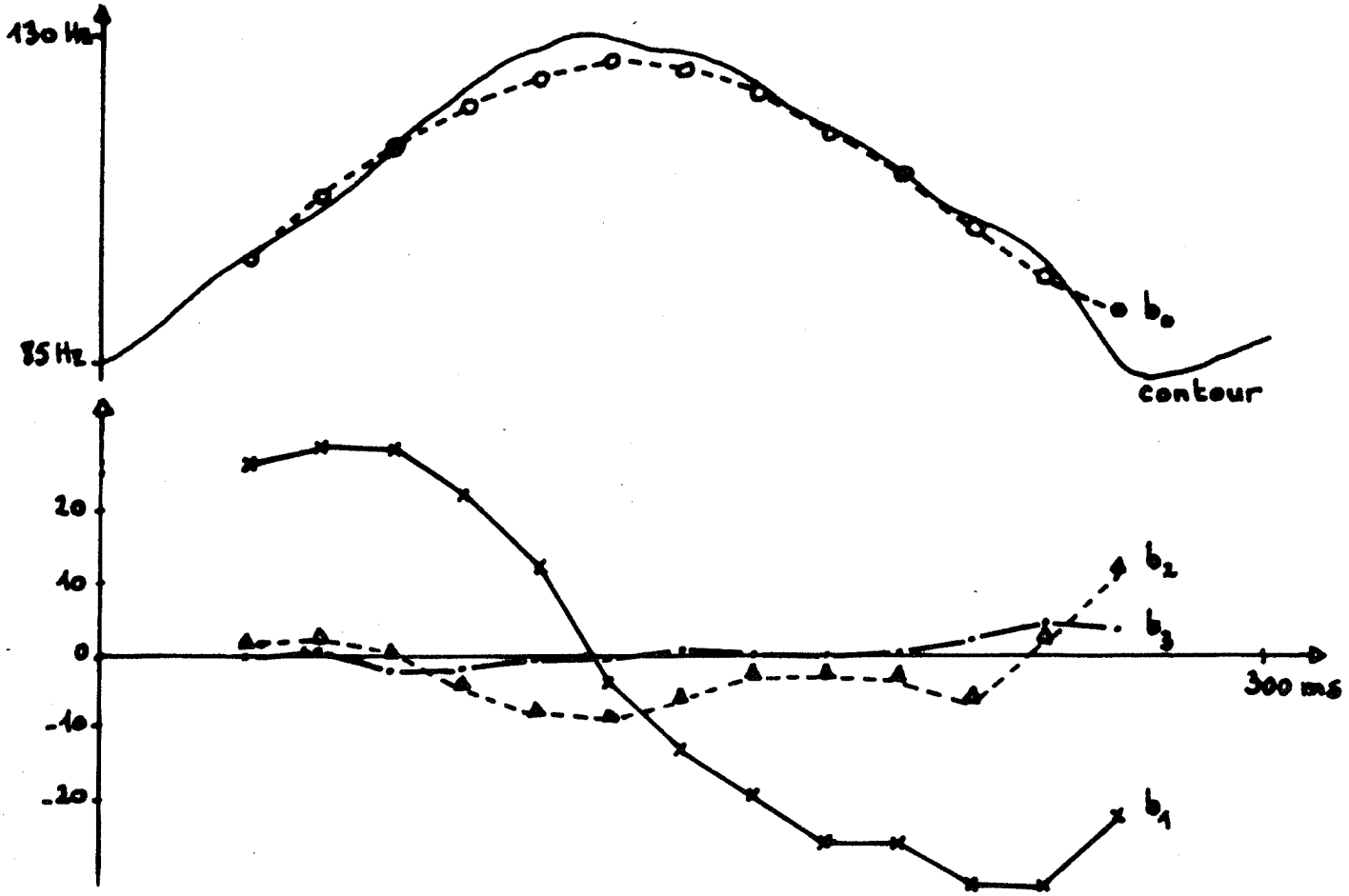
4. RESULTATS EXPERIMENTAUX

a) Exemple de représentation d'un contour mélodique :

La méthode a été appliquée au contour mélodique représenté fig. 1, d'une durée de 0,3s et correspondant à une portion d'une phrase affirmative. On a représenté l'évolution dans le temps des quatre premiers coefficients b_0, b_1, b_2 et b_3 du développement de Tchebycheff. Les coefficients d'ordre supérieur n'ont pas été portés car trop faibles et non significatifs.

Le coefficient b_0 fournit bien sûr une approximation de la valeur moyenne de la mélodie, en atténuant les variations locales. Les variations de pente du contour sont bien traduites par le coefficient b_1 : pour un contour normal les variations de b_1 sont assez lentes, comparées à celles du contour d'une voix de sourd (cf. Fig. 2). Le coefficient b_2 rend compte des extrema du contour mélodique et sa mesure fournit ainsi une mesure aisée des positions de ces extrema, tandis que b_3 renseigne sur les inflexions de la courbe.

De façon générale, plus les variations du contour étudié sont rapides, plus le nombre de coefficients à retenir est important, mais l'interprétation physique des coefficients d'ordre supérieur est de plus en plus difficile.



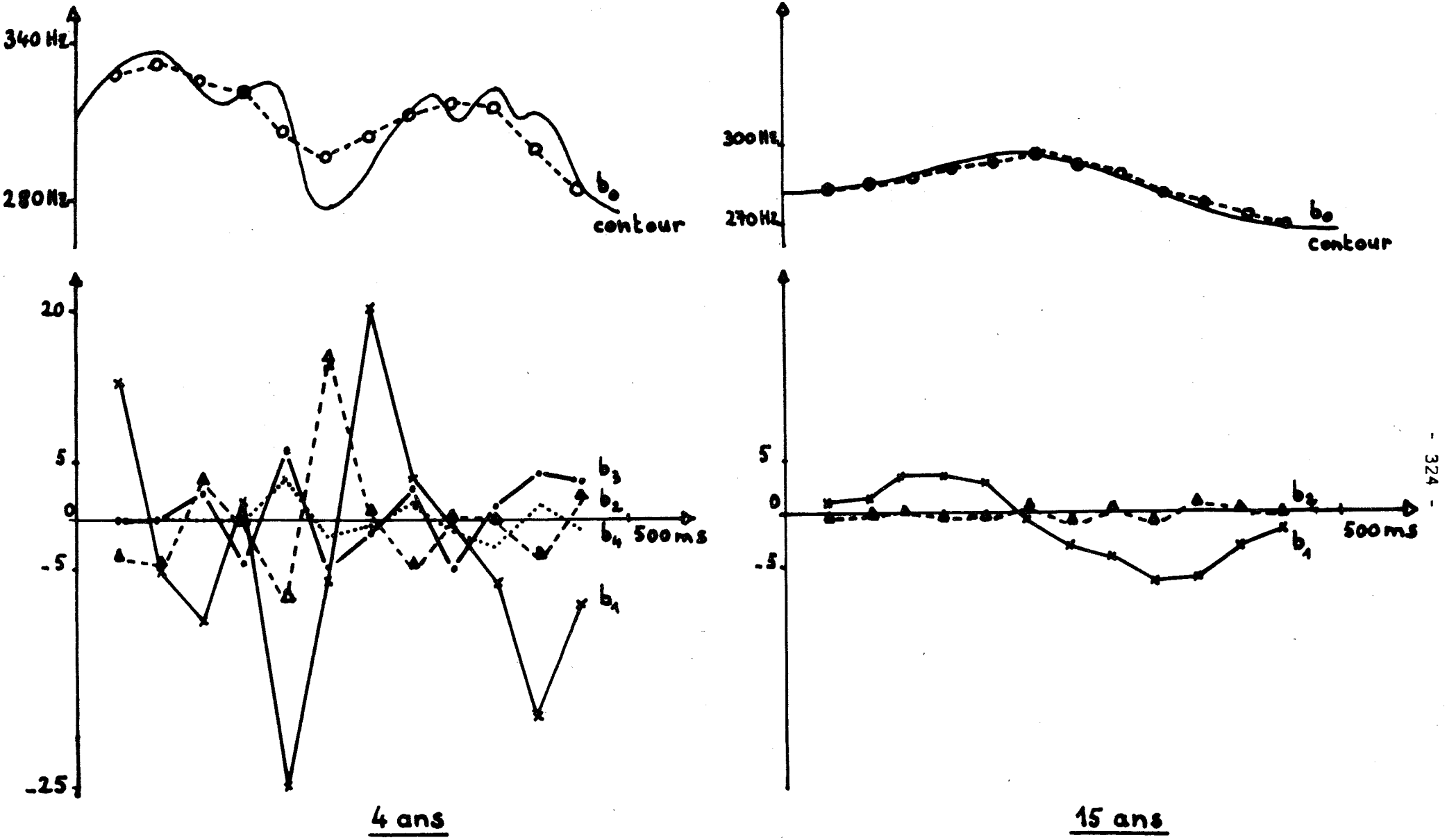
Analyse polynomiale d'un contour.

Figure 1.

b) Etude de l'évolution d'une voix :

L'analyse polynomiale de la mélodie permet d'étudier quantitativement l'évolution de la voix d'un sujet sourd au cours de la rééducation. A titre d'exemple, nous étudions la voix d'un enfant non-entendant à l'âge de 4 ans (début de la rééducation) et à l'âge de 15 ans. La fig. 2 montre, dans ces deux cas, l'évolution de la mélodie au cours de la prononciation d'une voyelle soutenue /a/ pendant 0,5s, ainsi que les analyses correspondantes, avec des échelles identiques.

La voix de 4 ans est caractérisée par des variations rapides et erratiques du pitch et il faut tenir compte en moyenne des 7 premiers coefficients du développement. L'évolution de b_1 est particulièrement caractéristique du manque de contrôle des cordes vocales.



Analyse polynomiale d'une voyelle soutenue

Fig. 2

La voix de 15 ans est déjà nettement mieux contrôlée ; les deux premiers coefficients sont suffisants pour rendre compte des variations de la courbe mélodique. L'examen de b_1 et b_2 fait encore apparaître une variation de la mélodie ; cette variation, due à l'effort d'émission fourni par l'enfant, est anormale pour une voyelle soutenue mais elle correspond cependant à une forme mélodique typique (cf. Fig. 1 l'évolution de b_1 et b_2).

5. CONCLUSION.

L'apprentissage du contrôle de la mélodie est une étape fondamentale de la rééducation des enfants non-entendants. Dans le cadre du système automatique de rééducation que nous développons actuellement, nous avons présenté une méthode d'analyse des contours mélodiques dans une base des polynômes de Tchebycheff. Cette analyse fournit une représentation plus intéressante qu'une analyse classique en monômes, en particulier en ce qui concerne la signification physique des coefficients obtenus. La méthode a été appliquée, en exemple, à l'étude de l'évolution de la voix d'un enfant en cours de rééducation.

Nous effectuons maintenant une analyse systématique d'un grand nombre de voix, en vue d'effectuer une classification des voix d'enfants à rééduquer, et une interprétation automatique des défauts.

THEME 3B

SORTIE PARLEE

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

SIMULATION DU CONDUIT VOCAL EN TECHNOLOGIES ANALOGIQUES

J.L. COURBON - Département Etudes et Techniques d'Acoustique

C.N.E.T. LANNION

RESUME

La méthode de simulation analogique qui est présentée est caractérisée par les deux fonctions distinctes : temps de propagation et calcul des transmissions et réflexions.

Le simulateur est essentiellement constitué de cellules de retard et de circuits d'interaction entre les ondes aller et retour dont le paramètre de commande est le rapport des deux aires adjacentes (plus récemment le taux de réflexion).

SUMMARY

SIMULATION OF THE VOCAL TRACT BY ANALOG CIRCUITS

Two different functions are used to simulate the vocal tract in this analog method : the propagation delay and the transmission - reflection values. Principal components of this simulator are delay lines and circuits for superposition of propagating and reflected waves.

The parameter controlling a section is the ratio between the two adjacent sections areas. (Recently the reflection coefficient).

SIMULATION DU CONDUIT VOCAL EN TECHNOLOGIES ANALOGIQUES

J.L. COURBON

C.N.E.T. LANNION

On suppose que le comportement du conduit vocal est identique à celui d'un tuyau rectiligne dans lequel ne se propagent que des ondes acoustiques planes normales à l'axe.

Le conduit est donc caractérisé en chaque point par sa section. Celle-ci est variable en fonction du temps. Il est donc assimilable, en première approximation, à une ligne électrique sans perte à constantes localisées (sous quelques réserves) (1).

I - CHOIX D'UNE METHODE DE SIMULATION ANALOGIQUE

1 - 1 - Eléments LC et inductances de synthèse.

La première méthode envisagée pour réaliser une ligne électrique sans perte est d'utiliser des cellules LC en Γ .

La difficulté est de construire des éléments capacitifs et selfiques de valeur variable en vue du fonctionnement dynamique.

La plage de variation de ces éléments doit être de l'ordre de celle de la fonction d'aire, c'est-à-dire de 1 à 100. La linéarité de variation doit aussi être suffisamment bonne devant les limites de validité de l'analogie.

A 3 kHz et pour des sections de conduit de 1 cm, l'impédance d'une cellule de l'analogue s'écarte de 5% de l'impédance caractéristique. Il est difficile d'atteindre des précisions supérieures dans toute la dynamique sur l'impédance d'une cellule (au moins deux éléments variables) de façon que cette nouvelle erreur soit négligeable.

Notons à ce sujet que le meilleur moyen d'obtenir des inductances de synthèse "en l'air" semble être le circulateur (circuit réalisable avec trois amplificateurs opérationnels en tension). Les éléments de commande sont alors des résistances variables.

1 - 2 - Convertisseurs d'impédance.

Plutôt que de chercher à réaliser des éléments variables, il est préférable d'interposer un transformateur ou un convertisseur d'impédance entre deux cellules successives à éléments fixes.

---

(1) J. GENIN - Simulation du Conduit Vocal.
Compte rendu des Journées d'Etudes sur la Parole
1972 - CNET LANNION pp 105 - 124.

.../...

Un transformateur d'impédance en courant est réalisable avec deux amplificateurs opérationnels, le rapport de transformation étant le rapport de deux résistances. Ainsi le nombre des éléments de commande peut être réduit pour un même nombre de cellules ou sections.

Le gyrateur, très intéressant aussi pour cette application, est dérivé directement du circulateur. Il peut être également obtenu avec une bonne qualité par un montage en antiparallèle de deux sources commandées (6).

La conversion d'impédance est du type $Z' = \frac{K}{Z}$, c'est-à-dire une inversion, aussi est-il nécessaire de changer d'analogie à chaque intersection donc d'alterner les cellules en π et en $\overline{T}$. Ceci conduit à un nombre impair de cellules.

I - 3 - Séparation des fonctions temps et calcul énergétique.

Le conduit vocal étant quantifié par portions de tuyaux cylindriques de 1 à 1,5 cm de long, les intersections sont les lieux de réflexion et de transmission des ondes incidentes.

La propagation dans la section cylindrique s'assimile au retard seul dans ce modèle. Aussi avons nous séparé les fonctions de retard et de transmission-réflexion.

Les sections sont simulées par des cellules de retard tandis que les intersections sont des circuits de désadaptation d'impédance, c'est-à-dire de calcul des cœfficients de réflexion et de transmission(2).

II - LE SIMULATEUR

Le temps de propagation du son dans l'air étant approximativement de 30 μs pour 1 cm, le signal de source vocale est échantillonné à 33 kHz et chaque échantillon est retardé d'une période pour chaque centimètre de conduit c'est-à-dire à chaque section. La première section est constante ; la dernière dépend de l'ouverture des lèvres qui est essentiellement variable.

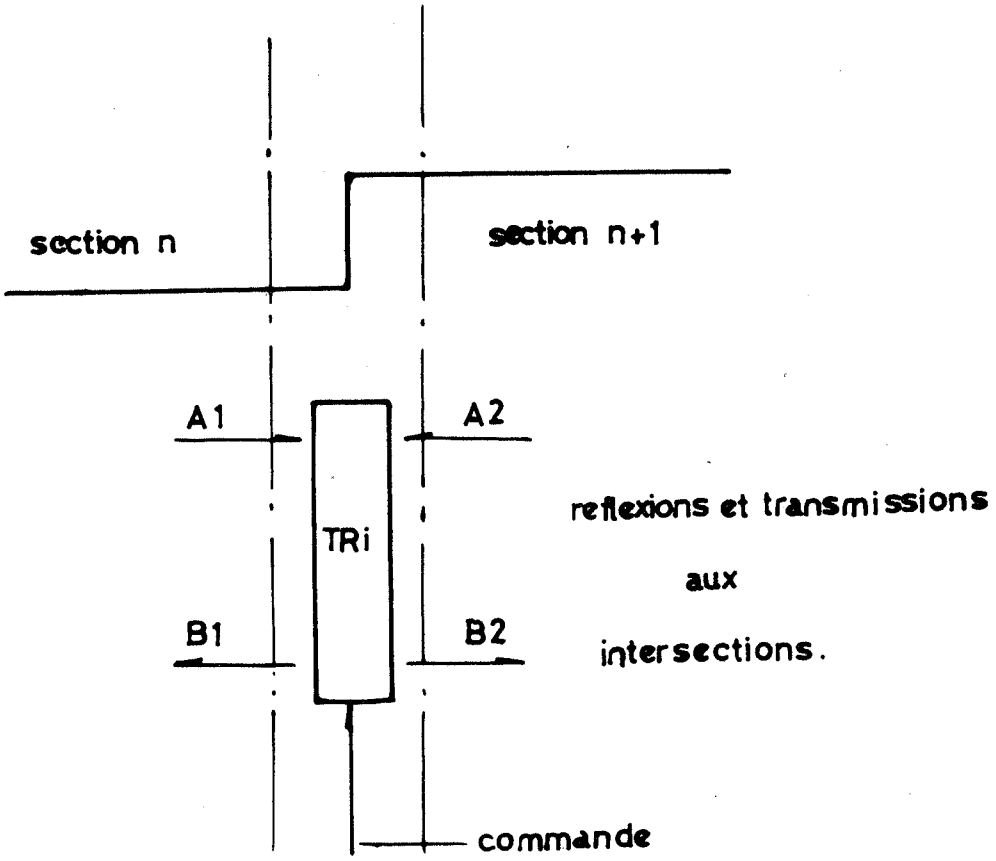
---

(6) J.C. MARCHAIS - L'amplificateur Opérationnel et ses applications. Editions MASSON et Cie.

(2) J.L. COURBON - La Simulation du Conduit Vocal- Analyse et Synthèse de parole CNET LANNION - 1972/73, 1pp 85-96.

.../...

II - 1 - Circuits de réflexion et de transmission d'ondes (TR_i)



Le schéma ci-dessus montre la fonction réalisée par ce circuit. La désadaptation d'impédance entre la section (n) et la section (n+1) est traduite en signaux électriques par une transmission d'énergie de A1 en B_2 et par une réflexion de A_1 en B_1 suivant des coefficients qui sont relatifs à la différence des aires des sections (n) et (n+1). De même pour A_2.

On a donc localisé en un point le changement d'aire d'une section à la suivante et l'on considère que l'aire est constante sur 1 cm (aire moyenne d'une section de 1 cm de conduit vocal).

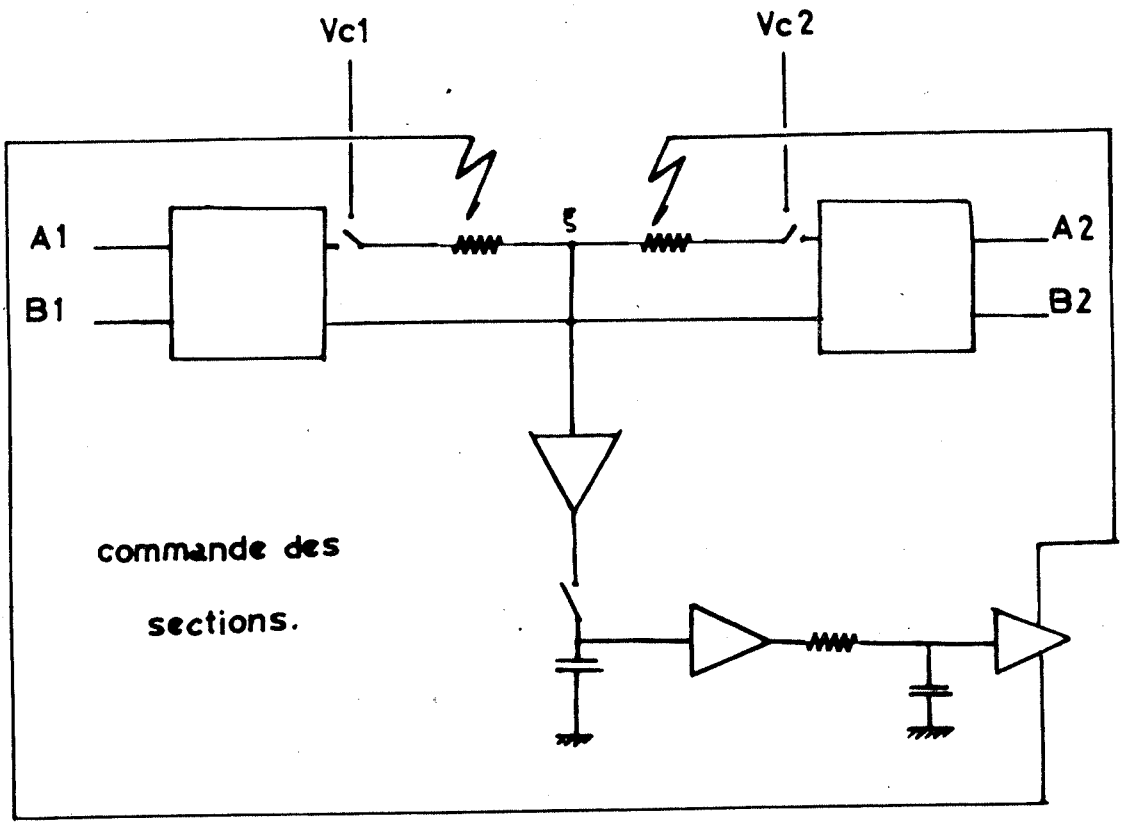
La commande des coefficients de réflexion et de transmission se fait au niveau du circuit par deux résistances.

Le problème que pose la commande de la valeur des résistances a été résolu en utilisant des photorésistances.

Entre deux échantillons successifs du signal, deux tensions de commande sont appliquées sur les bornes 1 et 2 (schéma ci-dessous) et les photodiodes alimentées en mode différentiel sont asservies par la tension d'erreur ε.

.../...

Cette tension ξ est mise en mémoireune fraction de période avant que les échantillons du signal soient présentés ; à ce moment les tensions de commande sont bloquées et la valeur ξ est maintenue en mémoire.

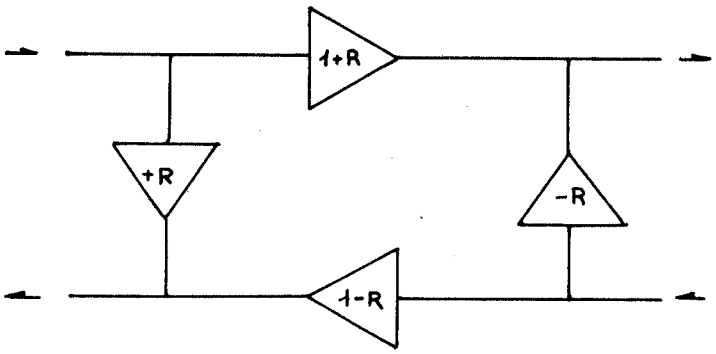


La valeur du rapport des aires des sections consécutives est égale à la valeur du rapport des résistances de ce circuit. Les tensions V_{c1} et V_{c2} commandent le rapport des aires de deux sections.

Une variation du rapport de 1 à 16 peut être effectuée par les résistances au bout d'un temps d'environ 40 ms avec une bonne précision, seul le rapport des résistances étant significatif et non leur valeur propre.

Une nouvelle version de ces circuits vient d'être réalisée. C'est le coefficient R de réflexion qui est le paramètre de commande de façon que la transmission aller soit (1 + R), retour (1 - R) et que les réflexions soient R et -R.

.../...



Ondes de pression.

Ce calcul repose sur des circuits intégrés multiplicateurs dont la linéarité et la dynamique sont suffisamment bonnes pour éviter l'utilisation de l'asservissement qui est un facteur de bruit important.

Parallèlement, une étude est en cours : l'utilisation des techniques de transport de charges ou chaîne à seaux.

La technologie MOS pourrait prétendre apporter dans cette optique un gain de volume et de coût de fabrication considérable.

II - 2 - Ligne à retard à échantillonnage

Un retard de l'ordre de 30 µs est réalisable sans difficulté pour un signal dont la bande ne s'étend guère au delà de 5 KHz. L'échantillonnage à 33 KHz laisse prévoir une bonne qualité du signal reconstitué.

Chaque échantillon est mis en mémoire analogique puis est tansféré dans une seconde mémoire au cours d'une seule période d'échantillonnage. L'étage suivant dont les temps de commutation sont synchrones prend en compte cet échantillon pendant la période suivante tandis que le premier étage prend simultanément un nouvel échantillon.

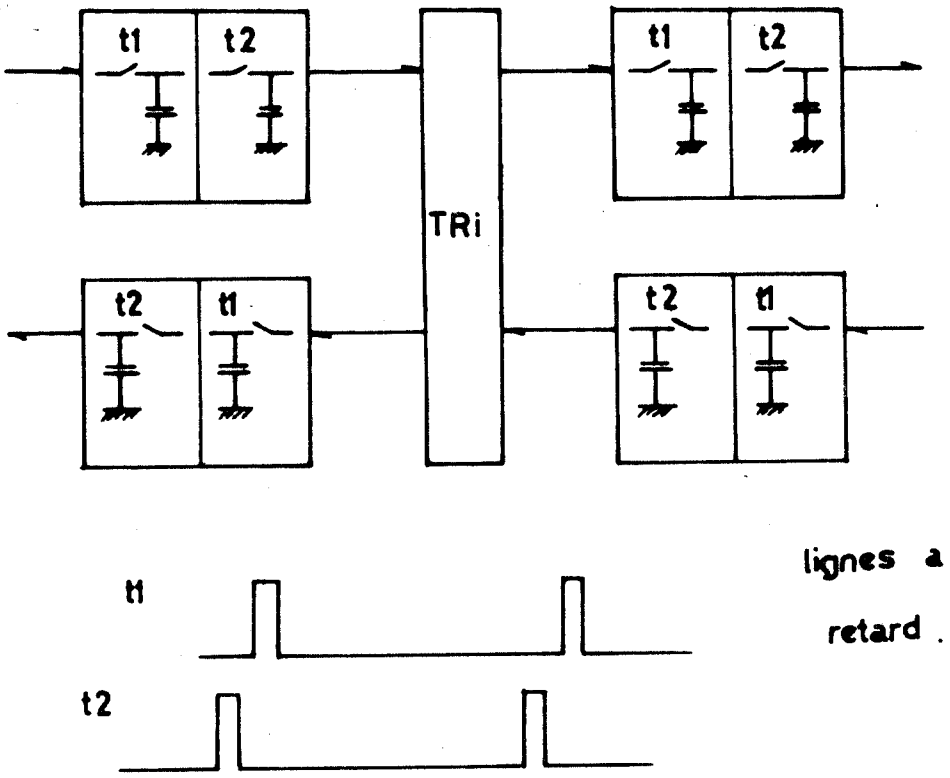
Les échantillons sont donc bloqués pendant une période à chaque période dans chaque étage.

Un retour des signaux réfléchis se fait de la même manière par une voie parallèle et de sens contraire.

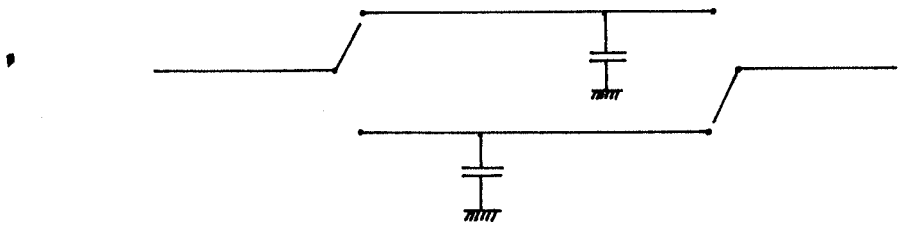
Le retard peut donc être ajusté très précisément à la valeur souhaitée par simple variation de la période d'horloge, mais toutes les cellules ont toujours un retard identique entre elles.

Les circuits de mémoire analogique sont des échantillonneurs-bloqueurs assemblés par groupe de 2 en série.

.../...



Le mode de commutation parallèle semblerait être plus favorable à l'heure actuelle grâce aux développements des circuits COS.MOS. en commutation analogique.



II - 3 - Les commandes des sections

Les photorésistances étant commandées par le support de deux tensions, l'une est fournie par un convertisseur digital-analogique, l'autre étant une tension fixe de référence.

Ces deux tensions sont de signes opposés.

.../...

Suivant la valeur du rapport désiré (> 1 ou < 1), un permutateur analogique aiguille les deux tensions selon "un bit de signe" en provenance du calculateur T 1600 ou du pupître manuel.

A chaque convertisseur est associé un registre mémoire de 8 bits à entrées et sorties parallèles, chacun d'entre eux pouvant être sélectionné par 4 bits d'adressage, de même que les permutateurs.

Bien que seulement 7 sections soient en oeuvre à l'heure actuelle, cette partie logique est prévue pour 17 sections en cas d'extension du nombre des sections.

Dans une deuxième version qui est en voie d'achèvement, c'est le coefficient de réflexion qui est le paramètre de commande. Les circuits de commande sont identiques.

II - 4 - Entrée des données de commande

L'information binaire peut être affichée directement par un jeu d'interrupteurs et adressée au numéro de la section.

Un ensemble de contrôle, adjoint depuis peu, permet de fournir les données par clavier et sous forme de valeur d'aire en cm2. C'est à partir d'une calculatrice de poche que le calcul du coefficient de réflexion est effectué $R = \frac{A_1 - A_2}{A_1 + A_2}$. Un programme câblé associé permet l'introduction de la valeur binaire soit directement au niveau des registres des intersections, soit dans une mémoire de 64 mots de 16 eb qui permet d'emmagasiner quatre configurations d'aire pour 17 sections de manière à obtenir un fonctionnement dynamique autonome.

L'accès par calculateur confère un fonctionnement à débit plus important pour l'étude des aspects dynamiques.

II - 5 - Source vocale

Le premier signal de source qui fut utilisé était issu d'un transistor unijonction monté en relaxateur et l'impulsion fournie était corrigée par une inductance (masse des cordes vocales) de façon à obtenir un spectre décroissant à 12 db/octave. Les transitions au démarrage et à l'extinction n'étaient pas possibles.

Une réalisation récente suivant le modèle à deux masses de FLANAGAN (3) permet de simuler l'onde émise par la glotte et l'interaction avec le conduit vocal. Les paramètres de commande sont :

- la pression sous glottique.
- la tension des cordes vocales.
- l'ouverture des cordes vocales au repos.

(3) M. CARCAUD - Simulation de la Source Vocale. Analyse et Synthèse de parole CNET LANNION 1972/73, 1 pp 97-103.

.../...

II - 6 - Charge aux lèvres, Conduit nasal, Sources de bruit

L'équivalent de la charge aux lèvres est un circuit RL, inductance de synthèse et résistance en parallèle simulant le rayonnement de la bouche.

Le même modèle selon RUIZ (4) est en cours de réalisation.

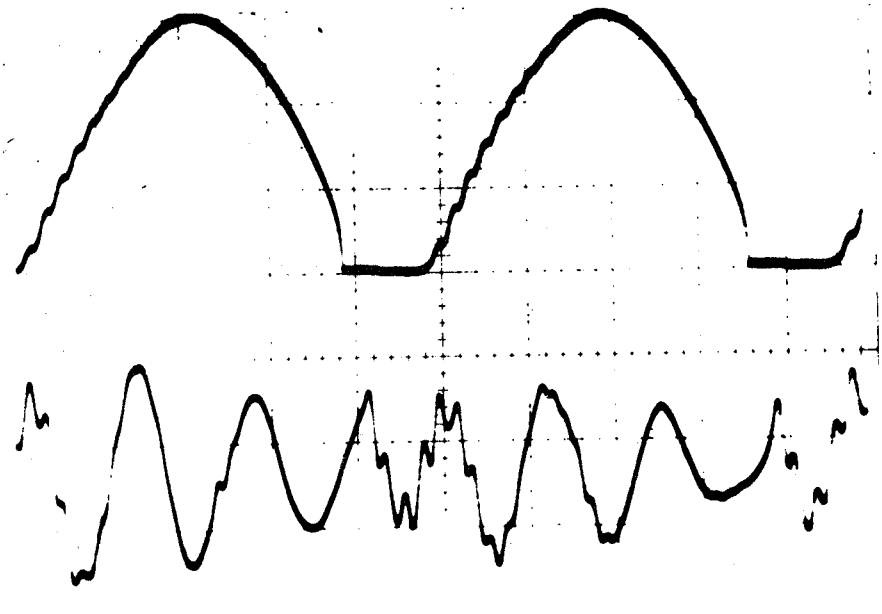
Le conduit nasal sera rapporté au conduit principal et ses sections seront probablement toutes fixes.

Quant aux sources de bruit, elles seront introduites aux constrictions du conduit lorsque certaines conditions de pression, de débit et de forme seront réalisées ; des points de mesure sont prévus au niveau des intersections pour permettre de prendre la décision d'insertion d'une source de bruit et pour le calcul de son niveau de pression et de son spectre (5).

III - POSSIBILITES ACTUELLES DE SYNTHESE

III - 1 - En fonctionnement statique (avec 7 sections)

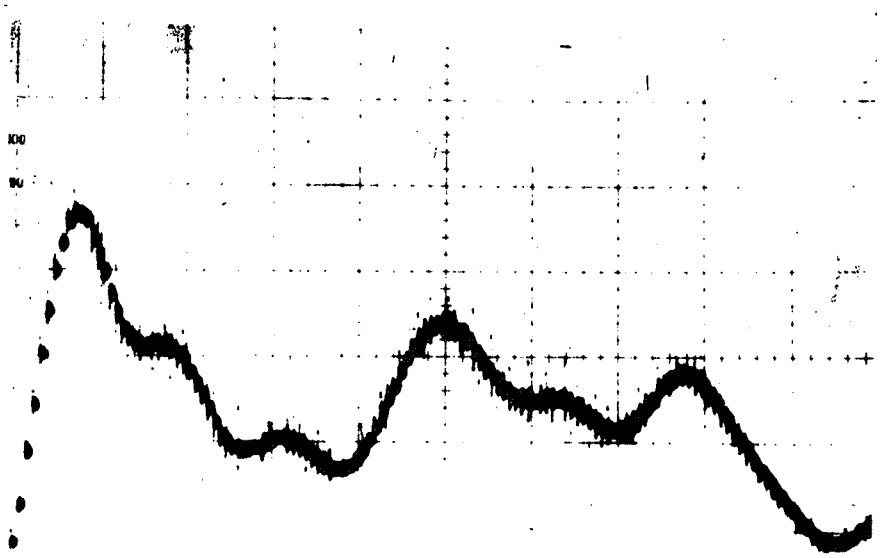
Les photographies ci-dessous donnent un exemple des spectres d'un i et d'un u simulés.



(5) B. GUERIN - Simulation du Conduit Vocal - Compte-rendu des Journées d'Etudes sur la parole 1972 - CNET LANNION pp 129-141.

(4) P.M. RUIZ - A program For speech synthesis by simulation of the vocal tract - 1968 - BELL TELEPHONE LABORATORIES -

.../...



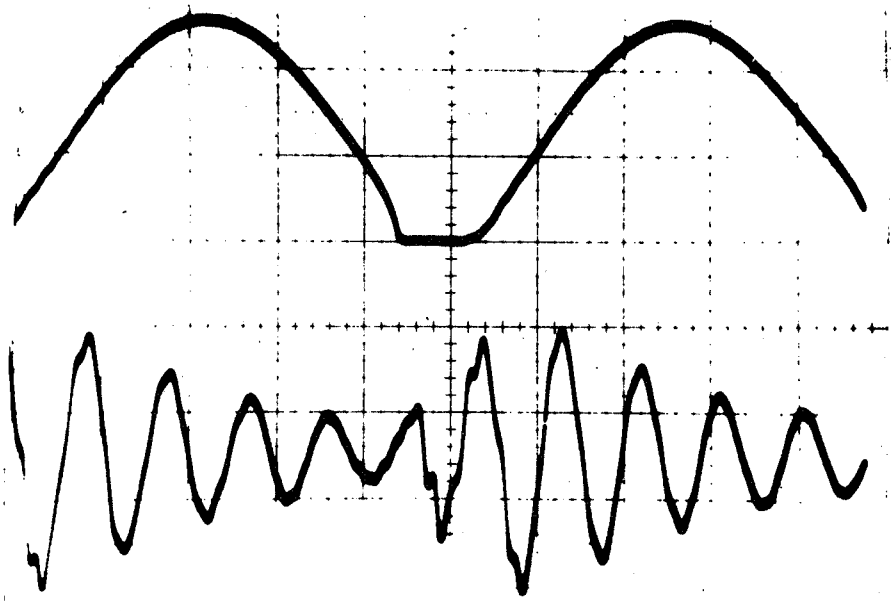
Spectre du signal

- U -

(4 kHz en abcisse)

L'axe horizontal est gradué linéairement de 0 à 4 kHz. Le signal d'excitation (onde de pression à l'entrée du conduit) montre l'effet du couplage entre la source et le conduit.

Pour obtenir le son u, le conduit doit être allongé de 1 à 2 centimètres (effet des lèvres). En fait toutes les sections sont augmentées simultanément d'un septième de l'allongement par variation de l'horloge (1 section = 1 période d'horloge). Le premier formant est trop élevé ; le nombre réduit de sections est une cause d'imprécision.

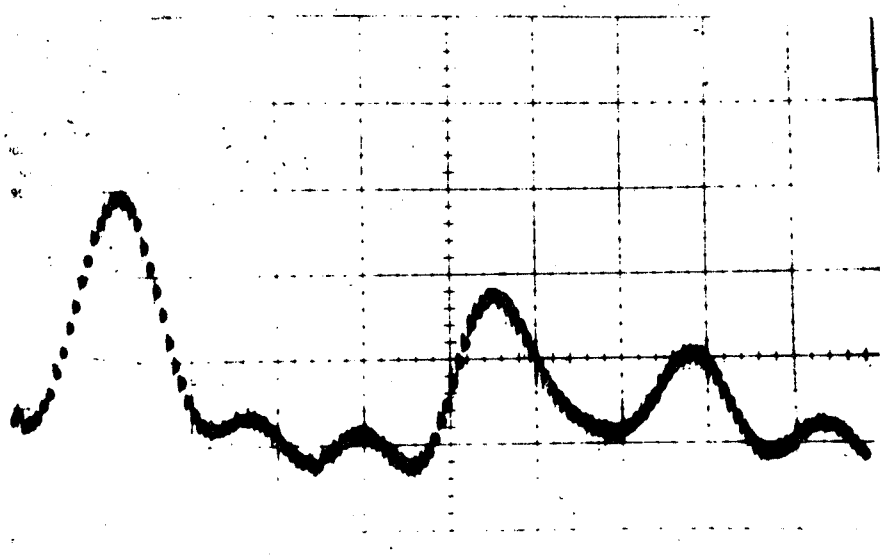


Signal de source

Signal de sortie du simulateur

I

.../...



Spectre du signal
- I -
(4 kHz en abcisse)

L'exploitation statique à l'oreille est pratiquement impossible car un son tel qu'une voyelle soutenue en permanence n'a plus guère de signification sans contexe auditif. Il faut nécessairement juxtaposer plusieurs sons dans le temps pour reconnaître un message vocal.

III - 2 - En fonctionnement dynamique (7 sections)

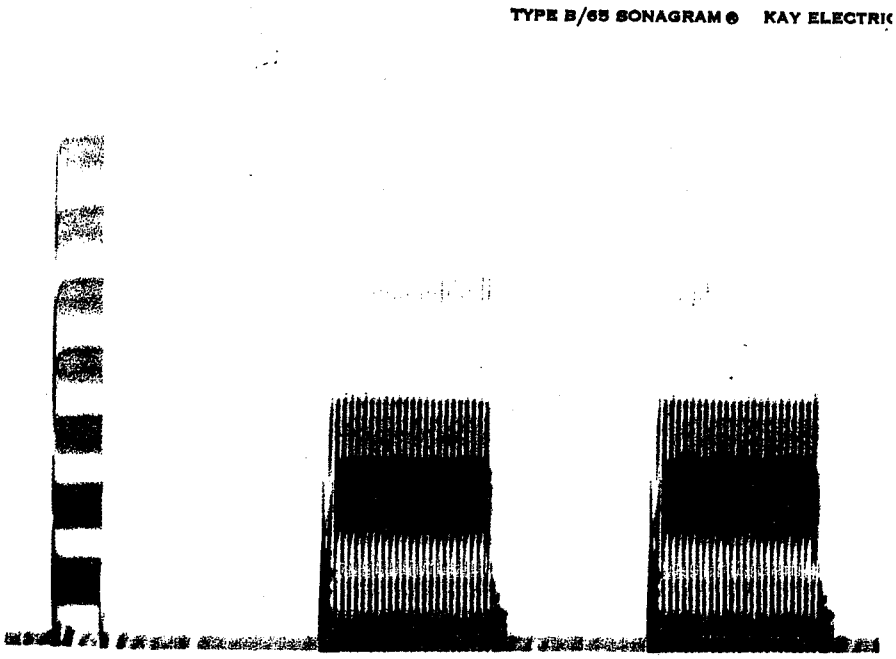
Une succession des voyelles A, O, OU, I, E, U à montré une certaine confusion entre A et O ouvert de même que la nécessité d'allonger le conduit pour une bonne distinction I, U.

Le sonagramme du mot PAPA montre la première tentative de fonctionnement dynamique : le passage d'une configuration dont la dernière section est fermée à une configuration de A est brutal si bien qu'il n'y a pas de période transistoire (on ne peut donc pas observer les locus).

La source, qui était bloquée lors de la fermeture, recommence à osciller dès l'ouverture et la surpression qui s'était formée dans le conduit (blocage en continu) produit un effet de plosive. Même lorsque le conduit n'est pas excité, la plosive est très bien marquée à l'ouverture car une simple tension de décalage d'un circuit se comporte comme une source de pression et porte le conduit au blocage.

En conclusion, la dernière réalisation à douze sections va apporter plus de finesse dans la définition des formants, une augmentation du rapport signal à bruit et une plus grande souplesse de commande.

-=-=-=-=-



Sonagramme de PA PA

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

"Introduction de paramètres prosodiques dans un programme de synthèse de nombres"

J. VAISSIERE - J. SAP

Laboratoires de Marcoussis

RESUME

Un programme de synthèse automatique de nombres à partir de mots isolés a été écrit pour la commande d'un synthétiseur à formants.

L'introduction de paramètres de rythme et d'intonation générés automatiquement à partir de règles simples améliore grandement le naturel des nombres synthétisés.

SUMMARY

A program of automatic synthesis of French numbers from separate words have been implemented for a formant synthesizer.

The authors describe the automatic generation of suprasegmental features using simple rules derived from a limited corpus.

The naturalness of synthesized speech is greatly improved by such rules.

INTRODUCTION DE PARAMETRES PROSODIQUES DANS UN PROGRAMME DE SYNTHESE DE NOMBRES.

Jacqueline Vaissière, Jacques Sap.

1.INTRODUCTION

Le but de ce rapport est de présenter l'ensemble des instructions qui ont été insérées dans un programme déjà existant de synthèse de nombres à partir de mots. Ces instructions sont destinées à conférer un rythme et une mélodie "naturels" aux nombres générés automatiquement par le programme.

Le rythme d'un nombre se traduit essentiellement par deux catégories de paramètres: premièrement, par la coupure éventuelle de ce nombre en deux troncons (ou plus) séparés par une pause (césure), et deuxièmement, par les variations relatives de durée entre les divers éléments mis en séquence dans le nombre. Quant à l'impression générale de mélodie, elle est essentiellement corrélée par l'évolution dans le temps des valeurs prises par la fréquence du fondamental (Fo): les variations de Fo engendrent la perception de hauteurs différentes sur les syllabes en séquence, et les variations de hauteur d'une syllable à l'autre confèrent au nombre une certaine mélodie.

Le programme complet se limite à la génération des nombres inférieurs à un million (c'est-à-dire de "zéro" à "neuf cent quatre-vingt dix neuf mille neuf cent quatre-vingt dix neuf"). Les instructions prosodiques insérées dans le programme suffisent cependant au traitement de n'importe quel chiffre, même si celui-ci contient des éléments de base tels que "million" ou "milliard".

L'étude a consisté dans un premier temps à composer un corpus d'analyse (ou liste de nombres), adapté exactement à l'analyse des deux facteurs prosodiques précédemment cités (rythme et mélodie). La méthode utilisée pour la composition de ce corpus est brièvement décrite dans la première partie de ce rapport. Nous avons procédé dans un second temps au choix d'un locuteur de référence. L'analyse des paramètres prosodiques s'est faite à partir de données tirées des sonogrammes des nombres du corpus lus par le locuteur sélectionné. Les résultats de cette investigation et les instructions insérées dans le programme qui en découlent sont présentés dans la seconde partie de ce rapport.

2. CORPUS D'ETUDE ET CHOIX D'UN LOCUTEUR:

2.1. Corpus d'étude:

Le corpus se compose de trois parties. Chacune de ces parties correspond à l'étude d'un seul facteur.

i. La première partie du corpus est destinée à l'étude de l'influence de la position d'un élément

dans un nombre sur la durée et la fréquence du fondamental de cet élément.

L'élément "deux" a été retenu comme premier élément de comparaison. L'élément "deux" peut en effet être commuté avec les dix-neuf premiers éléments ("un", "trois", "quatre" ... "dix-neuf"). Le corpus a été construit de facon à ce que l'élément 'deux' occupe dans des nombres de longueur variable toutes les positions possibles (nous entendons par 'longueur' le nombre d'éléments de base contenus dans le nombre). Par exemple, le corpus débute par le nombre "deux" (longueur 1), "deux cent" et "cent deux" (longueur 2), "deux cent mille", "cent deux mille" et "cent mille deux" (longueur 3) etc...

Les autres éléments ("cent", "mille" etc...) ont été étudiés à partir de listes composées de facon identique.

ii. La seconde partie du corpus est destinée à l'étude de l'influence de la nature intrinsèque d'un élément sur la durée et la Fo de cet élément.

Par exemple, les éléments de base de "un" à "dix-neuf" ont été étudiés dans l'environnement suivant:

X cent X mille X cent X

ou X représente un des dix-neuf premiers éléments.

iii. L'ensemble des chiffres précédants ont servi à l'étude de la présence et de la position de pauses éventuelles d'une part, et d'autre part de l'influence des pauses sur la durée et la fréquence du fondamental des éléments qui l'encadrent.

2.2. Choix du locuteur:

Douze locuteurs ont été enregistrés alors qu'ils prononcaient cinq phrases contenant neuf fois la séquence "soixante treize". Nous avons choisi au départ une voix naturelle (aisance du locuteur à parler devant un micro) et d'un rythme assez lent. En effet, la durée globale de la séquence en question variait de 66°/o entre un locuteur dont l'élocution était jugée comme "normale" par les auditeurs et un locuteur parlant "rapidement". Cependant l'évolution relative des valeurs de durée d'un élément à l'autre est sensiblement parallèle pour tous les locuteurs.

Et pour rendre la détection des formants moins complexe , une voix masculine a été selectionnée.

3. RESULTATS

3.1. Concernant le rythme:

i. Nous avons noté que le rythme se traduit en premier lieu par la coupure éventuelle du nombre en tronçons séparés par des pauses.

Nous pouvons résumer les résultats pertinents pour la synthèse des nombres de "un" à "un million" comme suit:

. une seule césure au maximum est apparue dans les

nombres inférieurs à un million. Si elle existe, cette césure est obligatoirement placée après l'élément "mille".

. les nombres de quatre éléments ou de moins de quatre éléments n'ont pas de césure (tout au moins dans le système de notre locuteur de référence).

. quelque soit le nombre d'éléments total, l'élément "mille" en position initiale ou en avant dernière position, n'est pas suivi d'une pause.

Le programme engendrera donc une pause après l'élément "mille" dans le cas des nombres de plus de quatre éléments, et où l'élément "mille" n'est ni en position initiale, ni en avant dernière position.

ii. Le rythme se traduit en second lieu par les variations relatives entre les divers éléments mis en séquence dans le nombre.

La durée d'un élément donné dans un nombre est fonction:

. premièrement, de la position de cet élément dans le nombre. Nos résultats confirment directement les conclusions de J. Génin (1) dans une étude qu'il a faite sur des nombres sans césure: durée longue en position finale ou si le nombre n'a qu'un seul élément; durée moyenne en position initiale ou après une césure, et durée courte dans les autres positions.

. et deuxièmement, de la position de la pause dans le nombre: l'élément suivant une pause a une durée moyenne, comparable à sa durée en position initiale dans le nombre.

Si n représente le nombre d'éléments dans un nombre à générer, et I le rang d'un élément dans ce nombre, l'organigramme du programme assignant pauses et durées peut se représenter comme indiqué sur la figure 1. Les éléments "million" et "milliard" ont un comportement prosodique équivalent à l'élément "mille" et M désigne un de ces trois éléments ("mille", "million" et "milliard").

3.2. Concernant la mélodie:

Appelons "tranche" une suite d'éléments de base encadrés par deux pauses. Un chiffre composé de plus de quatre éléments de base peut donc être divisé par la portion du programme présentée précédemment en plusieurs tranches.

Une tranche peut avoir deux contours mélodiques radicalement différents: soit la fréquence du fondamental dans son ensemble baisse du premier au dernier élément (contour 1), soit elle baisse du premier élément jusqu'à la fin de l'avant dernier élément et remonte sur le dernier élément (contour 2).

Si le nombre ne comporte pas de pause(s) et est en conséquence composé d'une seule tranche, soit le contour mélodique 1 ou le contour mélodique 2 lui est superposé, selon que ce nombre est prononcé par le locuteur avec une intonation descendante (contour 1) ou montante (contour 2)- affirmation versus enumeration, par exemple.

Si le nombre est composé de plusieurs tranches,

deux cas peuvent se produire: soit le nombre est prononcé avec une intonation descendante et dans ce cas, la tranche finale a un contour mélodique equivalent au contour1, les autres tranches (non finales) ayant un contour équivalent au contour 2; soit le nombre est prononce avec une intonation montante et dans ce cas, la tranche finale a un contour mélodique du second type et les autres tranches se terminent par une intonation descendante (contour 1).
Le locuteur signale clairement à son auditeur si une tranche est finale ou non finale en attribuant aux tranches finales une intonation différente des autres tranches. Le programme de synthèse utilisera le même procédé.

Nous avons choisi pour la synthèse le cas des nombres terminés par une intonation montante, car ce choix a été fait par le locuteur quand il a lu la liste des nombres du corpus. La figure 2 schématise les contours generes pour des chiffres de moins de quatre éléments et de plus de quatre éléments , avec et sans césure. Dans le cas des nombres avec césure(s), les valeurs maximales du fondamental sur chacune des tranches sont équivalentes. L'analyse révèle cependant une diminuation progressive des valeurs des maxima de la première à la dernière tranche. Il aurait été possible de simuler automatiquement cette diminution progressive, mais l'oreille semble se satisfaire d'une hauteur égale (mais si les valeurs sont supérieures, même très légèrement supérieures, sur la derniere tranche, l'oreille perçoit un "accent" d'insistance sur cette dernière tranche).

Pour simplifier le programme, les indices utilisés pour définir la durée des éléments sont aussi directement exploités pour la génération de la courbe de la fréquence du fondamental. Chaque élément de base a ete stocke dans trois tables différentes. Dans la table 1 et dans la table 3 sont stockées les informations sur les formants et la fréquence du fondamental des éléments de base en positions initiale et finale de nombre, respectivement (correspondant aux indices 1 et 3). Les éléments stockés dans la table 2 ont la fréquence fondamentale qu'auraient ces éléments s'ils avaient été prononcés avec un 'pitch' plat (on ne garde en fait que les informations concernant la micromélodie de ce nombre). Lorsqu'une tranche comporte plus de deux éléments (donc elle contient un ou plusieurs éléments d'indice 2), la fréquence du fondamental des éléments d'indice 2 est interpolée afin d'assurer la continuité entre le premier élément (obligatoirement d"indice 1) et le dernier élément (d'indice M ou d'indice 3). Lorsque la tranche ne comporte que deux éléments, soit l'élément M subit une translation vers le haut (dans le cas d'une tranche non finale) soit le premier élément subit une interpolation.

Une autre solution , plus économique du point de vue mémoire (mais se traduisant sans doute par une légère diminution de la qualité), aurait pu consister à ne stocker qu'une seule version de chaque élément et de modifier par calcul

leur durée. Plusieurs méthodes peuvent être utilisées: dans une méthode décrite par Rabiner, Schafer et Flanagan (2), des échantillons sont éliminés ou rajoutés automatiquement dans la partie stable du mot (elle même repérée automatiquement par le calcul de la dérivée du spectre), selon que le mot désiré est plus long ou plus court que le mot livré par le dictionnaire; dans une autre méthode décrite par Génin (1), et séduisante par sa simplicité un certain nombre de marqueurs sont disposés le long de l'élément stocké en mémoire. Cet élément représente la version la plus longue du mot (mot isolé). En fonction de la durée choisie pour l'élément intégré dans le nombre, certaines zones de l'élément sont synthétisées avec une vitesse plus ou moins rapide. Les méthodes que nous venons de décrire peuvent naturellement être employées avec l'algorithme assignant pauses, durée et fréquence du fondamental décrit dans le présent article.

4. CONCLUSIONS

Seule l'audition de nombres synthétisés avec l'aide du programme permet de juger de la validité des paramètres prosodiques générés. Le résultat essentiel réside dans le découpage en tranche(s) du nombre: alors qu'il a été souvent remarqué que le caractère naturel était souvent inversement proportionnel à la longueur des sègments synthétisés (que ce soit par synthèse complète par règles ou par mots), la qualité des nombres longs générés par ce programme peut être considerée comme égale à celle des nombres plus courts.

Cependant la qualité de la voix synthétique dans le présent programme pourrait encore être amelioréé par la réduction du pas de fréquence du fondamental (le pas actuel etant de 10 Hz) d'une part, et d'autre part par une meilleure répartition de l'énergie dans les graves.

5. DEMONSTRATION

...

REFERENCES:

1. J. Genin. "An Audio Response Unit for Telephone Needs", 1972, Conference on Speech Communication and Processing, Newton, Mass.

2. L.R. Rabiner, R.W. Schafer et J.L. Flanagan, "Computer Synthesis on Speech by Concatenation of Formant-Coded Words," Bell System Technical Journal, Vol. 50, NO. 5, 1971.

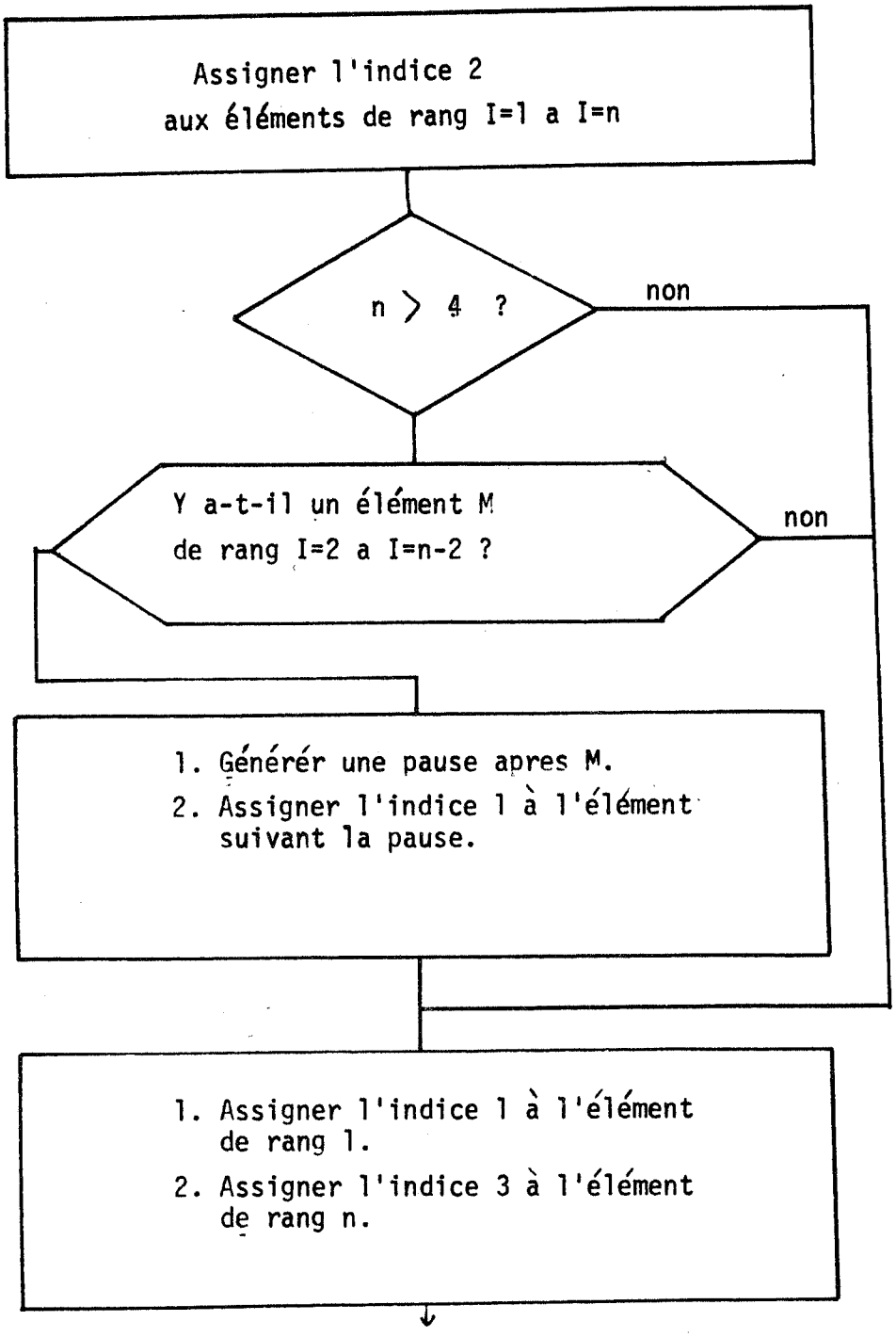


Figure 1: Algorithme représentant l'assignement des indices à chaque élément de base dans le nombre.

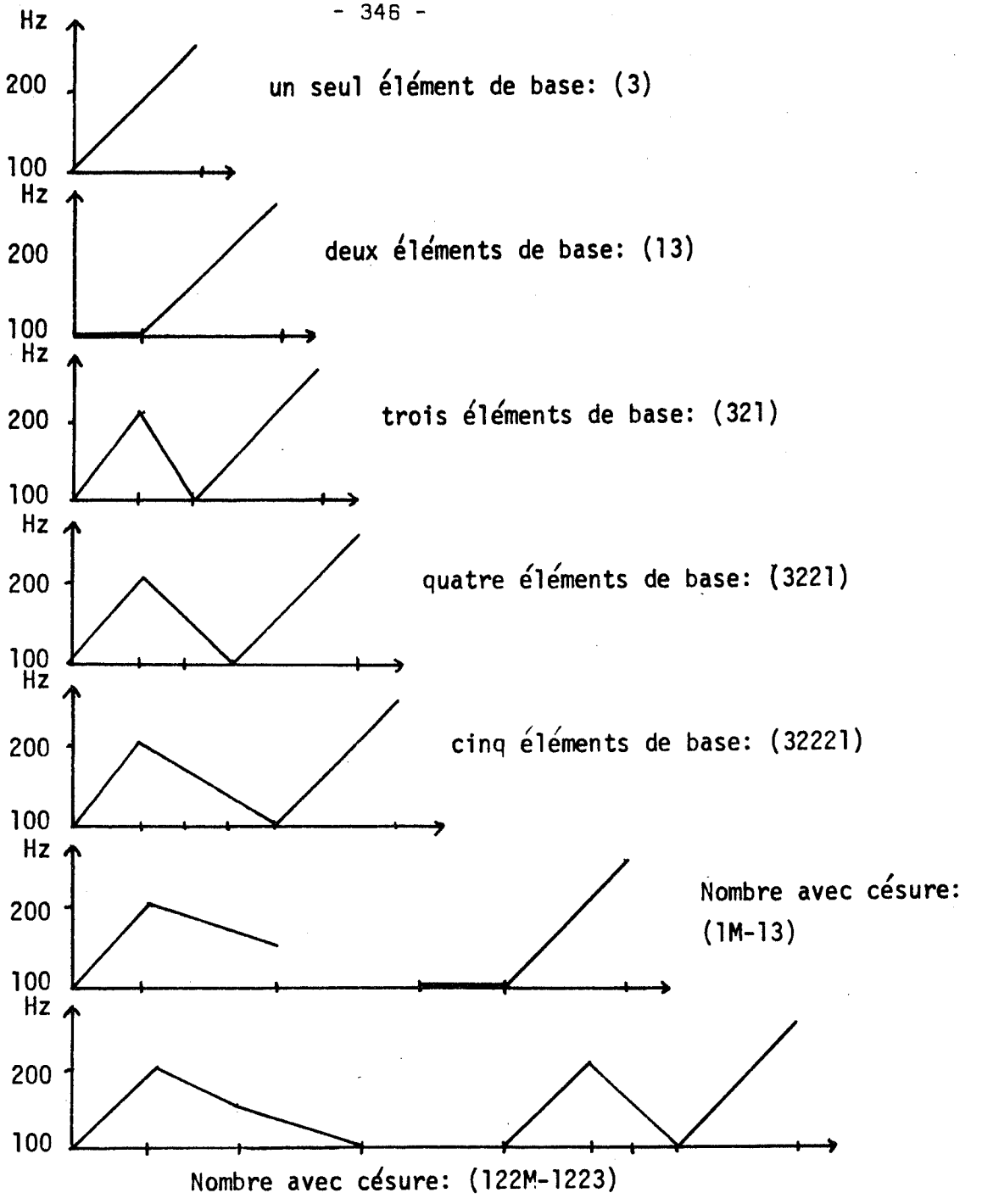


Figure 2: Exemples de courbes de fréquence du fondamental (micromélodie non comprise) générées pour des nombres avec et sans césure.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UN SYNTHETISEUR A STRUCTURE PROGRAMMABLE

C. GUEGUEN, J. LE ROUX, J. C. DOMENGER,

C. BENCHIMOL, J. F. BELLEC

RESUME

Le synthétiseur réalise, en technologie hybride, la fonction de transfert d'un système linéaire du dixième ordre sous la forme d'un filtre à structure programmée. Les paramètres de cette fonction de transfert peuvent être modifiés au cours du temps en commandant des multiplieurs digitaux-analogiques (DAM) par le calculateur. Utilisé comme synthétiseur le filtre est excité par un modèle de la source vocale. Plusieurs câblages standard des opérateurs élémentaires permettent d'accéder à diverses structures-type (filtre en treillis, vocoder à formants, ...). Ce synthétiseur peut être accessoirement utilisé comme filtre analogique ou numérique classique.

SUMMARY

This synthesizer realizes the transfer function of a 10th order linear system as an hybrid filter with programmable structure. These parameters can be modified by controlling some digital analog multipliers by the computer. When used as a speech synthesizer, the filter is fed with vocal source waveforms. Several pre-set wirings are available (lattice filter, formant vocoder, ...). This device can be used as a multipurpose digital or analog standard filter.

UN SYNTHETISEUR A STRUCTURE PROGRAMMABLE

C. GUEGUEN, J. LE ROUX, J. C. DOMENGER,
C. BENCHIMOL, J. F. BELLEC

Les méthodes modernes d'analyse du signal vocal peuvent être interprétées comme un processus d'identification du système de phonation. Le modèle couramment choisi est un système linéaire récurrent dont les coefficients sont ajustés au sens d'un critère de minimisation donné. Cette identification laisse cependant subsister un résiduel d'erreur qui peut être interprété comme l'excitation correspondante du canal vocal. Dans cette approche, la synthèse de parole peut être obtenue en simulant la fonction de transfert identifiée par un filtre récursif variable et en l'excitant par une forme schématique de la source vocale.

Dans le cadre de cette étude, le synthétiseur proposé réalise, en technologie hybride, une F.T. du dixième ordre sous la forme d'un filtre à structure programmée. Les paramètres peuvent être commandés par des multiplieurs digitaux-analogiques. Plusieurs câblages standards des opérateurs élémentaires permettent d'accéder à diverses structures-type (filtre en treillis, filtre transversal, cascade de second ordre, ...). Divers modèles de synthétiseur sont donc alors potentiellement disponibles (formants série, formants parallèles, PARCOR vocoder, ...).

Par ailleurs, du fait de son caractère non spécifique, le système peut être accessoirement utilisé comme un filtre numérique ou analogique classique.

1 - MODELISATION DU CANAL VOCAL PAR UNE FONCTION DE TRANSFERT

L'approche constituée par le codage prédictif du signal vocal peut être concrétisée par la figure 1 représentant le processus usuel d'une identification par une méthode du modèle. Le signal vocal s_n étant considéré comme la réponse du système phonatoire à une excitation e_n, un modèle paramétrique de structure fixée est alors soumis à la même excitation et produit une sortie $\hat{s}_n$. Un critère d'erreur est alors bâti sur un intervalle donné (moindres carrés, par exemple) et un algorithme d'optimalisation ajuste les paramètres du modèle au sens du critère d'erreur minimale. Une telle analyse peut être interprétée comme une analyse par synthèse et s'accorde donc de manière naturelle à la synthèse, a posteriori, de la parole.

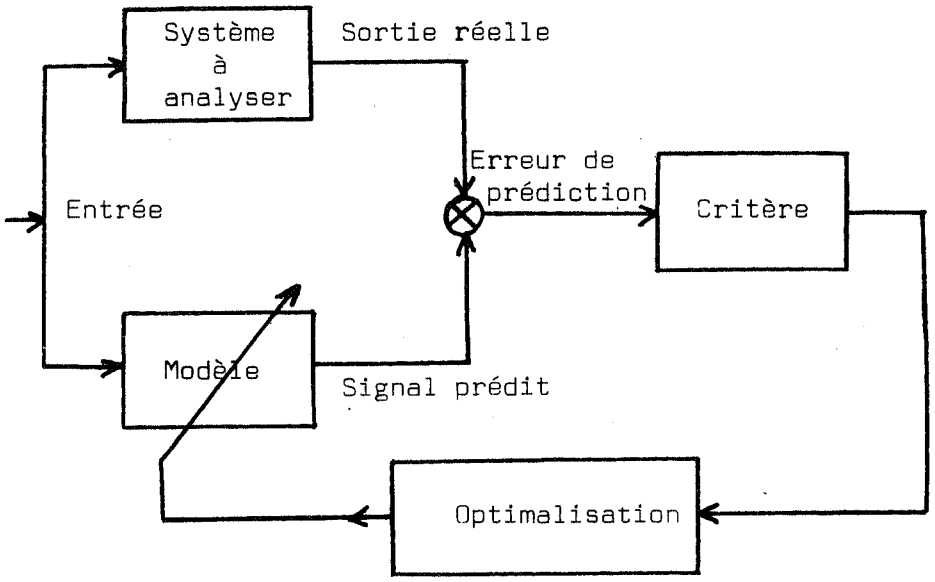


Figure 1 - OPTIMALISATION DU MODELE REPRESENTANT LE SIGNAL

Le modèle couramment utilisé est une équation de récurrence linéaire sans second membre :

$$s_n + a_1 s_{n-1} + \ldots + a_p s_{n-p} = e_n \qquad (1)$$

où s_n et e_n sont les échantillons du signal et de l'excitation à l'instant n et où les a_i représentent des coefficients inconnus à ajuster. Cette relation est usuellement interprétée comme une relation de prédiction sous la forme :

$$\hat{s}_n = - \sum_{i=1}^{p} a_i s_{n-i} \quad \text{avec } e_n = s_n - \hat{s}_n \qquad (2)$$

La fonction de transfert en z du modèle a donc pour expression :

$$F(z) = 1/A(z) \qquad \text{avec} \qquad A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}$$

On remarque que, sans la restriction (1), cette fonction de transfert ne comporte que des pôles mais l'approximation peut être rendue néanmoins suffisante par un choix convenable de l'ordre. La valeur p = 10 est, en général, un bon compromis.

Différents critères ont été suggérés (moindres carrés, maximum de vraisemblance, minimum de variance, ...) mais, dans le cas des bruits gaussiens, ces différents cas ne se distinguent que par des informations a priori différentes sur les paramètres a_i à estimer.

Les diverses méthodes se distinguent donc essentiellement au niveau de leur implantation et de leurs performances (temps calcul, encombrement mémoire, ...).

Les méthodes globales travaillent sur une fenêtre temporelle de l'ordre de plusieurs périodes fondamentales et utilisent comme étape intermédiaire le calcul des coefficients d'auto-corrélation de la suite s_n sur l'intervalle en question. Dans cette optique, les méthodes proposées par ITAKURA et SAITO /1/ à /3/, MARKEL et GRAY /4/ et /5/, utilisant l'algorithme de Levinson, semblent les plus performantes.

Les méthodes récursives travaillent de manière instantanée sur des schémas d'approximation stochastique. Elles participent des techniques auto-régressives utilisées dans les séries temporelles (ARMA). Parmi celles-ci, la méthode introduite par GUEGUEN et CARAYANNIS /6/ et /7/ se servant du filtrage de Kalman est très performante mais au prix de calculs coûteux. Des procédures sous-optimales peuvent être avantageusement implantées (divers types de gradient).

Le choix d'une technique d'analyse dépend donc du cadre de l'application envisagé. La transmission à faible coût de la parole nécessite une analyse en ligne (temps-réel) dont le manque d'acuité peut être compensé par un codage du résiduel de prédiction réel. Par contre, dans la synthèse proprement dite où les informations sur la source d'excitation sont plus schématiques, le temps consacré à l'élaboration d'un modèle plus précis peut être considérablement plus long (diminution de la longueur des mots mémorisés).

2 - REALISATION PRATIQUE DU SYNTHETISEUR

Dans l'optique précédente, l'instrument essentiel de la synthèse est un filtre numérique à coefficients commandables. Dans le cas présent, on a retenu, dans la réalisation, les options suivantes :

i - Au niveau des modules, utiliser une technologie hybride permettant une compatibilité complète entre les aspects numériques et analogiques ;

ii - Au niveau du câblage, ne pas spécifier une fois pour toute la structure du filtre mais permettre la programmation de diverses structures-type.

Dans ces conditions, le coeur du système est composé par une batterie de 10 modules représentant chacun une cellule de filtrage. Les fonctions principales sont représentées par un opérateur retard pur (1/z)(ou un intégrateur (1/p) au choix) et par un multiplieur digital-analogique figurant un coefficient variable. Les opérateurs de retard sont pour plus de sécurité composés par des doubles échantillonneurs-bloqueurs commandés à 10 kHz. Le reste des circuits tend à cadrer au mieux les coefficients dans leur plage de variation et à assurer un maximum d'interconnexions possibles entre les modules.

Les modules sont liés par un câblage extérieur assurant la structure particulière du filtre. Les divers programmes standard correspondent à l'association des cellules suivant une cascade de second ordre, un filtre transversal, un filtre en treillis. Cette dernière structure, utilisée, par ailleurs, par ITAKURA et SAITO, sous diverses formes, fait apparaître des coefficients k_i dotés des avantages suivants :

- la plage de variation est bornée comme celle des DAM au secteur $-1 < k_i < +1$
- leur codage requiert une moindre précision que les a_i

L'adjonction d'un module supplémentaire (finesse supérieure du modèle) ne modifie pas l'ajustement des sections précédentes.

Cette structure particulière explicitée par la figure 2 peut être obtenue par interconnexion des modules représentés par la figure 3. Pour assurer un nombre important de structures possibles, le module élémentaire possède diverses entrées et sorties mises en évidence sur la figure 4.

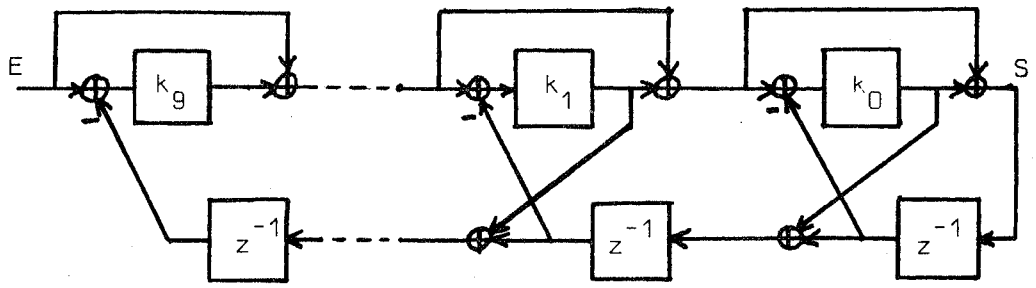


Figure 2 - SCHEMA DES MODULES DU SYNTHETISEUR ET DE LEUR LIAISON

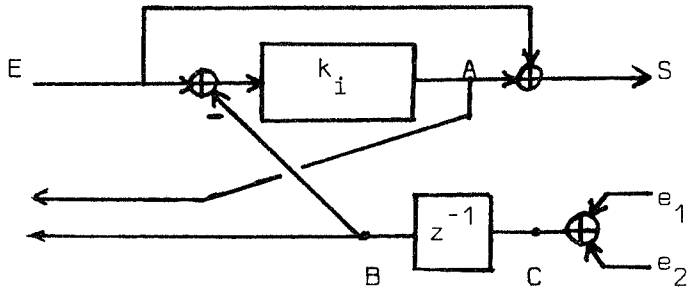


Figure 3 - SCHEMA D'UN MODULE DU SYNTHETISEUR

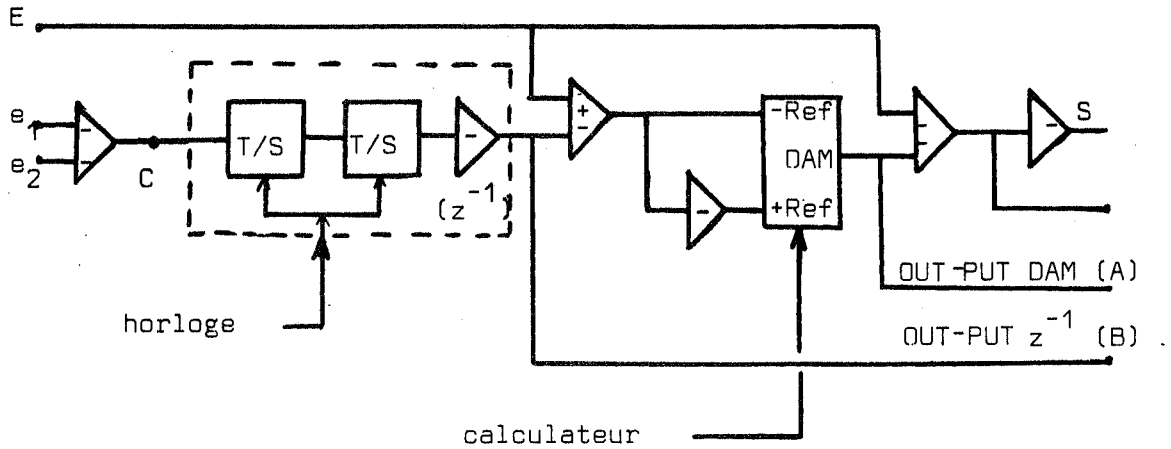


Figure 4 - SCHEMA SIMPLIFIE DU CABLAGE D'UN MODULE

Utilisant les notations de la figure 5, chacun des modules réalise pour i = 1, p, les fonctions :

$$s_n^{i-1} = s_n^i + k_i \, (s_n^i - x_{n-1}^{i-1})$$

$$x_n^i = x_{n-1}^{i-1} + k_i \, (s_n^i - x_{n-1}^{i-1})$$

avec

$$x_{n-1}^i = z^{-1} \cdot x_n \qquad \text{et} \qquad x_n^0 = s_n^0$$

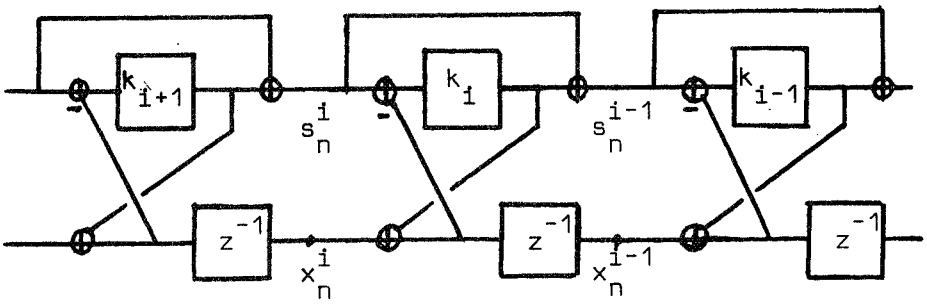


Figure 5 - LIAISON D'UNE PLAQUE AVEC SES VOISINES

4 - COMMANDE DU SYNTHETISEUR PAR LE CALCULATEUR

Le calculateur a pour rôle d'afficher les nouveaux paramètres k_i sur les multiplieurs digitaux-analogiques au moment voulu. Il le fait par l'intermédiaire d'un coupleur universel /8/. Il conserve dans une mémoire-tampon les nouveaux paramètres k_i à afficher, le nombre n_e d'échantillons pendant lequel ces paramètres seront utilisés.

Pour chaque nouveau groupe de paramètres, un compteur est réinitialisé à la valeur n_e. On le décrémente à chaque impulsion d'horloge. Quand sa valeur s'annule, les nouvelles valeurs n_e et k_i sont transmises de la mémoire-tampon au compteur et aux multiplieurs. Le calculateur charge ensuite les valeurs suivantes de n_e et des k_i dans la mémoire-tampon (figure 6).

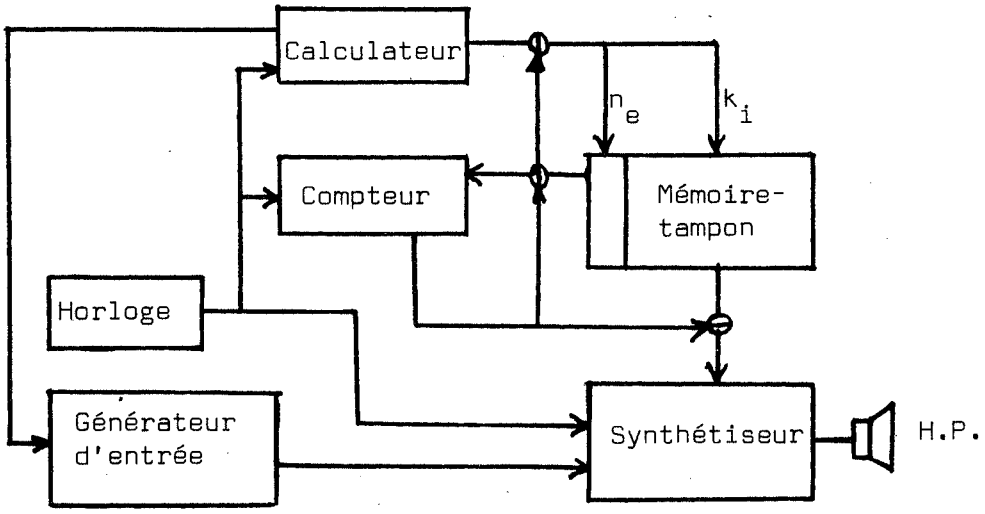


Figure 6 - TRANSMISSION DES INFORMATIONS DU CALCULATEUR VERS LE SYNTHETISEUR

A partir des informations d'amplitude, de fréquence fondamentale, de voisement de la parole à synthétiseur, le signal analogique d'excitation est produit par un mélange convenable de bruit et d'ondes glottales. En analyse-synthèse (communication numérique), ces informations sont prélevées sur le signal réel par un algorithme temps-réel permettant la détection synchrone du fondamental et l'estimation de l'énergie instantanée (LE ROUX /9/).

Dans les expériences présentes, on a utilisé la méthode de MARKEL pour la détermination des k_i pour un système d'ordre 10. La fréquence d'échantillonnage à la synthèse est 10 kHz et les paramètres sont renouvelés toutes les 20 ms. La parole synthétique est dans ces conditions de bonne qualité (8 kbits environ). Les efforts actuels sont consécrés à une meilleure modélisation de la source vocale et à une

réduction importante de la longueur des mots décrivant les k_i. Il semble qu'une telle réduction peut être envisagée sans détériorer notablement la qualité du système.

BIBLIOGRAPHIE

/1/ F. ITAKURA, S. SAITO : Analysis synthesis telephony based on the maximum likehood method

6 Int. congress on acoustics, C 5.5, Tokyo, 1968

/2/ F. ITAKURA, S. SAITO : Digital filtering techniques for speech analysis and synthesis

7 Int. congress on acoustics, 25 C 1, Budapest, 1971

/3/ F. ITAKURA, S. SAITO : A statistical method for estimation of speech and formant frequencies

Electronics and communication in Japan, Vol 53 A,, n°1, p. 36-43, 1970

/4/ J. D. MARKEL, A. H. GRAY : On autocorrelation equations as applied to speech analysis

IEEE Trans. on Audio. and Elec., Vol AU-21, n° 2, pp. 69-79, 1973

/5/ A. H. GRAY, J. D. MARKEL : Digital lattice and ladder filter synthesis

IEEE Trans. on Audio. and Elec., Vol AU-21, n° 6, pp. 491-500, 1973

/6/ C. GUEGUEN, G. CARAYANNIS : Analyse de la parole par filtrage optimal de Kalman

Automatisme (Dunod ED.), Tome 18, n° 3, 1973

/7/ G. CARAYANNIS : Analyse de la parole par identification récurrente d'un modèle du système de phonation

Thèse de Docteur-Ingénieur, Paris VII, Novembre 1973

/8/ NGUYEN CHI THANH : Conception et réalisation d'un interface multi-périphérique pour miniordinateur. Application à la synthèse de la parole et au traitement d'image

Thèse de Docteur-Ingénieur, Paris VII, 1975 (à paraître)

/9/ J. LE ROUX : Une méthode synchrone d'extraction en temps réel de fondamental

Compte-rendu des 6ème Journées d'Etude sur la Parole, Toulouse, Mai 1975

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

SYNTHETISEUR DE PAROLE A FORMANTS A CIRCUITS NUMERIQUES

Jean GREVIN Francis BERTIN Jacques SAP

Laboratoires de MARCOUSSIS

RESUME

Nous décrivons un synthétiseur de parole à formants à circuits numériques qui, connecté à un ordinateur, fonctionne en temps réel.

Onze paramètres sont transmis du calculateur au synthétiseur toutes les 8, 16 ou 24 milli secondes. Un point de l'onde de parole est calculé toutes les 125 micro secondes.

SUMMARY

We describe a hardware implemented digital speech formant synthesizer which, connected to a digital computer, produces speech in real time at a sampling rate of 8 KHz.

The computer transmits the values of eleven parameters to the synthesizer every eight, sixteen or twenty four milli seconds.

SYNTHETISEUR DE PAROLE A FORMANTS A CIRCUITS NUMERIQUES

Jean GREVIN Francis BERTIN Jacques SAP

Laboratoires de MARCOUSSIS

Cet article décrit un synthétiseur à formants à circuits numériques qui, connecté à un ordinateur, produit de la parole en temps réel.

L'ordinateur transmet au synthétiseur les valeurs de onze paramètres toutes les huit, seize ou vingt quatre millisecondes suivant la finesse désirée. Le synthétiseur produit un échantillon de l'onde de parole toutes les cent vingt cinq microsecondes codé sur dix éléments binaires.

Le principe du synthétiseur est celui des synthétiseurs à formants du type série. Les fonctions de transfert du synthétiseur sont réalisées grâce à un algorithme câblé effectuant en séquence tous les filtrages récursifs nécessaires. Pour cela, on dispose en particulier de mémoires, d'un additionneur et d'un multiplieur capable de réaliser la multiplication de deux nombres de seize éléments binaires chacun (résultat sur trente deux éléments binaires) en moins d'une demi-microseconde.

I - CARACTERISTIQUES DU SYNTHETISEUR

Le schéma équivalent du synthétiseur est celui du canal vocal des synthétiseurs à formants du type série (figure 1).

Une source vocale de fréquence variable F0, produit un signal dont la variation d'amplitude est obtenue par multiplication par un facteur variable A0, le signal résultant est ajouté au signal d'une source de bruit multiplié par un facteur d'amplitude variable AR puis transmis à une série de quatre filtres du second ordre dont les fréquences de résonance des trois premiers sont variables.

Les caractéristiques des filtres sont les suivantes :

Filtres du second ordre à bandes passantes constantes.

F1 : Fréquence de résonance variant de 100 Hz à 875 Hz, bande passante de 70 Hz ;

F2 : Fréquence de résonance variant de 500 Hz à 3 650 Hz, bande passante de 100 Hz ;

./...

F3 : Fréquence de résonance variant de 1 800 Hz à 4 950 Hz, bande passante de 120 Hz ;

F4 : Fréquence de résonance fixe de 3 500 Hz, bande passante de 200 Hz.

Le synthétiseur produit donc un son permanent lorsque les valeurs des six paramètres effectifs de sa commande sont fixes. (sons élémentaires)

On produit une élocution articulée en faisant varier dans le temps les valeurs des paramètres.

Un moyen commode de faire varier à la fois tous ces paramètres est d'utiliser un ordinateur dans la mémoire duquel on stocke les valeurs successives des paramètres.

II - PRINCIPE DU SYNTHETISEUR NUMERIQUE

Le schéma de la figure 1 n'est pas physiquement réalisé. Il est simulé par le synthétiseur numérique qui, fonctionnant comme un calculateur, calcule successivement sur 16 éléments binaires les valeurs des tensions aux divers points du schéma équivalent, ce calcul est effectué toutes les 125 microsecondes. Le résultat final, dont on ne garde que 10 éléments binaires, est transmis à un décodeur digital analogique.

La fonction de transfert des filtres est de la forme :

$$H(F) = \frac{1}{1 - (F/Frès)^2 + j\Delta F/Frès}$$

F : Fréquence
Frès : Fréquence de résonance
ΔF : Bande passante

On obtient une bonne approximation du fonctionnement d'un tel filtre en écrivant que sa sortie à l'instant t est fonction linéaire de ses sorties précédentes aux instants $t - T$ et $t - 2T$ ainsi que de son entrée à l'instant t

$$y(t) = K0.x(t) + K1.y(t - T) + K2.y(t - 2T)$$

T est le retard considéré ; dans la suite de nos opérations nous aurons évidemment T = 125 microsecondes.

Les coefficients K1, K2 et K0 sont liés aux paramètres du filtre ΔF et Frès ainsi qu'à la période d'échantillonnage T

$$K1 = 2e^{-\pi.\Delta F.T} \cos(2\pi.F_{rés}.T)$$
$$K2 = -e^{-2\pi\Delta F.T}$$
$$K0 = 1 - K1 - K2$$

./...

Une caractéristique essentielle du synthétiseur est que toutes les valeurs possibles des coefficients K0, K1 et K2 sont gardées en mémoires mortes.

F1 est codé à l'aide de 5 éléments binaires :

00000 représente 100 Hz

00001 représente 125 Hz

00010 représente 150 Hz

⋮

11111 représente 875 Hz

Les calculs de coefficients étant difficiles et ne se réduisant pas à un nombre limité d'additions et de multiplications, les 32 valeurs de K1, les 32 valeurs de K0 et la valeur unique de K2 obtenues pour toutes les valeurs de Frès avec F = constante = 70 sont disponibles par simples adressages de mémoires mortes.

Il en va de même pour F2 et F3 codés sur six éléments binaires chacun

F2 varie de 500 Hz à 3 650 Hz par pas de 50 Hz
F3 varie de 1 800 Hz à 4 950 Hz par pas de 50 Hz

A la demande du synthétiseur l'ordinateur présente les valeurs courantes des paramètres qui sont gardées sur des registres.

Le renouvellement de ces paramètres n'est pas nécessaire au synthétiseur qui synthétise le son permanent correspondant aux valeurs actuelles contenues dans les registres d'entrée. Si les valeurs ne sont pas renouvelées un son permanent est produit.

III - ELEMENTS CONSTITUTIFS DU SYNTHETISEUR

Le schéma est donné par la figure 2, il comporte :

- Générateur d'instructions.
  Il engendre 55 opérations élémentaires à la suite l'une de de l'autre. Ces opérations, représentées chacune par le changement d'état d'un fil physiquement distinct, consistent en des ordres de lectures ou écritures de mémoires, chargement ou décalages de registres.
- Multiplieur à fonctionnement simultané.
  On présente les deux opérandes à l'entrée sur 16 éléments binaires chacun et le résultat est disponible sur 32 éléments binaires à la sortie au bout d'un temps court qui est la somme des temps de propagation dans les divers éléments.
- Additionneur, soustracteur à fonctionnement simultané.
- Mémoire morte.

./...

- Inter-Face.
- Mémoire vive.
  On y garde les résultats intermédiaires des opérations.
- Source de bruit.
- Source vocale.

IV - FONCTIONNEMENT DU SYNTHETISEUR

Nous utilisons les symboles suivants :

- Sex : signal d'excitation
  c'est le signal qui pénètre dans le filtre de formant F4 (voir figure)
- S vocal : signal issu de la source vocale
- S bruit : signal issu de la source de bruit
- K_{0F4}, K_{1F4}, K_{2F4} : coefficients K_0 K_1 K_2 pour le formant F4
- $Y_{F4,0}$: signal issu du filtre de formant F4, à l'instant considéré
- $Y_{F4,-1}$: signal issu du filtre de formant F4 à l'instant précédant l'instant considéré
- $Y_{F4,-2}$: signal issu du filtre de formant F4 à l'instant précédant $Y_{F4,-1}$

Voici, sommairement, les opérations successives effectuées par l'algorithme :

- Calcul du signal d'excitation (signal à l'entrée du premier filtre)

  Sex = S vocal x A0 + S bruit x AR
  Sex = Signal d'excitation
  S vocal = Signal issu de la source vocale
  S bruit = Signal issu de la source de bruit

- Calcul de l'effet du filtrage du formant F4
  Adressages de la mémoire morte pour obtenir les coefficients K_{0F4}, K_{1F4}, K_{2F4}

  $$Y_{F4,0} = K_{0F4} \cdot Sex + K_{1F4} \cdot Y_{F4,-1} + K_{2F4} \cdot Y_{F4,-2}$$

  $$Y_{F4,-2} = Y_{F4,-1}$$

  $$Y_{F4,-1} = Y_{F4,0}$$

./...

- Calcul de l'effet du filtrage du formant F2
  on procède comme précédemment en utilisant $Y_{F4,0}$ comme signal d'entrée
- Calcul de l'effet du filtrage du formant F1
- Calcul de l'effet du filtrage du formant F3
- Transmission du résultat vers le registre de sortie
- Incrémentation de deux compteurs :
  - Le premier indique qu'une période du fondamental est écoulée, en ce cas, il faut le recharger et réinitialiser la source vocale.
  - Le second indique que soixante quatre points ont été écoulés et qu'il faut redemander des paramètres à l'ordinateur.

On remarque que l'on a choisi pour les filtres l'ordre F4 F2 F1 F3.

Cet ordre a été choisi à la suite de simulation du synthétiseur par programme d'ordinateur.

Il représente un compromis entre le meilleur rapport signal/bruit et l'absence de saturations.

CONCLUSION

Le synthétiseur numérique possède un nombre plus important d'éléments qu'une réalisation à l'aide de circuits analogiques.

Il a l'avantage de ne nécessiter aucun réglage et donc de donner un résultat sûr. Il permet d'avoir une très bonne dynamique et peut être considéré comme un instrument de travail puissant et précis pour tous les travaux de phonétique.

B I B L I O G R A P H I E

1 - J. GREVIN — "Etude et réalisation d'un synthétiseur numérique de parole". Thèse de 3ème cycle - Orsay 1974

2 - J.J. MASSOT — "Etude et réalisation d'une unité de réponse vocale". Thèse de 3ème cycle - Orsay 1971

3 - J. PAILLE — "Contribution aux études sur la synthèse paramétrique de la parole synthétiseur à formants - analogue de la source vocale". Thèse Docteur es Sciences Physiques Grenoble 1971

4 - R. CARRE — "Contribution aux études sur l'analyse et la synthèse de la parole - Rôle et importance des formants". Thèse Docteur es Sciences Physiques Grenoble 1971

5 - G. FANT — "Acoustic analysis and synthesis of speech with applications to swedish". Tiré à part du n° 1 de "Ericcson technics" 1959

6 - G. ROGER, E. KARSENTI, J. SAP — "Application des procédés de visualisation à l'étude d'un synthétiseur paramétrique". AUTOMATISME - mars 1973 n° 3

7 - RABINER, JACKSON, SCHAFER, COKER — "Hardware realisation of a digital formant speech synthesizer". I.F.F.E. Transactions on Communication Technology, Vol. COM 19 n° 6 Décembre 1971

8 - G. FANT — "Acoustic theory of speech production" Mouton - La Haye 1970

9 - J.L. FLANAGAN — "Speech analysis synthesis and perception" Springer Verlag

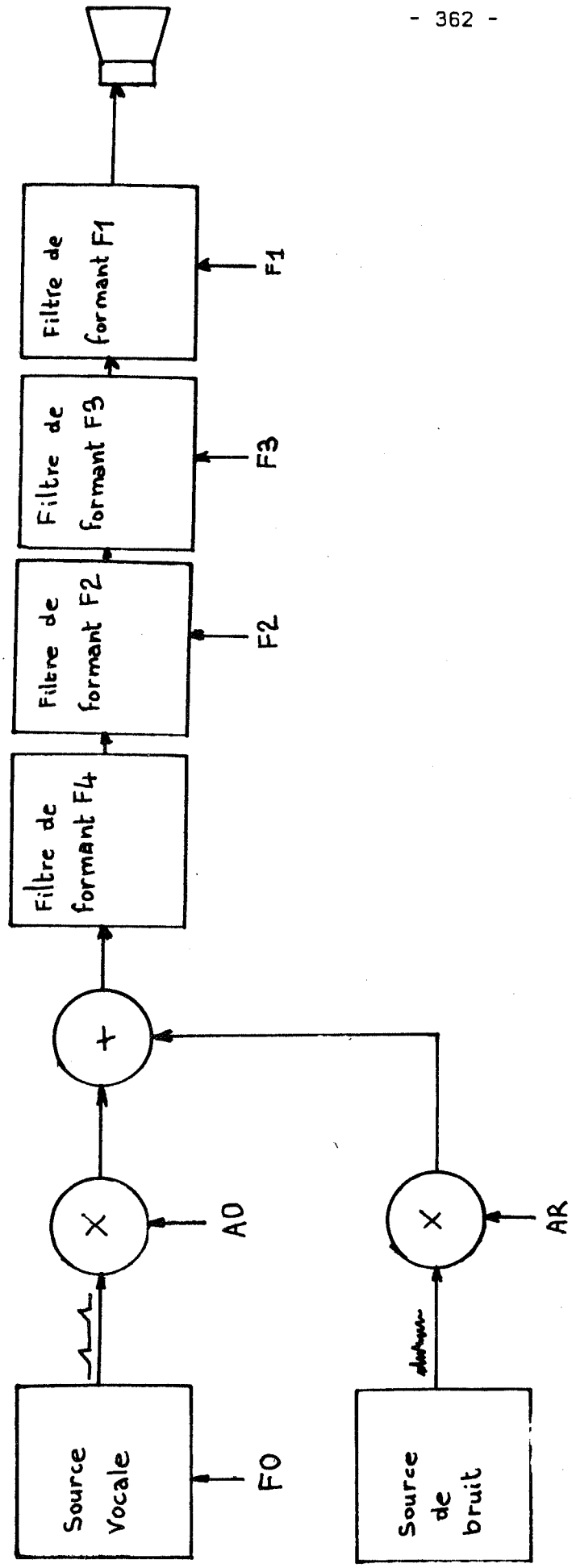


Figure 1 - Schéma du synthétiseur à formants série équivalent

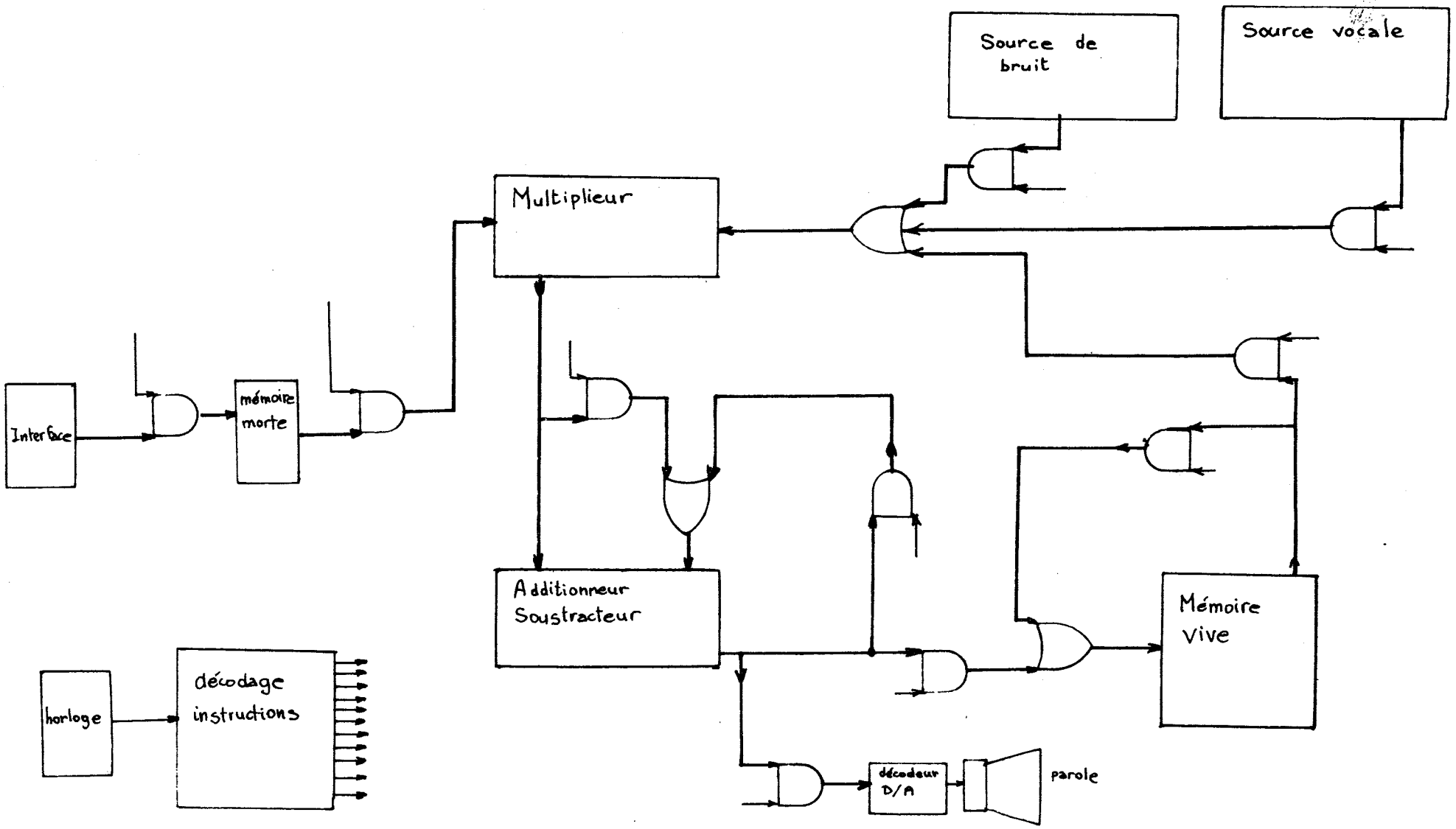


Figure 2 _ Schéma synoptique du synthétiseur

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

SYRUS

SYNTHESE, SUR UN MINI-ORDINATEUR, DU SIGNAL VOCAL DANS SA REPRESENTATION AMPLITUDE - TEMPS

X. RODET -- C. SANTAMARINA

RESUME

La parole est étudiée dans sa représentation amplitude - temps. Un ensemble de programmes permet l'enregistrement de la parole, son examen sur une console de visualisation et sa synthèse à partir d'un texte littéral en calculant la mélodie, l'accent et le rythme.

SUMMARY

Speech is studied in its amplitude - time representation. We describe a system which permits speech mémorisation, graphic display and synthesis by rule. Its calculates the fondamental frequency pattern and some prosodic features.

S Y R U S

SYNTHESE, SUR UN MINI-ORDINATEUR, DU SIGNAL VOCAL DANS SA REPRESENTATION AMPLITUDE - TEMPS

X. RODET -- C. SANTAMARINA

I / INTRODUCTION

Cette étude vise à la réalisation d'une unité à réponse vocale utilisant un minimum d'appareillage analogique et une capacité de mémoire aussi réduite que possible. Le système pourrait ainsi être mis en oeuvre sur de nombreux calculateurs. Nous avons donc traité le signal dans sa représentation amplitude - temps et écrit les programmes sur un mini-ordinateur MULTI 20. Les bases de la méthode de synthèse par régles, que nous avons élaborée ont été exposées aux 5émes Journées d'Etude sur la Parole à ORSAY. (1).

II / REPRESENTATION AMPLITUDE - TEMPS

Contraitement aux méthodes les plus courantes fondées sur une analyse en fréquence, un signal est représenté sous forme d'une suite de doublets (Δ_{ti}, A_i) où Δ_{ti} est l'intervalle de temps entre deux extrémums et A_i l'amplitude correspondante (fig. 1). Les Δ_{ti} sont comptés en nombre de périodes d'une horloge de référence. Les A_i sont discrétisés sur un faible nombre de niveaux.

III / MOYENS DISPONIBLES

Nous travaillons sur un mini-ordinateur MULTI 20 muni d'une mémoire de microprogrammes altérable (A. R. O. M.), relié à une console de visualisation et à un périphérique vocal (fig. 2). Ce dernier se compose d'un codeur pour l'enregistrement de la parole et d'un décodeur pour la synthèse. Le codeur effectue la conversion analogique digitale du signal issu de microphone. Il échantillonne le signal à 10 KHz et note les Δ_{ti} sur 4 bits. Les amplitudes sont discrétisées sur 15 niveaux seulement. Le décodeur transforme la suite des Δ_{ti}, A_i en un signal analogique par une simple interpollation linéaire entre les extrémums.

Le codage constitue une réduction très importante de la quantité d'information. Le signal demeure compréhensible, mais la qualité s'en ressent : le nombre de niveaux réduit la dynamique du signal à 12 db environ. Aussi nous contruisons un nouveau codeur à 64 niveaux et à fréquence d'échantillonnage plus élevée (jusqu'à 14 KHz). Cela permettra de réduire le bruit restitué, et l'imprécision sur la description du signal.

IV / METHODES

Nous rappelons brièvement la métnode de synthèse par règles (1).

Le signal est considéré comme une succession de zones quasi-stables voisées ou bruitées et de transitions. Les éléments voisés sont constitués par répétition d'une pseudo-période spécifique pour chaque phonème. (Une pseudo-période est définie comme la courbe enre-

.../...

gistrée du signal vocal pendant une période des cordes vocales). Fig. 3. Les variations de la fréquences fondamentale sont réalisées par troncature ou extension de la pseudo-période. Les éléments bruités sont calculés grâce à un tirage aléatoire, dans une courte séquence enregistrée. Pour les transitions, la pseudo-période du phonème voisé est déformée pour rendre compte des évolutions habituellement décrites en terme de variation de la fréquence fondamentale et variations des fréquences et amplitudes des formants.

La pseudo-période est donc séparée actuellement en deux composantes, l'une C_1 n'ayant pas de fréquence supérieure au premier formant, l'autre C_2 rendant compte des fréquences supérieures. (fig. 4). Ces deux courbes subissent séparément deux affinités, l'une suivant l'axe des amplitudes, l'autre suivant l'axe des temps. Chaque pseudo-période de la transition est la somme de ces deux courbes. Les paramètres d'anamophose évoluent au cours de la transition.

En raison de la fréquence d'échantillonnage et du nombre de niveaux choisis, la description de ces deux courbes est imprécise. De plus les formants d'ordre supérieur à un ne sont pas traités différemment les uns des autres. Nous espérons pallier à ces inconvénients avec le nouveau codeur en construction.

V / LE SYSTEME

Suivant cette méthode nous avons écrit le programme SYRUS.

- SYRUS transforme un texte littéral en parole synthétique (fig. 5).
- SYRUS permet l'enregistrement et l'examen de la parole sur une console de visualisation graphique.
- SYRUS permet ainsi l'étude des méthodes de synthèse et la mise au point des paramètres.

1) Transcription phonétique.

Le programme découpe au préalable le texte littéral en "mots" en fonction des blancs et des signes de ponctuation. Il rend compte de règles élémentaires de la prononciation pour arriver à une écriture phonétique tenant compte des liaisons. Les exceptions se trouvent dans un dictionnaire sous forme de "racines" ou de mots entiers.

2) Mélodie et accent.

En fonction de la longueur de la phrase, nous établissons un schéma intonatif standard de forme très simple (2), (7).

.../...

Un "accent tonique" a été introduit sur la dernière syllabe des phrases, en prolongeant la durée de la voyelle.

3) Construction du signal.

Pour chaque phonème SYRUS construit successivement la partie quasi-stable et la transition vers le phonème suivant. Il superpose sur la ligne mélodique calculée précédemment une micro-mélodie. Elle tient compte de la nature de la partie quasi-stable (consonantique ou vocalique), de la nature de la transition, du degré d'ouverture du phonème. De même SYRUS établit le rythme de l'énoncé en tenant compte des pauses, de l'accent et de la nature des phonèmes.

Le vocabulaire est constitué :

- des pseudo-périodes des phonèmes voisés.
- des segments de phonèmes bruités.
- des paramètres de transition.

Nous l'avons construit en cherchant à optimiser la qualité et la compréhension des messages synthétisés.

L'ensemble des programmes, vocabulaire compris, ne dépasse pas 16 K octets de mémoire vive et 2 K octets d'A. R. O. M.

VI / RESULTATS

La synthèse en temps réel est largement assurée car plusieurs modules ont été microprogrammés, ainsi le temps de calcul d'un message est à peu près le dixième de sa durée d'émission.

Actuellement la qualité du message souffre beaucoup du trop faible nombre de niveaux d'amplitude. Cependant, un auditeur habitué à la "voix" de SYRUS reconnaît des mots français isolés presque sans erreur. Mais l'énoncé présente un caractère "haché" peu naturel. Ce qui tient essentiellement à la limitation du contexte lors du calcul des zones stables et des transitions.

VII / DEVELOPPEMENTS

La qualité de l'émission et la précision des motifs seront améliorés grâce au nouveau codeur. Par ailleurs, pour pallier au caractère "haché" de l'énoncé nous cherchons d'une part à étendre le contexte dans lequel est calculé chaque phonème, d'autre part à effectuer un lissage des variations de paramètres au cours d'une phrase. Des notions de coarticulation pourront être introduites.

.../...

R E F E R E N C E S

(1) RODET X.
Une méthode de synthèse par règles du signal vocal dans sa représentation amplitude - temps.

Compte rendu des 5émes Journées d'Etudes du Groupe "Communication parlée" - Vol. I - ORSAY, Mai 1974.

(2) LARREUR D. et BOË L.J.
Synthèse paramètrique de la phrase énonciatrice en français.

Compte rednu des 5émes Journées d'Etudes du Groupe "Communication parlée" - Vol. II - ORSAY, Mai 1974.

(3) BOË L.J. et LARREUR D.
Les caractéristiques intrinséques de la fréquences laryngienne : production, réalisation et perception.

Compte rendu des 5émes Journées d'Etudes du Groupe "Communication parlée" - Vol. I - ORSAY, Mai 1974.

(4) SANTERRE L.
Transitions articulatoires et transitions acoustiques dans la parole réelle.

Compte rendu des 3émes Journées d'Etudes sur la Parole.
Mai-Juin 1972 - C.N.E.T. LANNION.

(5) KOZLENKO N.I. and RYZHKOVA R.N.
Articulation characteristics of extremally coded speech.

SOV. Phys. Acoust. - Vol. 19 n° 3 - Nov-Dec 1973

(6) LIENARD J.S. et TEIL D.
Les éléments phonétiques et la traduction automatique du message écrit en message parlé.

Automatisme Tome XV n° 10 - Octobre 1970

(7) VAISSIERE J.
Fréquences fondamentales des phrases déclaratives en français

Compte rendu des 5émes Journées d'Etudes du Groupe "Communication parlée" - Vol. I - ORSAY, Mai 1974.

(8) STRAKA G.
Album phonétique

Les Presses de l'Université Laval. QUEBEC 1965.

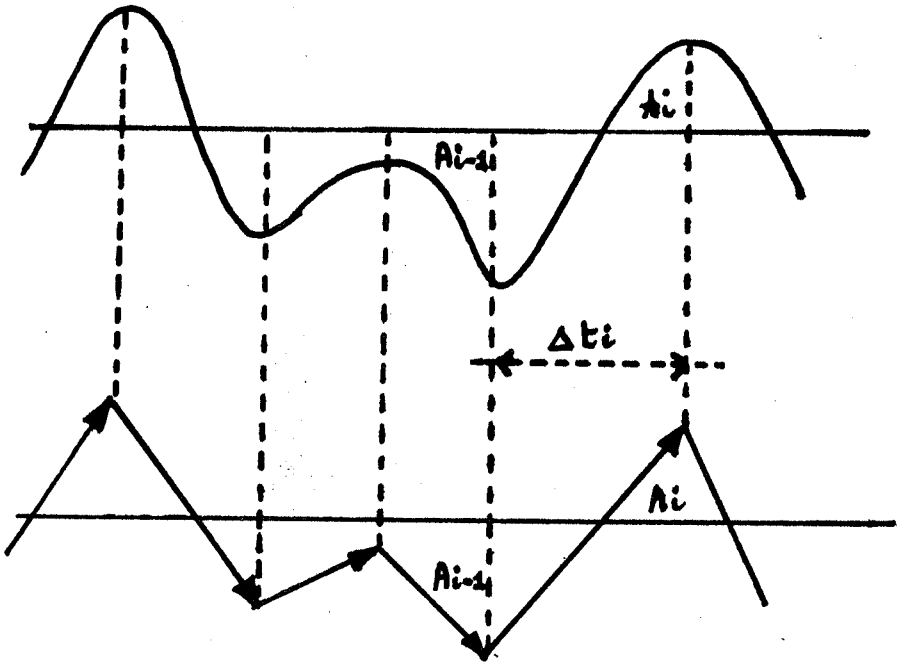

Ai
Ai-1
Δti
Ai
Ai-1

Fig 1

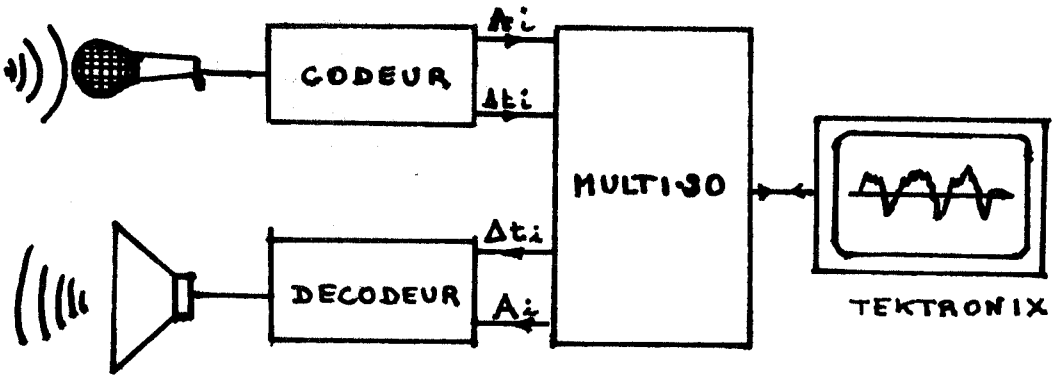

Ai
CODEUR
Δti
MULTIPLEX
Δti
DECODEUR
Ai
TEKTRONIX

Fig 2

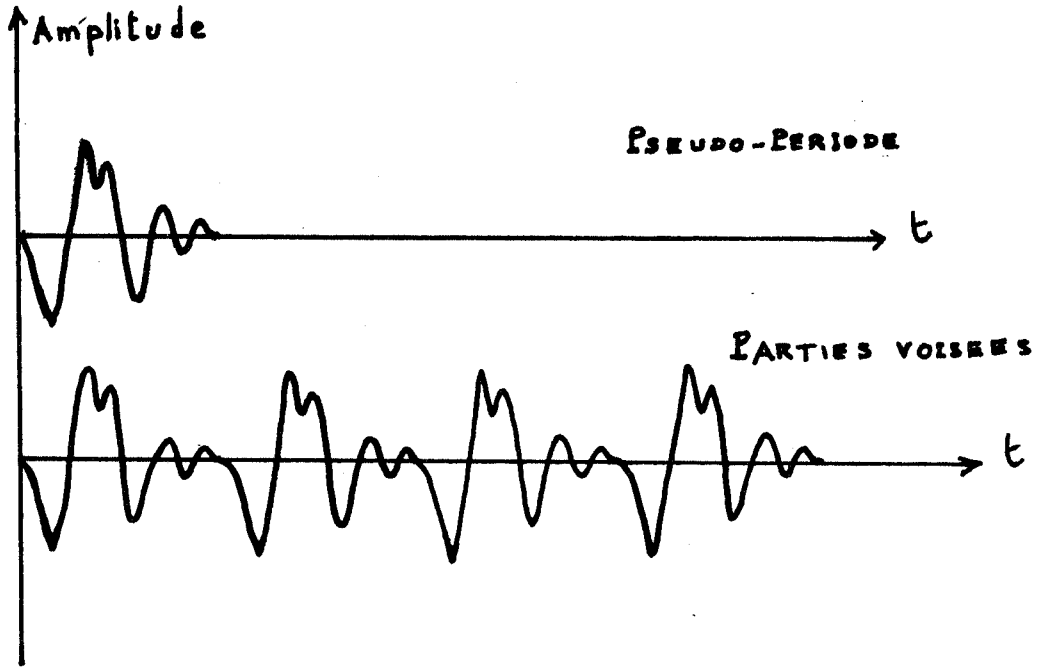

Amplitude
Pseudo-Periode
t
Parties voisées
t

Fig 3

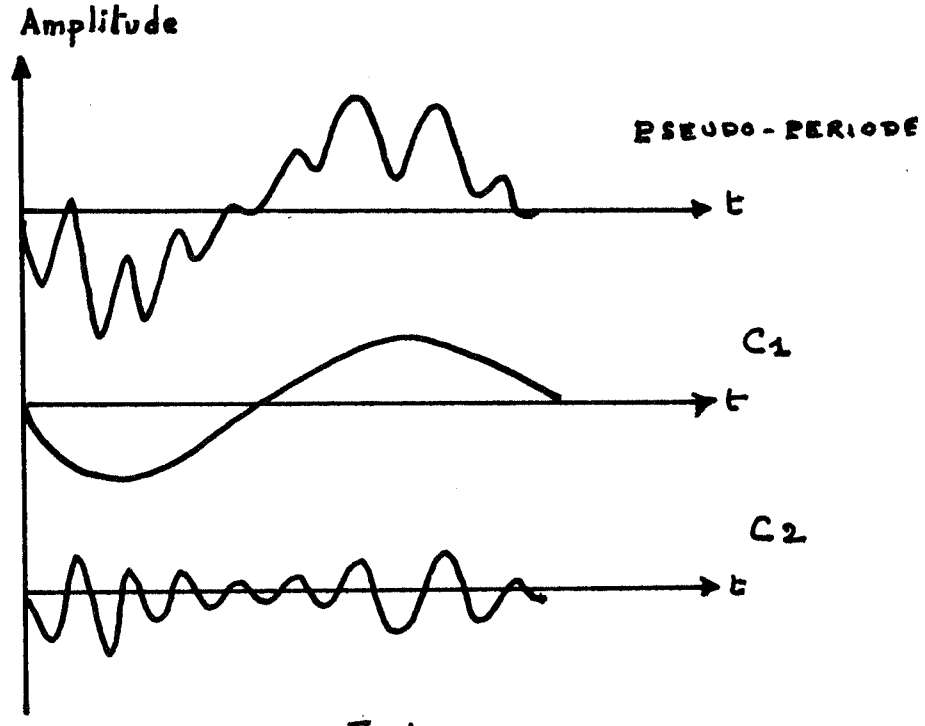

Amplitude
Pseudo-Periode
t
C1
t
C2
t

Fig4

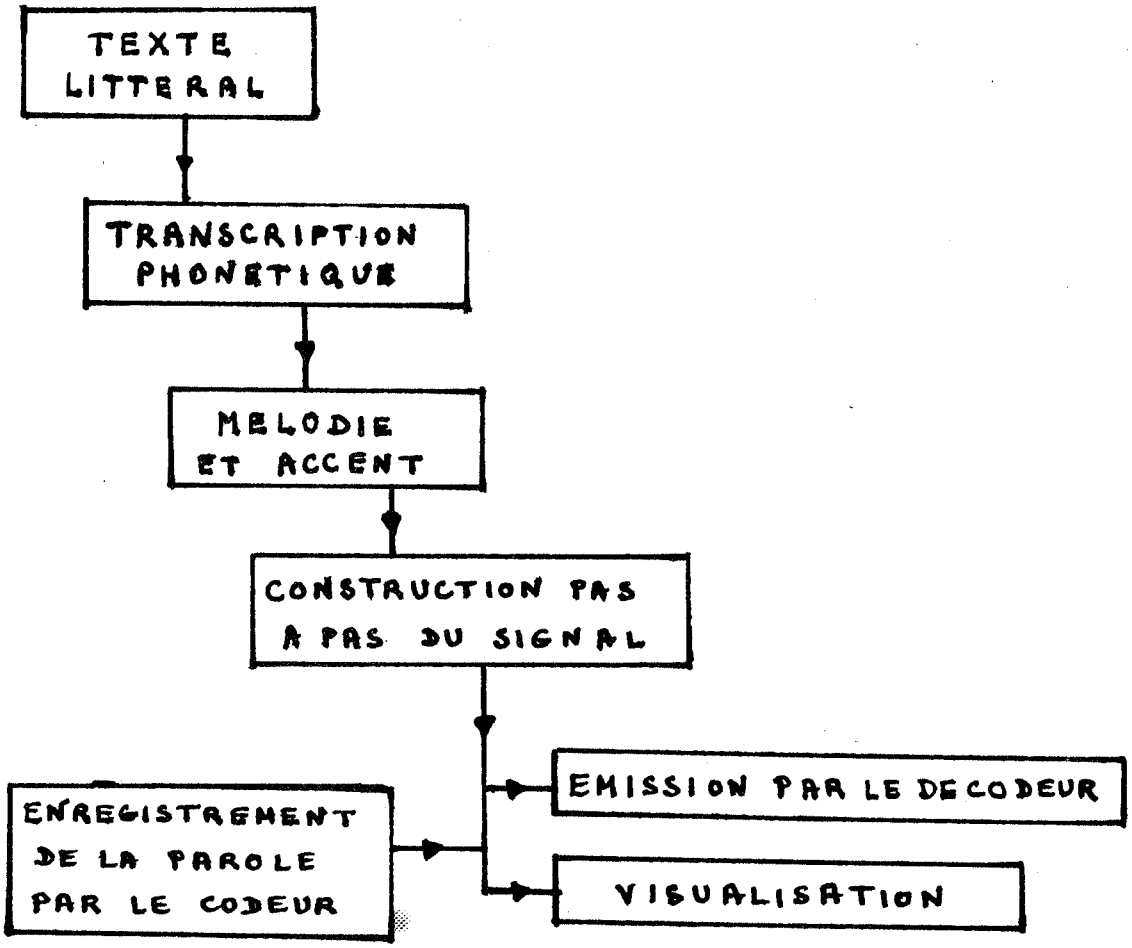

TEXTE LITTERAL
TRANSCRIPTION PHONETIQUE
MELODIE ET ACCENT
CONSTRUCTION PAS A PAS DU SIGNAL
EMISSION PAR LE DECODEUR
VISUALISATION
ENREGISTREMENT DE LA PAROLE PAR LE CODEUR

Fig 5

THEME 3C

COMMANDE DES PROCESSUS

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UTILISATION DE LA METHODE DES TRAJECTOIRES DANS UN SYSTEME DE RECONNAISSANCE VOCALE A VOCABULAIRE LIMITE

- - - -

L.F. PAU, M. LAGNEAU, Laboratoire de Théorie des Systèmes, E.N.S. des Télécommunications, 46 rue Barrault - 75634 PARIS CEDEX 13.

RESUME

Utilisant de manière convenable des méthodes de compression du signal, on représente des segments de parole continue par des trajectoires orientées à 2-3 dimensions. Une mesure de similarité entre trajectoires permet de faire la reconnaissance de ces commandes vocales sans segmentation préalable. Les méthodes de filtrage et lissage de trajectoires, particulièrement importantes pour cette application, sont décrites.

SUMMARY

By using adequately signal compression methods, it is possible to represent segments of continous speech by oriented trajectories in 2-3 dimensions -Vocal command recognition is performed without speech segmentation thanks to a measure of similarity between such trajectories- Filtering and smoothing methods for trajectories, of particular importance for this application, are described.

UTILISATION DE LA METHODE DES TRAJECTOIRES DANS UN SYSTEME DE RECONNAISSANCE VOCALE A VOCABULAIRE LIMITE

L.F. PAU, M. LAGNEAU.

= = = = =

1. PRINCIPE DE LA METHODE

On rappelle que les méthodes d'analyse des données, sans professeur (telles que l'analyse en composantes principales, et ses variantes, ou l'analyse factorielle des correspondances), peuvent être appliquées à la suite ordonnée des vecteurs obtenus en échantillonnant le signal vocal (ou un sous-produit de celui-ci) à cadence fixe ou modulée. La suite ordonnée des points représentatifs de ces vecteurs dans un espace de dimension réduite (par exemple, 2 ou 3) permet de représenter un enregistrement de parole sous la forme de la trajectoire orientée joignant ces points entre eux dans l'ordre imposé [2,4,5,6]

La reconnaissance de vocabulaires limités est l'un des sous-produits de ce procédé ; elle est basée sur la comparaison, au sens d'une mesure de similarité, de la trajectoire du segment de parole à classer avec chacune des trajectoires associées à chacun des segments de parole appris. Antérieurement, on a obtenu un taux de reconnaissance de 83 % dans un vocabulaire de 215 mots isolés segmentés en phonèmes [4]. Plus récemment, on a obtenu un taux de reconnaissance provisoirement de l'ordre de 88 % dans un ensemble de 15 commandes vocales courtes du type : langage géométrique pour la commande de robots ; la paramétrisation est ici celle d'un vocodeur simple, et l'échantillonnage est à période fixe de 10 ms ; on a également étudié la sensibilité vis-à-vis de l'apprentissage [2].

Le grand avantage de ce procédé est sa rapidité, et sa stabilité, alliées au fait que la reconnaissance se fait sans segmentation du signal vocal.

2. PHASE D'APPRENTISSAGE

Le but de l'apprentissage est d'appliquer au tableau des vecteurs-formes, décrivant les échantillons successifs du signal vocal, une méthode d'analyse des données sans professeur. L'ensemble des formes d'apprentissage est identique à l'ensemble des vecteurs-échantillons obtenus à partir de l'enregistrement des mots du vocabulaire (éventuellement répétés).

Dans la suite, X (t) désignera le vecteur échantillon prélevé à l'instant t, après numérisation de la sortie du vocodeur ; X(t) est un vecteur-ligne.

2.0. Matériel

Le matériel mis à la disposition de ce projet est le suivant:

../..

1) Calculateur CII 10070

2) Calculateur hybride EAI 640 pour le traitement du signal et sa numérisation au moyen de 15 convertisseurs AD

3) Vocodeur à N = 15 canaux, de faible coût, à base de filtres élémentaires du second ordre (THAI-SEREY [7]), la bande passante couverte est [210 - 3930 HZ] ; des modifications ultérieures de ce matériel ont conduit à un meilleur choix des fréquences centrales des filtres, en particulier vers les fréquences hautes, voir aussi 2.1.

2.1. Lissage des trajectoires en hardware. Il y a deux possibilités :

2.1.1. Le signal vocal est filtré par un filtre passe-bande du 1er ou 2nd ordre, dont la sortie est ensuite passée dans un redresseur (fournissant la valeur moyenne du signal filtré) ; la sortie du redresseur est enfin appliquée à un filtre passe-bas du 1er ordre.

2.1.2. Le signal vocal est d'abord redressé, puis filtré par un filtre passe-bande du 1er ou 2nd ordre.

2.1.3. Des deux lissages 2.1.1. et 2.1.2., ce dernier est le plus intéressant et de loin. Dans [2], on a utilisé 2.1.1.

2.2. Numérisation

Le problème s'est posé de câbler les portes et les bascules nécessaires à la synchronisation des conversions AD des niveaux d'énergie sortant du vocodeur, comme définis en 2.1. La solution suivante fut adoptée :

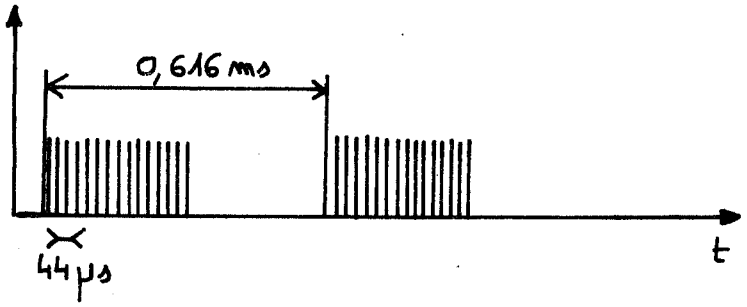


Figure 1 : Numérisation et échantillonnage.

Les signaux de sortie sont échantillonnés à une cadence de T = 10 ms, et numérisés sur 16 bits (longueur du mot machine sur calculateur EAI).

L'intérêt de cette numérisation est de rendre possible l'insertion de calculs annexes (analogiques ou numériques) dans la fenêtre inutilisée.

2.3. Assemblage des données numérisées

Les données numérisées X(t) représentent la distribution des niveaux d'énergie moyens sur les canaux du vocodeur.

Ici encore, les données donnant lieu à la représentation de trajectoires, peuvent être assemblées de deux manières différentes :

../..

2.3.1. Assemblage simple

Les tableaux constitués $U = \{U(i,j) ; i \in I, j \in J\}$ sont définis par :

I : vecteurs échantillons successifs X(t), rangés dans l'ordre du temps, avec la période T

J : mesures fournies par 2.2.

$U(i,\bullet) = X(iT)$

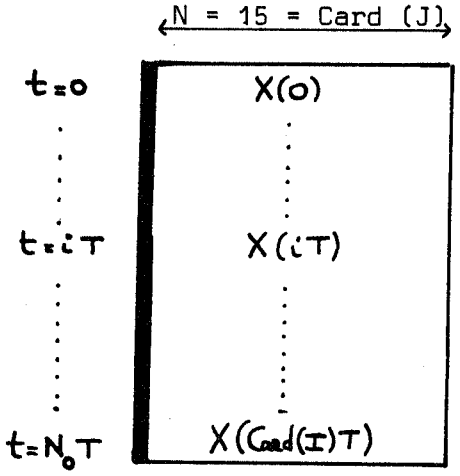


Figure 2 : Tableau U

dim U = N x Card (I)
N_0 = nombre d'échantillons
N_0T = durée totale des enregistrements.

2.3.2. Assemblage en vue du filtrage et lissage numérique (suite en 2.5.2.)

Si on considère un filtre numérique d'ordre k, les tableaux constitués $V = \{V(i,j) ; i \in I ; j \in J\}$ sont définis par :

I : vecteurs-échantillons étendus successifs y(t), rangés dans l'ordre du temps, avec la période kT
y(t) est donc obtenu en mettant côte à côte k vecteurs échantillons successifs, et en les représentant par un seul vecteur y(t)

J : mesures fournies par 2.2., chacune répétée k fois avec la période N par définition de y(t)

$V(i,\bullet) = y(ikT)$

$y(t) \triangleq [X(t) \quad X(t+T) \ldots\ldots X(t+(k-1)T)]$

../..

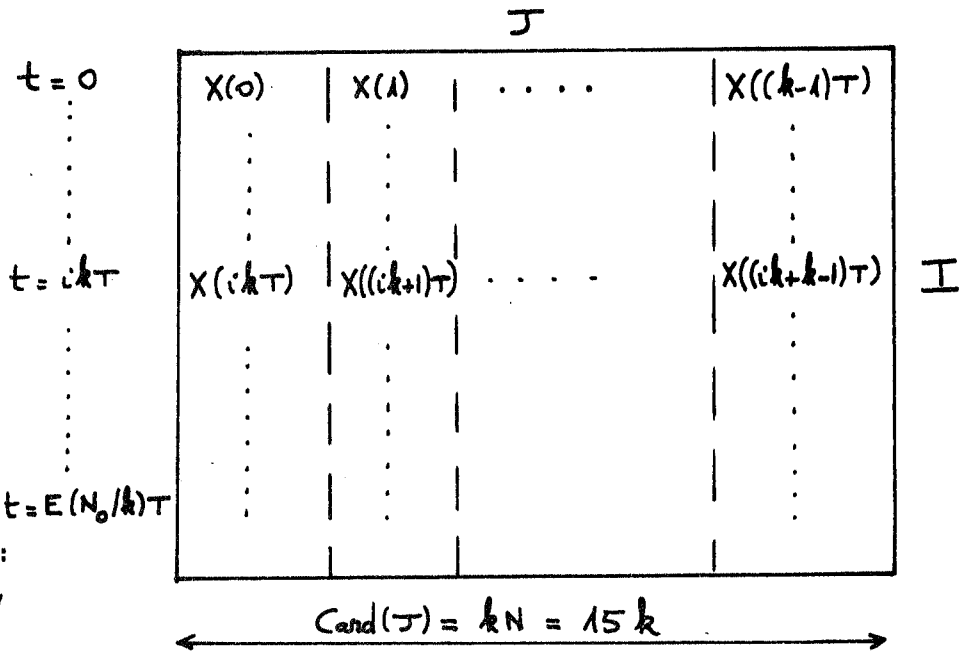


Figure 3 :

Tableau V

Voir aussi 2.5.2.

2.5. Compression des données

Selon le procédé adopté, on appliquera à l'un ou l'autre des tableaux U ou V , désignés globalement par $W = \{ w(i,j) ; i \in I, j \in J \}$ l'analyse en composantes principales normalisée, l'analyse des correspondances [1,2], ou l'analyse des correspondances paramétrisée [6].

Dans les 3 cas, le résultat sera un ensemble de paramètres, par lesquels un vecteur-échantillon w(i,•) peut être projeté dans un espace de dimension réduite $r \leq \text{Inf (Card (I), Card (J))}$,

grâce à une formule du type :

$$z(i,q) = \sum_{j=1,\ \text{Card}(J)} u_{jq}\, f(w(i,j) ; w(i,\cdot) ; W) \qquad (1)$$

$q = 1,\ldots,r$

Cette formule sera généralement linéaire .

2.5.1. Cas 1 : Analyse des correspondances

$$z(i,q) = \sum_{j=1,\ \text{Card}(J)} u_{jq}\, \frac{P_{ij}}{P_{i\cdot}\sqrt{P_{\cdot j}}} \qquad (2)$$

où | P : matrice de contingence associée à W

| $u_{\cdot q}$: q - ème vecteur propre normé associé à la valeur propre λ_q, selon les notations standard.

C'est cette formule qui sert dans [2].

../..

2.5.2. Cas 2 : Analyse des correspondances paramétrisée [6], avec choix du filtre numérique (faisant suite à 2.3.2.)

On rappelle que l'analyse des correspondances paramétrisée consiste à remplacer la loi marginal P.j associé à W par une distribution essentielle H.j ([6]), et à refaire la théorie de l'analyse des correspondances avec H.j. Ceci interdit la représentation simultanée de I et J, mais permet, par contre, de pondérer différemment les échantillons successifs X(t) de Y(t) (et leurs composantes) et de synthétiser par là-même un filtre numérique d'ordre k à coefficients positifs.

$$z(i,q) = \sum_{j=1,\,Card(J)} u_{jq} \, f\left(w(i,j)\, h_{.j} \; ; \; w(i,.) \; ; \; W\right) \qquad (2)$$

2.6. Résultats de l'apprentissage

Pour un réglage donné du dispositif, le résultat de l'apprentissage est constitué par les r x Card(J) paramètres u_{jq} de (1), servant au réglage du pont potentiométrique qui correspond au câblage de (1). L'utilisation de la compression des données évoquée au 25, nécessite l'emploi d'algorithmes de calcul de spectres de matrices, comme expliqué en Annexe.

3. PHASE DE RECONNAISSANCE

Le but de la reconnaissance est de comparer la trajectoire d observée, aux trajectoires d'apprentissage d(m) associées aux mots m = 1,..., M du vocabulaire, et de classer au moyen d'une mesure de similarité qui doit s'appliquer à des trajectoires orientées en fonction du temps.

$$d(m) \triangleq \{ z\,(i,\bullet) \;,\; i \in \text{(mot d'apprentissage m)} \}$$

$$d \triangleq \{ z_d(i,.) \;,\; i \in \text{mot inconnu} \}$$

$$z_d(i,q) = \begin{cases} \displaystyle\sum_{j=1,\,Card(J)} u_{jq} \, f\left(w(d,j) \; ; \; w(d,.) \; ; \; W\right) \\ \text{ou :} \\ \displaystyle\sum_{j=1,\,Card(J)} u_{jq} \, f\left(w(d,j)\, h_{.j} \; ; \; w(d,.) \; ; \; W\right) \end{cases}$$

où : W est l'ensemble des données d'apprentissage utilisé.

3.1. Choix du vocabulaire

Dans [2], on a pris M = 15 mots prononcés 4 fois par le même locuteur, afin de réaliser 4 ensembles indépendants de données d'apprentissage. En regard de chaque mot se situe le nombre d'échantillons X(t) à la période T = 10 ms :

1.	Le rond	78, 78, 74,77
2.	Le carré	88, 98, 93, 92
3.	Le losange	107, 105, 108, 122
4.	Le vecteur	140, 102, 132, 114
5.	A droite	71, 74, 72, 72
6.	A gauche	87, 87, 91, 95
7.	En bas	66, 74, 69, 72

../..

8. Au dessus	83, 86, 84, 90
9. Faire	47, 43, 47, 45
10. Une rotation	110, 118, 110, 115
11. Une symétrie	107, 109, 104, 107
12. Agrandir	101, 97, 93, 92
13. Diminuer	89, 97, 92, 88
14. Déplacer	92, 104, 98, 89
15. Incliner	75, 69, 77, 70

3.2. Traitement des silences

Bien entendu, le début et la fin d'un mot donnera un point identique $z_d(i,.) \equiv S$ correspondant au silence, c-a-d que les trajectoires d, d(m), m = 1,... M sont en principe fermées. Ceci étant préjudiciable à la classification (qui exploite les écarts entre trajectoires), on est amené à prendre deux mesures :

3.2.1. Seuil sur l'énergie cumulée sortant des N canaux du vocodeur

Les trajectoires seront tronquées, afin de ne comprendre que les seuls échantillons pour lesquels le cumul des sorties numérisées de 2.2. dépasse un seuil fixé. 3.2.1. doit être appliqué à l'apprentissage 2.5. également.

3.2.2. Transformation conforme de centre S

On transforme les trajectoires d et d(m) dans une anti-inversion de pôle S et de rapport $(-a^2)$ donné, afin de conserver les angles entre trajectoires, et d'envoyer à distance respectable les points correspondant à un signal vocal proche du silence. On sait que ces branches de courbe ne seront pas prises en compte dans le calcul de la mesure de similarité entre trajectoires orientées, comme expliqué dans [3]. Les trajectoires transformées sont encore notées d.

3.3. Mesure de similarité entre trajectoires

On rappelle les mesures étudiées dans [2,3]. Leur principe est de discriminer les mots d'après la différence des transitoires entre phonèmes, tant au point de vue distance qu'orientation ("ab" est différent de "ba"). En vue de tests rapides, on a choisi une mesure de similarité entre trajectoires non orientées, basée sur la distance L^1 entre points :

$$D(d, d(m)) = \frac{1}{\mathrm{Card}(I)} \sum_i \left\| z(i,.) - z_d(i,.) \right\|_{L^1(\mathbb{R}^2)}$$

3.4. Taux de reconnaissance [2] Voir matrice de confusion de la figure 4 , où W est la liste 1 de M mots, et les mots à reconnaitre sont ceux de la liste 2.

3.5. Temps unitaire de reconnaissance [2] N = M = 15 CII 10070

../..

FIGURE 4

MATRICE DE CONFUSION CORRESPONDANT A LA LISTE D'APPRENTISSAGE N°1

MOTS RECONNUS \ MOTS D'APPRENTISSAGE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	100%														
2		100%													
3			66%	33%											
4				0%											
5					100%										
6						100%				33%					
7							100%								
8								100%							
9									0%L2 100%	L1					
10										66%					
11											100%				
12									100% 0%L1	L2		100%			
13			33%										100%		
14				66%										100%	
15															100%

$$\text{Taux de reconnaissance avec L1} = \frac{\text{Eléments diagonaux}}{15} = 88\ \%$$

Taux de reconnaissance avec L2 =82 %

Secondes CPU/mot	Espace $\mathbb{R}^k$ = r
0,37 s	2
0,48 s	3
0,62 s	5

3.6. Lissage et simplification des trajectoires

Il est bien évident que si la structure géométrique d'une trajectoire est simplifiée (approximée ou lissée, le temps unitaire de reconnaissance est réduit, mais aux dépends du taux de reconnaissance.

3.6.1. Trajectoire simplifiée en prenant 1 point sur n de d et d(m)

r = 2

1 point sur n: n=	Secondes CPU/ mot	Taux de reconnaissance
2	0,20 s	88 %
3	0,17 s	88 %
4	0,14 s	65 %
5	0,10 s	65 %

3.6.2. Détermination de l'enveloppe des trajectoires lissées

Le temps de calcul nécessaire au calcul de l'enveloppe est trop grand.

3.6.3. Suppression des redondances de points et squelettisation

Etant donné une trajectoire d = $\{ z_d(i,\cdot)$, $i \in$ mot inconnu $\}$, et $\varepsilon<0$, on supprime les échantillons (j) de d tels qu'il existe un point précédant $z_d(i,\cdot)$ de d vérifiant $\| z_d(j,\cdot) - z_d(i,\cdot) \|_{L^1(\mathbb{R}^n)} < \varepsilon$.

Cette méthode de squelettisation ne supprime pas l'information, car si deux points sont confondus, alors l'un est redondant ; ceci est le cas pour les échantillons successifs d'un même phonème à l'intérieur d'un mot. Pour 0,003 < ε < 0, 06, les taux de reconnaissance ont été inchangés, mais le temps global de reconnaissance a légèrement augmenté.

4. STABILITE DE L'APPRENTISSAGE ET DE LA RECONNAISSANCE

L'utilisation pratique de la formule de projection $z(i,q)$ (2) demande que le système de vecteurs $\{ u_{iq} \}$ pour une matrice associée à un ensemble de données issues d'un même locuteur soit invariant, quelque soient les données.

Une étude expérimentale a été réalisée en deux étapes :

- étude de la stabilité de l'apprentissage
- étude de la stabilité de la reconnaissance

../..

4.1. Stabilité de l'apprentissage

Cette étude a porté essentiellement sur la liste de 15 mots prononcée 4 fois par le même locuteur. On a extrait, pour ces 4 analyses, 4 systèmes de vecteurs propres dont on a comparé les directions.

Les résultats numériques sont les suivants :

Figure 5 = Vecteurs propres pour 4 listes du même locuteur (2).

LISTE 1		LISTE 2		LISTE 3		LISTE 4	
$u_{.1}^{(1)}$	$u_{.2}^{(1)}$	$u_{.1}^{(2)}$	$u_{.2}^{(2)}$	$u_{.1}^{(3)}$	$u_{.2}^{(3)}$	$u_{.1}^{(4)}$	$u_{.2}^{(4)}$
.24971	-.70801	.26427	-.69002	.28780	-.71395	.29531	-.7160
.42084	-.23103	.37779	-.25592	.41745	-.18439	.36515	-.2020
.28884	.41413	.25103	.40328	.25459	.41641	.26807	.3561
.20074	.40863	.21287	.43156	.18856	.41255	.22112	.4463
.08613	.17616	.12990	.16741	.10186	.17681	.11548	.1852
.07342	.22234	.08769	.22820	.07196	.22893	.08857	.2512
-.05840	.09430	-.03445	.10741	-.06194	.09440	-.05577	.0960
-.12439	.05070	-.11730	.06821	-.13168	.04803	-.13322	.0523
-.19265	.03843	-.20215	.03117	-.20509	.01899	-.18707	.0190
-.21972	.02098	-.21932	.03420	-.23823	.01830	-.22293	.0141
-.29371	.02219	-.24724	.00880	-.29710	-.00193	-.25424	.0054
-.29987	-.04834	-.30665	-.04934	-.29773	-.05848	-.32456	-.0514
-.28710	-.06997	-.34194	-.06696	-.28681	-.07178	-.32788	-.0602
-.35285	-.05899	-.39268	-.04765	-.35360	-.06531	-.35992	-.0447
-.36963	-.04282	-.35266	-.04284	-.35043	-.06540	-.35038	-.0454

On a calculé $\cos\left(u_{.1}^{(i)}, u_{.1}^{(j)}\right)$ pour $(i) \neq (j)$

$\cos\left(u_{.2}^{(i)}, u_{.2}^{(j)}\right)$ $i = 1, \ldots, 4$

$j = 1, \ldots, 4$

Tous ces cosinus sont expérimentalement compris entre 0,98 et 1, ce qui est assez satisfaisant pour conclure à l'invariance des vecteurs propres pour le même locuteur. De plus, les trajectoires correspondantes aux mêmes mots de chaque liste sont pratiquement identiques.

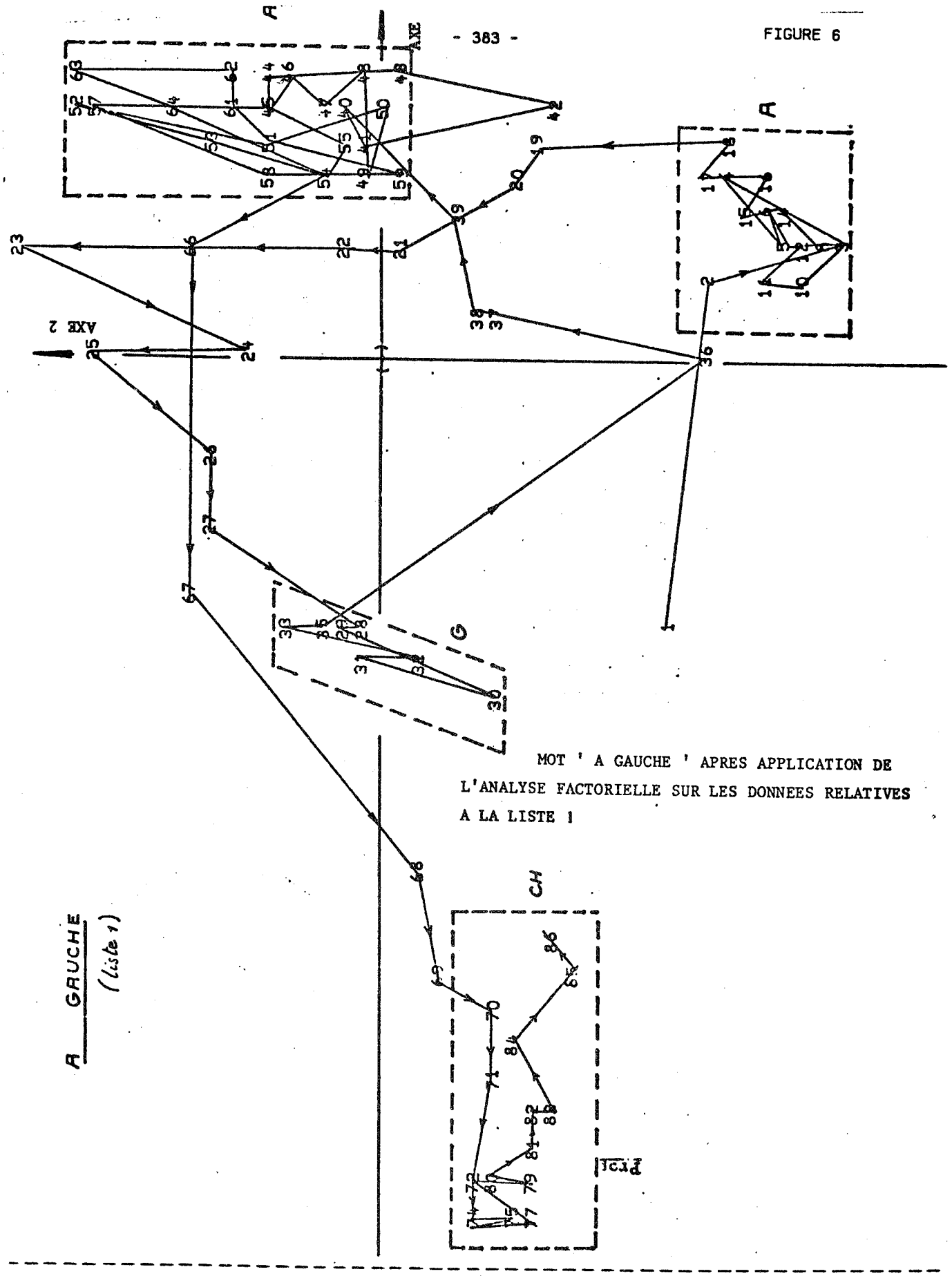
4.2. Stabilité de la reconnaissance

Soit une nouvelle liste de mots prononcée par le même locuteur

1 - ATTENDS	54	11 - UNE MAISON	55
2 - UN CAILLOU	63	12 - UNE NOIX	49
3 - LE CRAYON	67	13 - GENTIL	58
4 - LA POUPEE	65	14 - LE FUTUR	61
5 - L'ECOLE	55	15 - CHOCOLAT	60

../..

FIGURE 6



MOT ' A GAUCHE ' APRES APPLICATION DE L'ANALYSE FACTORIELLE SUR LES DONNEES RELATIVES A LA LISTE 1

6 - UN OISEAU	64	16 - UNE CHAISE	61
7 - DES GANTS	51	17 - CONTENT	51
8 - PEUREUX	40	18 - FATIGUE	55
9 - UN VELO	55	19 - UNE VOITURE	68
10 - UN AVION	67	20 - UN CHAT	45

Il a été vérifié expérimentalement les deux points suivants

a) Si on appelle $\vec{w}_i$ les vecteurs propres calculés dans cette analyse et $\vec{v}_i$ les vecteurs propres calculés précédemment, alors les calculs ont montré que $0{,}98 < \cos(\vec{v}_i, \vec{w}_i) < 1$

b) La projection des 15 mots (vus précédemment) d'un locuteur quelconque parmi les 4 dans les axes $\vec{w}_i$ fournissent approximativement les mêmes facteurs; par conséquent nous retrouvons les mêmes trajectoires.

REFERENCES

[1] J.P. BENZECRI et al, l'analyse des données, Dunod, Paris, 1973

[2] M. LAGNEAU, Application d'une méthode statistique à la reconnaissance d'un vocabulaire limité, Mémoire C.N.A.M., E.N.S.T., septembre 1974.

[3] L.F. PAU, Méthode des trajectoires, Journées GALF d'étude sur la parole, LIMSI, Orsay, 15-17 mai 1974

[4] L.F. PAU, Statistical reduction of speech patterns, in : Machine perception of patterns and pictures, Institute of Physics, Conf. Puble n° 13, Londres, pp 126 - 133, avril 1972.

[5] L.F. PAU, Brevet français 72-22.958 du 26.06.1972

[6] L.F. PAU, Méthodes statistiques de réduction et de compression des formes, Thèse docteur-ingénieur, Paris-Orsay, mai 1972.

[7] M. THAI SEREY, Vocodeur à canaux, E.N.S.T., décembre 1974.

ANNEXE : Procédures de diagonalisation

Nous allons passer en revue la plupart des méthodes de diagonalisation qui ont été étudiées sur le plan théorique et qui sont les plus utilisés sur le plan pratique.

Il existe trois grandes orientations qui dominent les méthodes de recherche des valeurs propres et des vecteurs propres des matrices réelles symétriques.

- Les méthodes directes conçues par une diagonalisation totale de la matrice
- Les méthodes itératives construites par la recherche des plus grandes valeurs propres (ou des plus petites en inversant la matrice).

../..

- Les méthodes d'approximation stochastique qui sont relatives aux sources aléatoires de potentiel infini.

I.) METHODES DIRECTES

I.1. Jacobi (1846) : on fait subir à la matrice une suite de rotation planes.

La convergence est lente relativement aux autres méthodes qui seront citées plus loin.

Pour le calcul des valeurs propres, on utilise la procédure "EIGEN" qui est une généralisation de la méthode de Jacobi.

I.2. GIVENS (1954)

On fait subir à la matrice une suite de (n-2) rotations planes qui la réduit à une forme tridiagonale. Cette méthode est plus rapide, et plus précise que Jacobi car elle ne repose pas sur un principe de convergence à la limite, mais sur la recherche de solutions algébriques.

I.3. HOUSEHOLDER (1958)

En (n-2) étapes, on réduit la matrice initiale à une forme tridiagonale, mais la méthode est deux fois plus rapide et moins encombrante en location mémoire que GIVENS.

a) On applique ensuite l'algorithme QR de Francis avec accélération de la convergence par déplacements de l'origine de KOBLANOUKAYA.

b) On peut appliquer la méthode de GIVHO (Méthode de bissection ou séparation des racines par les chaînes de STURM).

I.4. SVD (1970)

Singular Value Decomposition and Least Squares Solutions.

En analyse factorielle des correspondances, on fera directement les calculs d'après le tableau de contingence, sans calculer de matrice de contingence.

La méthode utilisée est une réduction à la forme bidiagonale du tableau rectangulaire (I, J), telle que : partie(J * J) bidiagonale, et : partie ((I - J), * J) = 0.

II.) METHODES ITERATIVES

II.1. Hotelling (1933)

On applique d'abord la récurrence $X = AX_n$, puis on calcule la matrice $A' = A - \lambda X {}^t X$ avec $\lambda = \lim_{m \to +\infty} \|X_m\|$

$X = \lim_{m \to +\infty} |X_m|$

et on recommance le processus.

Inconvénients - difficultés du choix du vecteur initial
- erreurs répercutées.

../..

II.2. Polynômes orthogonaux de Tchebytchev

(STIEFEL 1958) (ENGELL 1959)

III.) APPROXIMATION STOCHASTIQUE (1969 - BENZECRI)

BIBLIOGRAPHIE DE L'ANNEXE

1 - BENZECRI J.P. Algorithmes de géométrie euclidienne en analyse des données--L.S.M.--

2 - BENZECRI J.P. Approximation stochastique dans une algèbre normée non commutative--Bulletin de la société mathématique de France.1969, pp 225 - 244.

3 - FENELON J.P. Deux contributions à une programmathèque d'analyse des données : analyse factorielle des préférences, approximation stochastique de l'analyse des correspondances. Thèse de 3ème cycle, Université de Paris VI, le 26 Février 1973.

4 - FRANCIS J. The QR transformation. A unitary analogue to the LR transformation. Comput. J. 4.1961-62--

5 - GOLUB G. H. & REINSCH C. Singular Value Decomposition and Least Squares Solutions--Handbook Series Linear Algebra. Num. Math. 14 , 1970--

6 - GANTMACHER Théorie des matrices--Vol I & II. Dunod--

7 - STEWART G.W. Incoporating origin shifts into QR algorithm for symmetric tridiagonal matrices
Eigenvalues and eigenvectors of a real symmetric matrix
-- Communications of the A.C.M./Vol 13/n°6/June 1970--

8 - WILKINSON J.H. The algebric eigenvalue problem--Clarendon Press. Oxford. 1965--

9 - WILKINSON J.H. Global convergence of QR-algorithm--Processing of IFIP Congress. 1968--

10- WILKINSON J.H. Global convergence of tridiagonal QR algorith with origin shifts--Lin. Alg. and Appl. 1. 1968--

11- WILKINSON J.H. & REINSCH C. Handbook for Automatic Computation -- Vol II 1971 --

12- LEBART L. & FENELON J.P. Statistique et informatique appliquées -- 2è édition. Dunod 1973 --

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

UN ALGORITHME DE PROSODIE AUTOMATIQUE SANS ANALYSE SYNTAXIQUE

C. CHOPPY, J.S. LIENARD, D. TEIL

L.I.M.S.I. (C.N.R.S.) - B.P. 30 - 91406 ORSAY

RESUME

L'unité à réponse vocale ICOPHONE V fournit une voix parfaitement intelligible mais comportant un aspect artificiel et monotone. Un algorithme de prosodie automatique a été élaboré pour remédier en partie à ces défauts. Le texte est découpé en groupes intonatifs, à partir de considérations portant sur la seule longueur des mots, à l'exclusion de toute analyse syntaxique. Des schémas intonatifs simples, appliqués au niveau de ces groupes et au niveau de la phrase entière, sont combinés pour fournir le contour intonatif définitif. Le traitement est rapide et nécessite un faible volume de mémoire. Les résultats auditifs donnent à penser que la syntaxe a été prise en compte.

SUMMARY

The voice produced by the voice response unit ICOPHONE V is perfectly intelligible but posseses artificial and monotous features. An automatic prosody algorithm has been worked out so as to remedy these shortcomings in some degree. The text is cut up into intonative groups according to the word length to the exclusion of any syntax analysis. Simple intonation patterns applied to these groups and to the whole sentence are combined so as to give the definitive intonation outline. This processing is not time-consuming and requires but little computer memory capacity. The resulting utterance, as perceived, makes one think that syntax has been taken into account.

UN ALGORITHME DE PROSODIE AUTOMATIQUE SANS ANALYSE SYNTAXIQUE

C. CHOPPY, J.S. LIENARD, D. TEIL
L.I.M.S.I. (C.N.R.S.) - B.P. 30 - 91406 ORSAY

I - INTRODUCTION

La parole synthétique que nous obtenons par assemblage de diphonèmes (bib 1, 2) est parfaitement intelligible, mais présente un caractère artificiel très marqué, dû en partie au synthétiseur utilisé jusqu'à présent (Icophone, du type "relecteur de sonagramme"), et en partie à l'absence de facteurs prosodiques, notamment intonation et durée. L'unité à réponse vocale ICOPHONE V (bib 3) fournit donc une voix fonctionnelle, parfaitement utilisable grâce à la non-limitation du vocabulaire et au très faible temps de réponse. Nous cherchons, tout en conservant ces qualités, à introduire une prosodie déduite du texte, dans un double but :

- améliorer le découpage perçu du texte en éléments linguistiques, sans perturber l'intelligibilité au niveau phonétique
- rompre la monotonie engendrée par la succession de diphonèmes de durée constante.

De plus, devant être ultérieurement intégré à une unité de réponse vocale, l'algorithme devra être rapide et peu encombrant en mémoire.

Le synthétiseur paramétrique qui nous permettra de mettre en valeur la prosodie n'est pas encore en service. Nos expérimentations ont donc été menées sur l'ICOPHONE IV, pourvu d'une commande programmée de la fréquence fondamentale des oscillateurs et de la durée de chaque diphonème.

Les éléments prosodiques assument, entre autres, une fonction syntaxique pour le message parlé, il semblerait logique d'aborder ce problème par la mise en oeuvre d'une analyse syntaxique. Celle-ci, passablement complexe, a été réalisée pour le français par Jacqueline VAISSIERE (bib 4). Elle nécessite tout d'abord l'analyse du texte au niveau morphologique, c'est-à-dire la décomposition de chaque mot en ses composants (préfixes, radicaux, désinences, etc...), et une analyse grammaticale donnant les diverses fonctions possibles de chaque mot. On peut ensuite effectuer l'analyse syntaxique, qui permet au moyen d'arborescences de choisir à l'échelle de la phrase entre les diverses structures de chaque mot. Ce n'est qu'après avoir effectué tout ce travail que l'on peut associer un contour intonatif-type à chaque élément syntaxique, et introduire un découpage temporel au moyen de pauses.

La prosodie ainsi définie comprend également l'aspect micromélodique lié à la structure articulatoire des consonnes.

Ce processus nous semble difficilement applicable à notre problème de synthèse en temps réel, à cause du volume de calcul nécessaire. De plus il semble que l'analyse syntaxique puisse être facilement mise en défaut si l'on introduit en machine des phrases syntactiquement incorrectes, ambiguës ou incomplètes, ou simplement si le texte est entaché de fautes d'orthographe. Nous avons donc adopté une approche beaucoup plus simple, visant à fournir à l'auditeur une information syntaxique rudimentaire, mais cependant suffisamment variée d'une phrase à l'autre.

II - PRINCIPE DE L'ALGORITHME

La prosodie nous semble devoir convoyer deux grands types d'informations :

a) Les informations relatives au découpage de la phrase en mots ou groupes de mots, que nous appelerons "groupes intonatifs", ou élémentaires. Ces informations permettent également de classer les groupes élémentaires selon une hiérarchie de nature syntaxique (agencement des syntagmes) et surtout sémantique (insistance sur tel ou tel groupe élémentaire, visant à communiquer une nuance particulière). Mais un groupe élémentaire ne peut pas, sauf exception, transporter une idée, une signification, à lui tout seul.

b) Les informations relatives au découpage du texte en entités plus larges (phrases, propositions, membres de phrases) dont la caractéristique est leur autonomie sémantique. La prosodie devra à ce niveau comporter certaines informations sur la succession de ces unités (essentiellement : une suite est attendue, ou non).

II - 1. Découpage en groupes intonatifs

Dans cette première approche nous nous limiterons aux phrases énonciatives isolées. Il faut donc rechercher le découpage en groupes élémentaires. Soit par exemple la phrase

"Elle a acheté une robe très courte"

On peut la prononcer, sans en changer le sens, de différentes façons :

"Elle a acheté / une robe très courte"
"Elle a acheté / une robe / très courte"
"Elle a / acheté une robe / très courte" etc ...

Par contre le découpage

"Elle a acheté une / robe très courte"

est impossible, sauf s'il s'agit d'introduire une nuance particulière (amorce d'une question, puis réponse).

Nous baserons notre découpage sur la seule longueur des mots, estimée en nombre de syllabes. Les mots sont classés en deux catégories: mots courts (une syllabe), mots longs (plus d'une syllabe). Les associations sont faites entre deux mots courts successifs, ou entre un mot court et le mot long suivant.

La phrase précédente sera ainsi scindée en 4 groupes élémentaires :

"Elle a / acheté / une robe / très courte"

De même on aurait :

"Il faut / savoir / partir / à temps",

"Si vous / voulez / venir / il faut / téléphoner", etc ...

II - 2. Schémas intonatifs élémentaires

Il semble, après diverses études (bib 5, 6) qu'un schéma intonatif à 3 niveaux soit suffisant pour donner les nuances prosodiques que nous recherchons. Nous devrons cependant distinguer deux cas (fig. 1):

a) A chaque groupe élémentaire en position non finale nous appliquons le schéma (a), dont le maximum coïncide avec la dernière voyelle du groupe ou, si c'est un /ə / muet, sur l'avant dernière voyelle.

b) Le schéma (b) est appliqué au groupe final, mais le maximum est appliqué à la première voyelle du groupe.

II - 3. Schéma intonatif de phrase et schéma définitif

Pour marquer le découpage prosodique au niveau de la phrase nous appliquons un schéma d'ensemble (c) dont le maximum coïncide avec le maximum du 1er groupe.

Les schémas élémentaires et le schéma d'ensemble sont ensuite combinés pour former le schéma définitif (d), qui représente la moyenne entre la suite des schémas élémentaires et le schéma d'ensemble.

II - 4. Répartition des durées

En français, en dehors des phrases comportant des accents d'insistance, la durée des sons ne semble pas être une grandeur critique sur le plan phonétique, mises à part la durée des voyelles nasales, légèrement supérieure à celle des voyelles orales, la durée des occlusions, légèrement inférieure dans les occlusives voisées, et la durée de la dernière voyelle de la phrase, qui est souvent deux à trois fois plus longue que les autres. Ces règles très simples sont déjà prises en compte dans notre système de synthèse, de manière systématique, soit au niveau de la transcription orthographique-phonétique, soit dans la représentation même des diphonèmes.

Pour aller plus loin nous avons essayé de corréler la durée de chaque son avec sa fréquence de mélodie. Ceci entraînait un allongement choquant de la première partie de la phrase (accent traînant). Nous avons donc supprimé cette corrélation pour le premier groupe élémentaire, en la conservant pour le corps de la phrase.

III - RESULTATS ET CONCLUSION

Les résultats actuels ne sont que partiels, pour les raisons évoquées au § I. La figure 2 montre les contours intonatifs calculés pour les phrases

"Si vous voulez venir, il faut téléphoner"
et "Il faut savoir partir à temps".

La figure 3 montre les contours intonatifs de ces phrases relevés sur les spectrogrammes de parole synthétisée au moyen de l'ICOPHONE IV. On constate ainsi, comme à l'audition, que les petits accidents de la courbe sont estompés, et que les deux niveaux prosodiques (groupes intonatifs et ensemble de la phrase) sont perceptibles ainsi que nous l'avions souhaité.

On constate également
- qu'il est impossible, à l'écoute de ces phrases, d'affirmer qu'aucune analyse syntaxique n'a été faite,
- que les courbes obtenues ne présentent, d'une phrase à l'autre, aucun caractère systématique. Ce dernier point provient du fait qu'il est extrêmement rare de rencontrer deux phrases identiques quant au nombre des mots et au nombre de syllabes de chaque mot.

Il subsiste un certain nombre de points à étudier :
- la constitution des groupes élémentaires : d'une part certains mots (conjonctions, articles, pronoms) ne peuvent être regroupés que vers l'avant ; ces mots peuvent être facilement répertoriés. D'autre part nos investigations n'ont porté jusqu'à présent que sur des groupes élémentaires minima, composés au plus de deux mots ; il faut maintenant étudier les modalités de regroupement de plus de deux mots, ainsi que d'éventuelles règles de regroupement de deux groupes élémentaires successifs
- les schémas intonatifs élementaires qui doivent pouvoir se raccorder sans discontinuités (cas de dégénérescence à définir)
- les nuances prosodiques usuelles autres que l'énonciation, qui mettent en jeu, notamment, la relation entre le contour intonatif et la répartition des durées.

Nos premiers résultats nous permettent cependant de mettre en doute la nécessité d'une analyse syntaxique dans l'élaboration de la prosodie du français, du moins si l'on se place dans l'optique fonctionnelle que nous avons définie au § I.

- IV - BIBLIOGRAPHIE

1 - E. LEIPP, M. CASTELLENGO, J.S. LIENARD - La synthèse de la parole à partir de digrammes phonétiques - Comptes-Rendus du 6e Congrès International d'Acoustique - Tokyo, août 1968

2 - J.S. LIENARD, D. TEIL - Les éléments phonétiques et la traduction automatique du message écrit en message parlé - Revue "Automatisme", n° 10 - octobre 1970

3 - D. TEIL - Conception et réalisation d'un terminal à réponse vocale - Thèse de Docteur-Ingénieur - Université de Paris VI - 1975

4 - J. VAISSIERE - Contribution à la synthèse par règles du français - Thèse de troisième cycle - Université de Grenoble - novembre 1971

5 - P. DELATTRE - Studies in French and Comparative Phonetics - Mouton and C° - London - The Hague - Paris - 1966

6 - L.J. BOE - Synthèse paramétrique de la phrase énonciative en français - 5e Journées d'Etude sur la Parole, tenues au L.I.M.S.I. (ORSAY) du 15 au 17 mai 1974

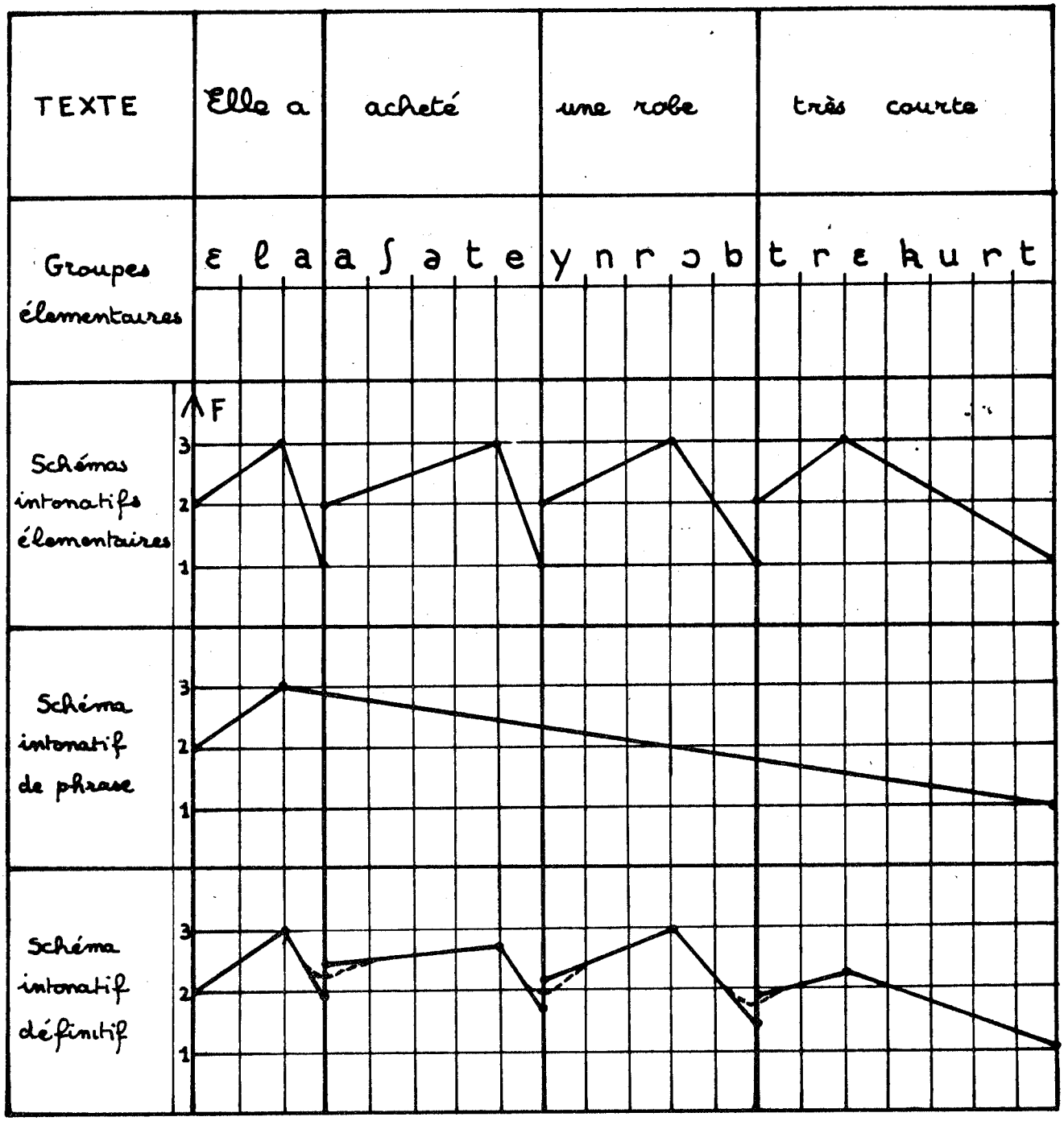


Fig 1 - Elaboration du schéma intonatif de la phrase "Elle a acheté une robe très courte"

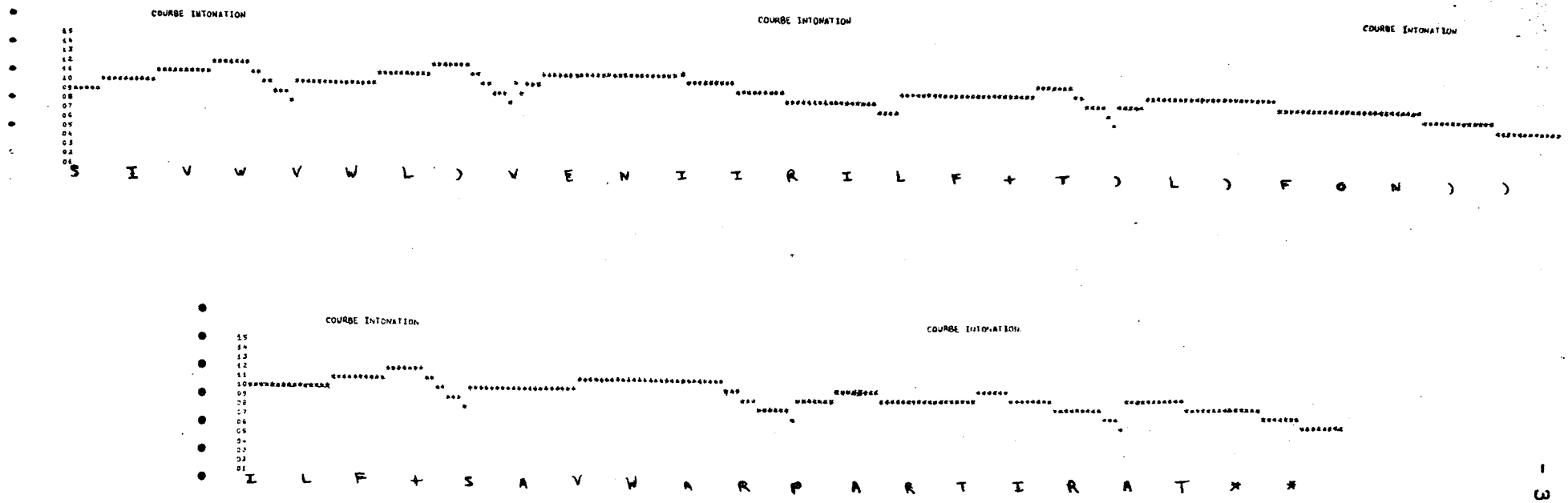


Fig 2 - Profils intonatifs définitifs obtenus par calcul pour les phrases " Si vous voulez venir , il faut téléphoner " et " Il faut savoir partir à temps " .

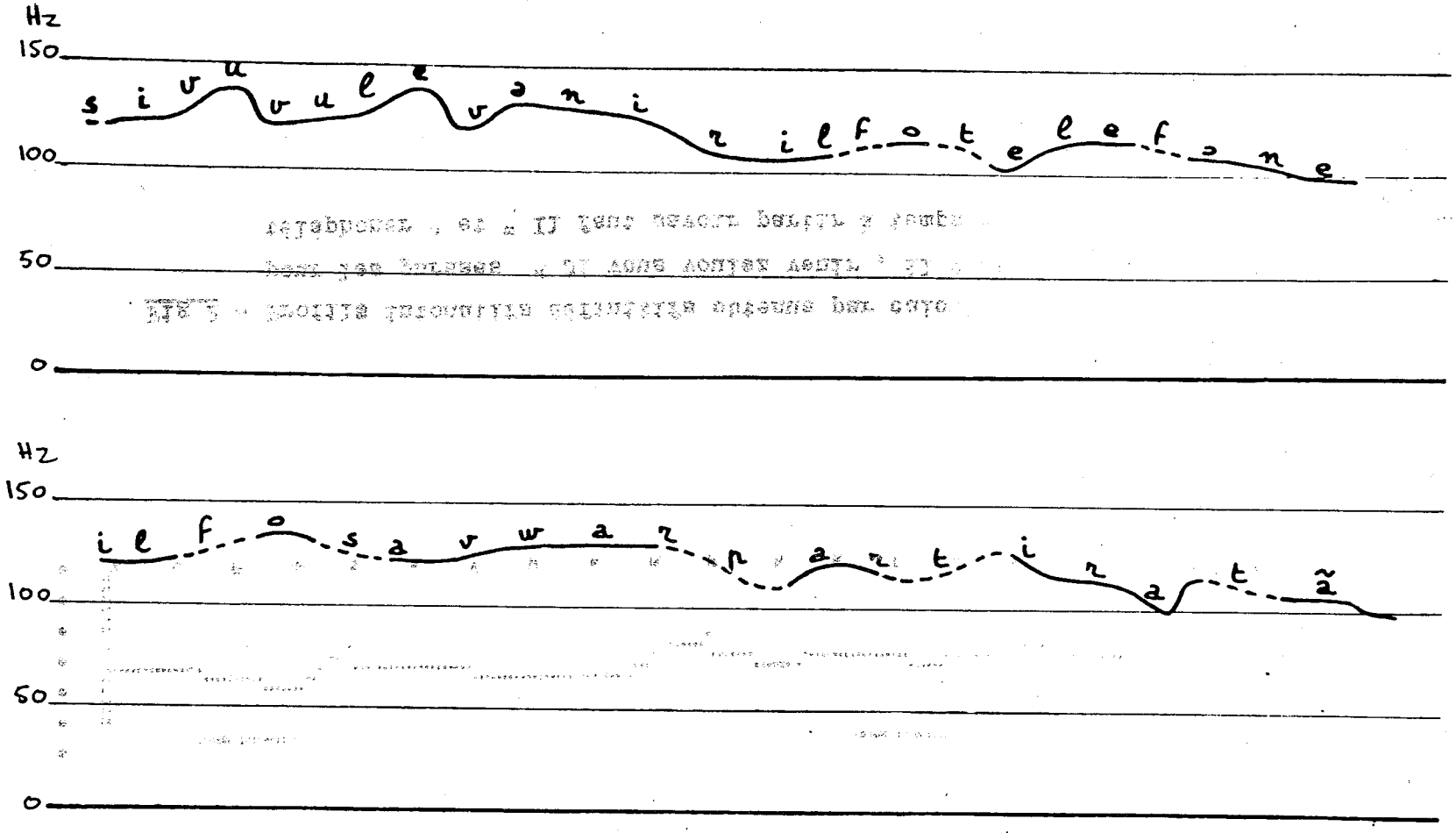


Fig 3 - Profils intonatifs relevés sur les spectrogrammes des phrases "Si vous voulez venir, il faut téléphoner" et "Il faut savoir partir à temps" synthétisées sur ICOPHONE IV avec prosodie automatique.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

Codes phonétiques (phonocodes) et commandes verbales

par J.A. Dreyfus-Graf, Genève, et équipes du C.N.E.T. Lannion

RESUME

Un appareil de reconnaissance automatique de mots phonocodés, le Phono-Décodeur I, nommé CHARLES, est actuellement en cours de développement au C.N.E.T., Lannion, Département de M. LORAND, équipe de M. CARTIER. Il permet d'évaluer la possibilité de construire des machines qui obéissent, en temps réel, à des voix très diverses. Les résultats actuels établissent qu'il n'y a pas de différence systématique entre voix masculines et féminines, que les transitoires ne sont pas significatives pour les phonèmes non-plosifs, et que la segmentation est entièrement réalisable sur la base du dispositif en cours d'expérimentation.
D'autres résultats plus complets pourront être communiqués prochainement.

SUMMARY

Phonetic Codes (Phonocodes) and spoken commands

An apparatus called Phono-Decoder I, or CHARLES, for the recognition of coded words, is presently in course of development at the C.N.E.T., Lannion (Department of Mr LORAND, team of Mr CARTIER). It opens the way for building machines controlled by various voices in real time. According to the present results, there is no systematic difference between male and female voices, the transitory parts of phonemes are not significant, except for stop consonnants, and the segmentation is entirely achievable on the basis of the device now under experiment.

Further and more complete results will be communicated in the near future.

Codes phonétiques (phonocodes) et commandes verbales

par J.A. Dreyfus-Graf, Genève, et équipes du CNET, Lannion

1. COMMANDES VERBALES EN TEMPS REEL

Est-il possible de construire des machines relativement simples qui obéissent immédiatement à la parole humaine, c'est-à-dire sans exiger la mémorisation préalable d'échantillons vocaux prononcés individuellement par chaque usager ?

Pour répondre à cette question, un appareil de reconnaissance automatique, nommé "Phono-Décodeur I", ou CHARLES, est en cours de développement au Centre National d'Etudes des Télécommunications (CNET), à Lannion.

Ces travaux sont effectués dans le Département de M. LORAND, par l'équipe de M. CARTIER, avec MM. COURBON, DAGORNE, DOUCEN, LUCAS, et autres collaborateurs.

CHARLES est un sigle signifiant "Capteur Hybride Adapté à la Reconnaissance d'un Langage d'Eléments Spéciaux".

Il est conçu en vue d'obéir immédiatement à des mots en nombres illimités, prononcés par des usagers masculins ou féminins appartenant à n'importe quel groupe linguistique. Ces mots universels sont formés à partir de codes phonétiques, nommés PHONOCODES [1][2][3][4], tels que SOTINA ou SOKINA. Ces phonocodes sont basés sur les classes de phonèmes les plus internationales et les plus distinctes, telles que les 3 groupes de voyelles, O (=o,ou), I (=i,é), A(=a,â), et les 3 groupes de consonnes S (=s) ou Ŝ (=ch), T ou K, N ou M.

2. SCHEMA DE PRINCIPE DU PHONO-DECODEUR I = CHARLES

Les positions P1 à P85 du schéma de la Fig. 1. situent les principales fonctions du Phono-Décodeur I = CHARLES

Les mots phonocodés P1 peuvent être des monophones O,A,I,S, K,N, des diphones 00,01,S0,0S, des triphones KIO, ANA, OSA, INO, ou des polyphones quelques tels que SOSIN, NAKOS, KANIS, etc.

Les 4 entrées, P2 à P5, peuvent émaner soit d'un microphone, soit d'un magnétophone, soit d'un téléphone, ou d'un vocoder.

Un préamplificateur P6 régularise les spectres de fréquence de ces 4 entrées dans la bande téléphonique, de 100 à 3400 Hz environ. La dynamique d'entrée admissible est de 40 dB. Les dépassements sont signalés par les lumières de 2 lampes, "max" et "min", P7.

Les 2 compresseurs d'amplitude I et II (P8, P9) résuisent à 2 dB la dynamique admissible de 40 dB. Le compresseur I sélectionne les plosives, grâce à des constantes de temps de montée T_m=10 ms, et de descente T_d= 30 ms. Le temps T_v, plus ou moins grand que 10 ms, qui s'écoule entre l'apparition de la plosive (détecteur P10) et celle d'un voisement (détecteur P11) permet de séparer les consonnes plosives des voyelles explosées. Le compresseur II égalise les niveaux des phonèmes quasi-stationnaires, grâce à des constantes de temps de montée T_m=1 ms, et de descente T_d= 20 ms.

Chaque compresseur peut comprendre 2 boucles de réglage, "avant" et "arrière". Il délivre un signal doublement logarithmique, le rapport R entre les variations de niveaux d'entrée et de sortie pouvant être ajusté entre 1/5 et 1/20, [5].

Six filtres de bande, P21 à P26, suivis de 6 redresseurs, P31 à P36, et de 6 passe-bas 0-20 Hz, P 41 à P 46, analysent les composantes de fréquences (formants) des mots prononcés, ou d'autres sons captés. Les 6 canaux C1 à C6 d'analyne spectrale sont complétés par un 7ème canal, C7 caractérisant l'énergie globale, avec ou sans compression. On peut encore prévoir 7 couvertisseurs LIN/LOG, P51 à P57, fournissant les 7 sorties de canaux C1' à C7'-

Sept combinateurs de canaux, P61 à P67, permettent d'effectuer des sommes et des différences d'énergie, émanant de 8 canaux : les 6 canaux d'analyse spectrale C1 à C6 (ou C1' à C6'), le canal global C7 (ou C7') et le canal de plosives C8. On obtient ainsi 7 paramètres spectraux R1 à R7, qui, associés à 7 seuils, P71 à P77, délivrent 7 paramètres logiques L1 à L7.

Le traitement digital, P81 à P85, comprend les parties suivantes :

Le convertisseur analog-digital P81 échantillonne et quantifie les sorties des canaux d'analyse décrits précédemment. La période d'échantillonnage est T=2,5 ms, ou 5 ou 10 ms. La quantification digitalise jusqu'à 256 degrés d'amplitude, c'est-à-dire 8 bits.

Le multiplexeur P82 transmet les signaux digitaux à l'unité de traitement logique, qui est actuellement un calculateur T 1600 ou CII, 10070, mais qui sera finalement un micro-processeur, P83, intégré dans CHARLES.

Le traitement logique comprend 2 parties :
La logique I (des phonèmes), P84, qui effectue la reconnaissance des 6 classes de phonèmes (O,I,A,S,K,N), et leurs segmentations à l'intérieur des mots. Elle se base sur une matrice CR de canaux, une matrice RL de seuils et un système d'équations logiques EL.
La logique II (des mots), P85, opère la reconnaissance des combinaisons de phonèmes formant des mots, ainsi que leurs segmentations, définies par des durées de silences.

3. MATRICES ET EQUATIONS LOGIQUES

Les matrices de canaux CR4' et de seuils RL4, expérimentées actuellement, avec les équations logiques Rl4, figurant dans les Tableaux I à III.

La matrice de canaux CR4' comprend 6 combinaisons des 7 canaux Cl' à C6' et C8. Elle délivre 6 paramètres spectraux Rl à R6, permettant de distinguer les classes de phonèmes K et S, ainsi que les groupes I+N+O et A+O, puis de séparer N de I, et O de I.

La matrice des seuils RL4 détermine les 6 valeurs logiques (oui-non) à partir desquelles les équations logiques EL4 opèrent la reconnaissance des 6 classes de phonèmes désirées.

Dans la présente expérimentation, les canaux spectraux Cl' à C6' délivrent des signaux triplement logarithmiques, ceux-ci étant régulés d'abord par le compresseur d'amplitude II, P9 à double-boucle, puis traités par les convertisseurs LIN/LOG, P51 à P57.

Par contre le canal C7 de l'énergie globale délivre un signal linéaire, dont le seuil, fixé à 50 degrés d'amplitude, détermine les débuts et fins de mots.

4. LE MOT "KIO" PRONONCE PAR UN HOMME ET PAR UNE FEMME

Les Fig. 2 et 3 reproduisent les listings de reconnaissance du mot KIO, prononcé par un homme et par une femme, en utilisant la matrice spectrale CR4' (Tableau I), la matrice de seuils RL4 (Tableau II) et les équations logiques EL4 (Tableau III).

La Fig. 2 montre les numéros impairs (interlignes 5 ms) des 288 échantillons (période 2,5 ms) qui représentent le mot KIO prononcé par un homme.

Les durées sont les suivantes : K=47,5 ms, I=342,5 ms, 0=307,5 ms. Parmi les 19 échantillons du K, les 137 échantillons du I et les 123 échantillons du 0, il ne s'est pas produit un seul échantillon éronné.

La transition entre le I et le 0 présente 9 échantillons N (Nos 157 à 165), dont la durée de 22,5 ms est négligeable comparativement aux durées des parties quasi-stationnaires.

De même, la Fig. 3 montre les numéros impairs des 196 échantillons qui représentent le mot KIO, prononcé par une femme.

Les durées sont les suivantes : K=42,5 ms, I=305 ms, 0=132,5 ms, sans aucun échantillon erroné. La transition entre le I et le 0 est marquée par 4 échantillons N (Nos 140 à 143) = 10 ms, qui sont encore plus négligeable que dans le cas précédent,de la voix masculine.

La Fig. 4 montre le détail d'une partie du KIO émanant de la voix féminine de la Fig. 3.

Il s'agit de la transition entre le I et le O, analysée par les 6 canaux spectraux C1' à C6' et le canal global C7, puis par les paramètres combinatoires R2 à R6. On constate que la distinction des classes de phonèmes s'opère indépendamment des variations d'énergie globale, représentant 26 dB, entre 82 et 129 degrés d'amplitude, dans le canal C7.

5. QUATRE MOTS PRONONCES PAR 5 HOMMES ET 3 FEMMES

Des essais ont été effectués aussi sur les 4 mots KIO, ANA, OSA, INO prononcés par 5 hommes et 3 femmes, soit sur 32 mots. La segmentation des voyelles a été effectuée à 100 % grâce à un algorithme comparant des groupes successif de plusieurs échantillons.

Des méthodes d'optimisation, concernant notamment les matrices de spectres et de seuils, sont actuellement en cours d'expérimentation.

6. CONCLUSIONS ACTUELLES

Dans l'état actuel de son développement, le Phono-Décodeur I, nommé CHARLES, permet de tirer les conclusions suivantes concernant la reconnaissance automatique de mots phonocodés :

a) Il n'y a pas de différence systématique entre les voix d'hommes et de femmes.
b) Les durées des consonnes plosives sont inférieures à 50 ms, celles des autres phonèmes sont supérieures à 100 ms, ce qui permet déjà une distinction entre plosives et fricatives, par exemple.
c) Les phonèmes non-plosifs sont caractérisés par l'analyse spectrale de leurs parties quasi-stationnaires (durées plus grandes que 100 ms et leurs parties transitoires ne sont pas significatives (durées plus petites que 25 ms).
d) Les variations de dynamique sont à éliminer dans l'analyse spectrale, ceci avec des constantes de temps de montée supérieures à 8 ms pour les plosives, et inférieures à 3 ms pour les autres phonème.
e) L'utilité des convertisseurs LIN/LOG reste à examiner.
f) Quand les matrices de sepctres et de seuils seront optimisées, le convertisseur analog-digital se contentera d'un échantillonnage à 10 ms, et d'une quantification à 4 ou 2 niveaux (2 ou 1 bit.).
g) La segmentation est entièrement réalisable sur la base du dispositif en cours d'expérimentation.

Des résultats plus complets pourront être communiqués prochainement.

R E F E R E N C E S

(Publications J.A. Dreyfus-Graf)

[1] Recognition of Natural and of Artificial Speech (Phonocode) Reports of the ICSP, IEEE-AFCRL, Newton-Boston, 1972

[2] Reconnaissance automatique de la parole codée (phonocode), sonore et chuchotée, Revue d'Acoustique No 25, Paris 1973

[3] La reconnaissance subjective et objective de la parole codée (phonocode), 5èmes Journées d'Etude sur la Parole, GALF, AFCET, LIMSI, Orsay 1974

[4] Coded Speech (Phonocode) and Recognition Machines, 8 th ICA, London 1974

[5] Reconnaissance de signaux et régulation d'information Revue d'Acoustique No 18, Paris 1972.

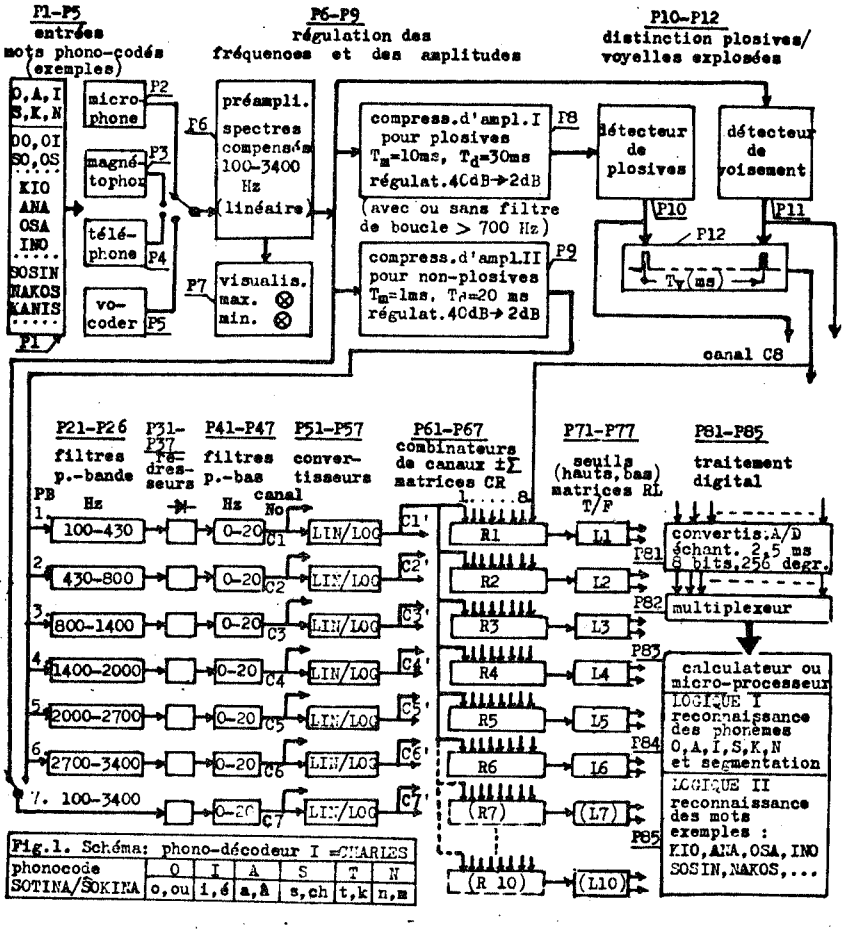


Fig.1. Schéma: phono-décodeur I =CHARLES

phonocode	O	I	A	S	T	N
SOTINA/SOKINA	o,ou	i,é	a,à	s,ch	t,k	n,m

Tableau I. Matrice des combinaisons de canaux

CR4 / CR4'	C1 / C1'	C2 / C2'	C3 / C3'	C4 / C4'	C5 / C5'	C6 / C6'	C7 / C7'	C8	distinctions
R1	0	0	0	0	0	0	0	1	K /reste
R2	-1/3	-1/3	-1/3	1/3	1/3	1/3	0	0	S /reste
R3	1	0	-1	0	0	0	0	0	I+N+0/reste
R4	0	1/2	1/2	-1/2	-1/2	0	0	0	A+0 /reste
R5	0	0	1	0	0	-1	0	0	N/I
R6	0	1/2	1/2	0	-1	0	0	0	O/I

Tableau II. Matrice des seuils

RL4	R1L1	R2L2	R3L3	R4L4	R5L5	R6L6'	
	50	35	40	20	-60	0	(degrés)

Tableau III.

équations logiques EL4
K = L1
S = L2.$\overline{L3}$.$\overline{L4}$
A = $\overline{L3}$.L4.L5.L6
O = L3.L4.L5.L6
I = L3.$\overline{L4}$.$\overline{L5}$.$\overline{L6}$
N = L3.L5

seuil début de mot : E= 50(degrés)

	équations logiques EL4 : T = true = oui F = false = non																
échant. No	R1-L1	R2-L2	R3-L3	R4-L4	R5-L5	R6-L6	R7-L7	KIO class. recon.	échant. No	R1-L1	R2-L2	R3-L3	R4-L4	R5-L5	R6-L6	R7-L7	KIO class. recon.
1	F	F	T	F	F	F	T	K	113	F	F	T	F	F	F	F	I
3	F	F	T	F	F	F	F	K	115	F	F	T	F	F	F	F	I
5	F	F	T	F	F	F	F	K	117	F	F	T	F	F	F	F	I
7	F	T	T	F	F	F	F	K	119	F	F	T	F	F	F	F	I
9	F	T	F	F	F	F	F	K	121	F	F	T	F	F	F	F	I
11	F	T	F	F	F	F	F	K	123	F	F	T	F	F	F	F	I
13	F	T	F	F	F	F	F	K	125	F	F	T	F	F	F	F	I
15	F	T	F	F	F	F	F	K	127	F	F	T	F	F	F	F	I
17	F	T	F	F	F	F	F	K	129	F	F	T	F	F	F	F	I
19	F	F	T	F	F	F	F	K	131	F	F	T	F	F	F	F	I
21	F	F	T	F	F	F	F	I	133	F	F	T	F	F	F	F	I
23	F	F	T	F	F	F	F	I	135	F	F	T	F	F	F	F	I
25	F	F	T	F	F	F	F	I	137	F	F	T	F	F	F	F	I
27	F	F	T	F	F	F	F	I	139	F	F	T	F	F	F	F	I
29	F	F	T	F	F	F	F	I	141	F	F	T	F	F	F	F	I
31	F	F	T	F	F	F	F	I	143	F	F	T	F	F	F	F	I
33	F	F	T	F	F	F	F	I	145	F	F	T	F	F	F	F	I
35	F	F	T	F	F	F	F	I	147	F	F	T	F	F	F	F	I
37	F	F	T	F	F	F	F	I	149	F	F	T	F	F	F	F	I
39	F	F	T	F	F	F	F	I	151	F	F	T	F	F	F	F	I
41	F	F	T	F	F	F	F	I	153	F	F	T	F	F	F	T	I
43	F	F	T	F	F	F	F	I	155	F	F	T	F	F	F	T	I
45	F	F	T	F	F	F	F	I	157	F	F	T	F	T	F	T	N
47	F	F	T	F	F	F	F	I	159	F	F	T	F	T	F	T	N
49	F	F	T	F	F	F	F	I	161	F	F	T	F	T	F	T	N
51	F	F	T	F	F	F	F	I	163	F	F	T	F	T	F	T	N
53	F	F	T	F	F	F	F	I	165	F	F	T	F	T	F	T	N
55	F	F	T	F	F	F	F	I	167	F	F	T	T	T	T	T	0
57	F	F	T	F	F	F	F	I	169	F	F	T	T	T	T	T	0
59	F	F	T	F	F	F	F	I	171	F	F	T	T	T	T	T	0
61	F	F	T	F	F	F	F	I	173	F	F	T	T	T	T	T	0
63	F	F	T	F	F	F	F	I	175	F	F	T	T	T	T	T	0
65	F	F	T	F	F	F	F	I	177	F	F	T	T	T	T	T	0
67	F	F	T	F	F	F	F	I	179	F	F	T	T	T	T	T	0
69	F	F	T	F	F	F	F	I	181	F	F	T	T	T	T	T	0
71	F	F	T	F	F	F	F	I	183	F	F	T	T	T	T	T	0
73	F	F	T	F	F	F	F	I	185	F	F	T	T	T	T	T	0
75	F	F	T	F	F	F	F	I	187	F	F	T	T	T	T	T	0
77	F	F	T	F	F	F	F	I	189	F	F	T	T	T	T	T	0
79	F	F	T	F	F	F	F	I	191	F	F	T	T	T	T	T	0
81	F	F	T	F	F	F	F	I	193	F	F	T	T	T	T	T	0
83	F	F	T	F	F	F	F	I	195	F	F	T	T	T	T	T	0
85	F	F	T	F	F	F	F	I	...	...	...	...	...	...	...	...	...
87	F	F	T	F	F	F	F	I	...	...	...	...	...	...	...	...	...
89	F	F	T	F	F	F	F	I	...	...	...	...	...	...	...	...	...
91	F	F	T	F	F	F	F	I	267	F	F	T	T	T	T	T	0
93	F	F	T	F	F	F	F	I	269	F	F	T	T	T	T	T	0
95	F	F	T	F	F	F	F	I	271	F	F	T	T	T	T	T	0
97	F	F	T	F	F	F	F	I	273	F	F	T	T	T	T	T	0
99	F	F	T	F	F	F	F	I	275	F	F	T	T	T	T	T	0
101	F	F	T	F	F	F	F	I	277	F	F	T	T	T	T	T	0
103	F	F	T	F	F	F	F	I	279	F	F	T	T	T	T	T	0
105	F	F	T	F	F	F	F	I	281	F	F	T	T	T	T	T	0
107	F	F	T	F	F	F	F	I	283	F	F	T	T	T	T	T	0
109	F	F	T	F	F	F	F	I	285	F	F	T	T	T	T	T	0
111	F	F	T	F	F	F	F	I	287	F	F	T	T	T	T	T	0

Fig.2 Reconnaissance du mot KIO (voix masculine, M. R.D.), avec matrice RL4 (févr.1975) et équations logiques EL4. Echantillonnage 2,5 ms. Interligne 2x2,5= 5 ms.
Mot No 5, 288 échantillons x 2,5 ms = 720 ms. Calculateur CII, 10070.
Durées : K= 19 x 2,5= 47,5 ms, I= 137 x 2,5 = 342,5 ms, 0= 123x2,5=307,5ms
Transition I→0 : 9 échantill."N"(Nos 156 à 165) x 2,5= 22,5 ms (négligeables)

	équations logiques EL4 : T = true, F = false																	
T=5ms échant. No	R1L1	R2L2	R3L3	R4L4	R5L5	R6L6	R7L7	KIO classes recon.	T=5ms échant. No	R1L1	R2L2	R3L3	R4L4	R5L5	R6L6	R7L7	KIO classes recon.
1	F	T	F	F	F	F	T	K	95	F	F	T	F	F	F	F	I
3	F	T	F	F	F	F	T	K	97	F	F	T	F	F	F	F	I
5	F	T	F	F	F	F	T	K	99	F	F	T	F	F	F	F	I
7	F	T	F	F	F	F	T	K	101	F	F	T	F	F	F	F	I
9	F	T	F	F	F	F	T	K	103	F	F	T	F	F	F	F	I
11	F	T	F	F	F	F	T	K	105	F	F	T	F	F	F	F	I
13	F	T	F	F	F	F	T	K	107	F	F	T	F	F	F	F	I
15	F	T	F	F	F	F	F	K	109	F	F	T	F	F	F	F	I
17	F	F	T	F	F	F	F	K	111	F	F	T	F	F	F	F	I
19	F	F	T	F	F	F	F	I	113	F	F	T	F	F	F	F	I
21	F	F	T	F	F	F	F	I	115	F	F	T	F	F	F	F	I
23	F	F	T	F	F	F	F	I	117	F	T	T	F	F	F	F	I
25	F	F	T	F	F	F	F	I	119	F	T	T	F	F	F	F	I
27	F	F	T	F	F	F	F	I	121	F	T	T	F	F	F	F	I
29	F	F	T	F	F	F	F	I	123	F	T	T	F	F	F	F	I
31	F	F	T	F	F	F	F	I	125	F	F	T	F	F	F	F	I
33	F	F	T	F	F	F	F	I	127	F	F	T	F	F	F	F	I
35	F	F	T	F	F	F	F	I	129	F	F	T	F	F	F	F	I
37	F	F	T	F	F	F	F	I	131	F	F	T	F	F	F	F	I
39	F	F	T	F	F	F	F	I	133	F	F	T	F	F	F	T	I
41	F	F	T	F	F	F	F	I	135	F	F	T	F	F	F	T	I
43	F	F	T	F	F	F	F	I	137	F	F	T	F	F	F	T	I
45	F	F	T	F	F	F	F	I	139	F	F	T	F	F	F	T	I
47	F	F	T	F	F	F	F	I	141	F	F	T	F	T	F	T	(N)
49	F	F	T	F	F	F	F	I	143	F	F	T	F	T	T	T	(N)
51	F	F	T	F	F	F	F	I	145	F	F	T	T	T	T	T	9
53	F	F	T	F	F	F	F	I	147	F	F	T	T	T	T	T	9
55	F	F	T	F	F	F	F	I	149	F	F	T	T	T	T	T	8
57	F	F	T	F	F	F	F	I	151	F	F	T	T	T	T	T	9
59	F	F	T	F	F	F	F	I	153	F	F	T	T	T	T	T	9
61	F	F	T	F	F	F	F	I	155	F	F	T	T	T	T	T	9
63	F	F	T	F	F	F	F	I	157	F	F	T	T	T	T	T	8
65	F	F	T	F	F	F	F	I	159	F	F	T	T	T	T	T	9
67	F	F	T	F	F	F	F	I	161	F	F	T	T	T	T	T	8
69	F	F	T	F	F	F	F	I	163	F	F	T	T	T	T	T	9
71	F	F	T	F	F	F	F	I	165	F	F	T	T	T	T	T	9
73	F	F	T	F	F	F	F	I	167	F	F	T	T	T	T	T	9
75	F	F	T	F	F	F	F	I	169	F	F	T	T	T	T	T	9
77	F	F	T	F	F	F	F	I	171	F	F	T	T	T	T	T	8
79	F	F	T	F	F	F	F	I	173	F	F	T	T	T	T	T	8
81	F	F	T	F	F	F	F	I	175	F	F	T	T	T	T	T	9
83	F	F	T	F	F	F	F	I	177	F	F	T	T	T	T	T	9
85	F	F	T	F	F	F	F	I	179	F	F	T	T	T	T	T	9
87	F	F	T	F	F	F	F	I	181	F	F	T	T	T	T	T	9
89	F	F	T	F	F	F	F	I	183	F	F	T	T	T	T	T	9
91	F	F	T	F	F	F	F	I	185	F	F	T	T	T	T	T	8
93	F	F	T	F	F	F	F	I	187	F	F	T	T	T	T	T	9
									189	F	F	T	T	T	T	T	9
									191	F	F	T	T	T	T	T	8
	196 échantillons x 2,5 = 490 ms								193	F	F	T	T	T	T	T	5
	mot: 25 ,Mlle S.								195	F	F	T	T	T	T	T	9

Fig.3 Reconnaissance du mot KIO (voix féminine Mlle D.), avec matrice RL4 et équations logiques EL4. Interlignes 2x2,5 = 5 ms. Durées : K= 42,5 ms, I= 305 ms O=132,5 ms; transition I→O = 4 échantillons "N" (Nos 141 à 143) = 10 ms (négligeables). Listings fournis par calculateur CII, 10070, févr.1975.

	analyse spectrale (tripl.logarithm.)						énergie	combinaisons de canaux, matrice CR4'				
canal→	C1'	C2'	C3'	C4'	C5'	C6'	C7	R2	R3	R4	R5	R6
T=2,5ms échant.	100-430Hz	430-800	800-1400	1400-2000	2000-2700	2700-3400	100-3400	S/reste	I+N+O/reste	A+O/reste	N/I	O/I
NO												
115 ↓	245	232	46	156	241	236	82	34.	199.	-56.	-190.	-102.
116 I	245	233	50	153	243	238	85	35.	195.	-56.	-188.	-101.
117 I	245	235	54	156	245	240	88	35.	191.	-56.	-186.	-100.
118 I	245	236	58	160	247	241	91	36.	187.	-56.	-183.	-100.
119 I	244	238	62	165	249	241	94	37.	182.	-57.	-179.	-99.
120 I	244	239	66	169	249	241	97	36.	178.	-56.	-175.	-97.
121 I	244	241	69	175	249	240	100	36.	175.	-57.	-171.	-94.
122 I	244	242	72	181	249	238	103	36.	172.	-58.	-166.	-92.
123 I	244	243	74	186	249	235	105	36.	170.	-59.	-161.	-90.
124 I	244	245	76	189	249	232	107	35.	168.	-58.	-156.	-88.
125	245	246	77	191	249	228	108	33.	168.	-58.	-151.	-87.
126 I	246	248	79	194	248	223	111	30.	167.	-57.	-144.	-84.
127 I	247	249	80	197	246	217	113	28.	167.	-57.	-137.	-81.
128 I	248	249	81	200	244	211	115	25.	167.	-57.	-130.	-79.
129 I	249	249	81	203	240	204	116	22.	168.	-56.	-123.	-75.
130 I	249	249	82	203	236	196	117	18.	167.	-54.	-114.	-70.
131 I	249	249	82	203	231	188	119	14.	167.	-51.	-106.	-65.
132 I	249	249	83	203	226	181	121	10.	166.	-49.	-98.	-60.
133 I	250	249	83	200	219	174	122	4.	167.	-44.	-91.	-53.
134	250	249	83	196	212	168	124	-2.	167.	-38.	-85.	-45.
135 I	250	249	83	191	205	162	125	-8.	167.	-32.	-79.	-39.
136 I	249	249	82	184	197	157	126	-14.	167.	-25.	-75.	-31.
137 I	249	249	82	178	190	153	127	-19.	167.	-18.	-71.	-25.
138 I	250	249	81	172	183	148	129	-25.	169.	-12.	-67.	-18.
139 ↓	250	249	84	167	179	146	129	-30.	166.	-7.	-62.	-12.
140	250	249	90	164	177	145	129	-34.	160.	-1.	-55.	-7.
141 (N)	250	249	95	161	175	144	129	-38.	155.	4.	-49.	-3.
142 (N)	250	249	98	156	173	143	129	-41.	152.	9.	-45.	1.
143	250	249	104	152	170	140	129	-47.	146.	15.	-36.	7.
144 ↓	250	249	108	147	167	136	127	-52.	142.	21.	-28.	11.
145 O	250	249	114	142	162	132	126	-58.	136.	29.	-18.	19.
146 O	250	249	118	136	158	126	125	-65.	132.	36.	-7.	25.
147	250	249	122	130	153	118	123	-73.	128.	44.	4.	32.
148	249	249	124	123	148	110	121	-80.	125.	51.	14.	38.
149	249	248	127	117	145	103	119	-85.	122.	56.	24.	42.
150	249	247	130	110	142	95	117	-92.	119.	62.	35.	46.
151	249	245	135	106	140	89	115	-97.	114.	67.	46.	50.
152	248	243	138	101	137	83	113	-102.	110.	71.	55.	53.
153	247	242	142	98	135	78	111	-106.	105.	75.	64.	57.
154	247	241	147	95	132	74	109	-110.	100.	80.	73.	62.
155	246	240	150	94	131	72	107	-112.	96.	82.	78.	64.
156	245	240	155	93	130	70	105	-115.	90.	86.	85.	67.
157	245	239	159	92	129	70	104	-116.	86.	88.	89.	70.
158	244	239	165	92	129	71	102	-117.	79.	92.	94.	73.
159	243	239	169	92	129	72	101	-118.	74.	93.	97.	75.
160	243	239	176	91	127	74	99	-121.	67.	98.	102.	80.
161	242	239	178	91	129	77	97	-119.	64.	98.	101.	79.
162	242	239	179	89	129	78	96	-120.	63.	100.	101.	80.
163 O	241	239	179	88	131	80	95	-119.	62.	100.	99.	78.
164	240	239	179	86	132	81	94	-118.	61.	100.	98.	77.
165 ↓	239	238	177	85	134	84	93	-116.	62.	98.	93.	73.

Fig.4. Analyse spectrale et combinaisons de canaux (matrice CR4') d'une partie du mot KIO (même voix féminine que Fig.3, échant. Nos 115 à 165). Transition I-O = 4 échant. "N" = 10ms. Canaux C1' à C6' avec compression+ convert. lin/log. Canal 7 (énergie) sans compres., lin. Interligne= 2,5 ms= T. Listings fournis par calculateur CII, 10010. (Févr. 1975. Mlle S).

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

Optimisation de phonocodes par tests d'intelligibilité avec des sujets allemands

Jens-Peter Koester

Jean A. Dreyfus-Graf

RESUME

Les phonocodes joueront un rôle important pour la commande verbale de machines. Dans la contribution suivante, nous donnons la description de l'optimisation du code SOTINAKEMUŜ dans cinq conditions externes différentes (le signal naturel et quatre sortes de distortions) avec des sujets allemands. L'investigation a été restreinte à des mono-, di- et triphones.

Les résultats reflètent des lois générales de la perception de signaux, construits d'une manière formelle, en fonction des sons mêmes, du contenu sémantique des signaux, de la structure des syllabes et de la qualité du signal aussi bien que de l'âge et du sexe des personnes interrogées.

SUMMARY

Optimization of phonocodes by means of perception tests with German subjects

Phonocodes will play an important role in the development of machines controllable by spoken commands. The following contribution describes the subjective optimization of the SOTINAKEMUŜ-code under five diverse external conditions: the natural signal and four different kinds & degrees of distortion. The investigation was restricted to mono-, di-, and triphones and German subjects were used.

The results in addition provide data on general laws governing the perception of signals constructed according to formal principles - laws which take into account the sounds themselves, the semantic loading of the signal, its syllable structure, the quality of the signal as well as age and sex of the subjects.

Optimisation de phonocodes par tests d'intelligibilité avec des sujets allemands

Jens-Peter Koester
Jean A. Dreyfus-Graf

Dans différentes publications, J.A. Dreyfus-Graf a discuté la grande utilité de l'application de phonocodes à des problèmes techniques et humains (1). Cette utilité, cependant, n'est garantie qu'au moment où la qualité d'un phonocode a été minutieusement optimisée.

Le développement d'un phonocode comprend trois opérations principales, à savoir:

1. la sélection de l'inventaire de classes internationales de phonèmes
2. la combinaison logique de celles-ci selon des règles phonétiques et linguistiques
3. l'optimisation du système théorique.

La sélection de l'inventaire des classes internationales de phonèmes est fonction des distances vectorielles entre elles et du système que l'on se propose de coder. L'ensemble de l'inventaire, un nombre agencé de voyelles et de consonnes, caractérise le matériel de base d'un des phonocodes possibles.

La combinaison des éléments de base se fait selon des règles logiques qui fixent plus ou moins rigoureusement le nombre et l'ordre des éléments à l'intérieur d'un élément fonctionnel afin de faciliter la segmentation (et par cela la reconnaissance) du message articulé codé. Les règles permettent aussi de distinguer différentes sortes d'éléments combinés (élémentaire, non élémentaire) portant des fonctions diverses (ils représentent des nombres ou des graphèmes, ils signifient des opérations ou des instructions) et de définir le rôle de quelques mots élémentaires qui ont une fonction particulière (ils indiquent des zéros explicites/implicites, des fractions décimales; ils représentent des mots anti-répétitifs etc.).

L'optimisation du système a pour but de réduire l'inventaire théorique (mono-, di-, triphones etc.) au nombre restreint des éléments fonctionnels nécessaires au codage des systèmes de caractère différent (binaire, ternaire etc.) qui apparaissent les plus satisfaisants en ce qui concerne:

1. la façon dont ils sont discriminables et identifiables
2. la situation externe dans laquelle le stimulus est donné
3. la vitesse de production qu'ils garantissent
4. la façon dont ils se laissent mémoriser.

L'équipe du Département de Phonétique de l'Université de Trèves, en collaboration avec J.A. Dreyfus-Graf, s'est chargée d'optimiser le phonocode SOTINAKEMUŜ sur la base du comportement perceptif des auditeurs allemands.

La procédure de sélection, étant une opération aboutissant à l'inventaire des phonèmes du code choisi, reste en dehors des réflexions se référant à l'optimisation proprement dite. Produisant les éléments de base (le code SOTINAKEMUŜ nous offre cinq voyelles et six consonnes: i, e, a, o, u; t, k, m, n, s, ŝ), elle n'entre en jeu que de façon indirecte au début et, à la fin pour profiter des résultats de l'optimisation grâce à la correction de son inventaire.

Théoriquement, le processus de combinaisons des éléments de base produit un nombre illimité d'éléments fonctionnels qui serviront à constituer un code définitif à l'aide de l'optimisation. Cependant, la vitesse de production et la mémorisation justifient une limitation de la longueur des éléments. Avant que des tests d'optimisation n'aient trouvé des valeurs optimales de longueur en fonction de la vitesse de production et de la mémorisation, il faut se contenter de fixer, de façon spéculative, le nombre de phonèmes par élément fonctionnel. Nous avons été amenés à nous limiter, dans notre travail, à des mono-, di- et triphones, poussés par la nécessité de ne pas apporter aux personnes de tests et aux expérimentateurs un matériel difficilement maîtrisable. Par conséquent, le nombre théorique d'éléments fonctionnels s'élève à 11^3=1.331. Nous n'excluons pas qu' une optimisation fondée sur une mémorisation optimale, pourrait, éventuellement, corriger cette limitation en faveur d'éléments fonctionnels plus longs.

Quant à la structure des éléments fonctionnels, nous nous sommes orientés à la macro-structure syllabique de l'allemand sans, cependant, refuser des exceptions dans la mini-structure pour que les résultats de nos expériences ne soient pas uniquement limités à l'allemand mais permettent une intégration dans une conception plus vaste qui envisagera une optimisation par rapport à des besoins internationaux. La liste ci-dessous donne l'inventaire des macro-structures admises et refusées:

structures possibles	utilisées	refusées
VV	X	VV
VC	X	
CV	X	
CC		X
VVV		X
VVC	X	VVC
VCV	X	
VCC	(X)	VCC
CVV	X	CVV

CVC	X	
CCV	X	CCV
CCC		X
C		X
V	X	

V = voyelle — = éléments identiques successifs
C = consonne ( ) = certaines éliminations à cause d'une articulation difficile

Ensuite ont été éliminés les éléments qui sont difficilement articulables (2) ce qui, à la fin, donne un ensemble de 798 éléments fonctionnels, dont 5 monophones, 80 diphones et 713 triphones.

Ces éléments ont été mis en suites stochastiqueset enregistrés, sous forme de cinq conditions externes différentes, par une voix d'homme, avec une intonation descendante:

	liaison	micro	filtrage	bruit	S/B
1.	directe	dynamique	sans	sans	
2.	directe	dynamique	sans	72 dB	- 6 dB
3.	directe	dynamique	300 Hz - 3.400 Hz	sans	
4.	téléphonique	charbon	300 Hz - 3.400 Hz	sans	
5.	téléphonique	charbon	300 Hz - 3.400 Hz	60 dB	+ 6 dB

La prononciation des éléments a été adaptée à la phonétique allemande, à savoir:

- voyelles initiales + CC = courtes (a, ɛ, I, ɔ, U)
- voyelles intermédiaires + C =

Sur les bandes de tests (une par série), les stimuli ont été séparés par des pauses décroissantes de 10 à 5 secondes pour la première bande et de 5 secondes pour les autres. Les tests ont été écoutés dans un laboratoire de langues équipé d'écouteurs de haute qualité. Pendant les pauses, les sujets ont noté leur impression auditive sur une fiche de protocole.*

*Entre la première et la dernière pause, le laps de temps décroît successivement de 10 à 5 secondes pour compenser l'effet d'apprentissage.

Le groupe des personnes testées se composait de huit femmes et huit hommes âgés de plus de 30 ans et du même nombre de femmes et d'hommes plus jeunes (âgés de 20 à 30 ans). Pendant la procédure de recrutement, ceux-ci ont été soumis à un test audiométrique et à un test d'usure.*

Les résultats qui vont être présentés en détail lors des 6èmes Journées d'Etude sur la Parole, rendent compte d'une série d'informations intéressantes dans le cadre de l'optimisation du phonocode choisi et révèlent des tendances générales pour traîter tout autre phonocode. En plus, ils permettent de déduire des traits significatifs dans le comportement perceptif d'un individu confronté à des stimuli construits de manière purement formelle.

Dans une des listes établies par l'ordinateur, les phonèmes du code sont classés en fonction du taux de reconnaissance. En même temps, cette liste donne, pour chacun des éléments de base la série des phonèmes par lesquels il a été substitué. Cette information indique la hiérarchie de la valeur des éléments de base pour un code définitif, tenant compte d'une part de toutes les conditions externes et d'autre part des cinq conditions externes prises isolément. Ces résultats permettent, en plus, un pronostic sur l'amélioration du code par la neutralisation de certaines distinctions (par exemple $m \neq n \rightarrow m = n$).

Dans un deuxième type de listes, les éléments fonctionnels sont classés selon leur taux de reconnaissance et donnent, comme pour les éléments de base, en ce qui concerne chaque unité incorrecte, la substitution pour l'ensemble des éléments et pour les groupes de conditions externes. Cette liste permet aussi de décrire la dépendance de la valeur d'un élément fonctionnel en fonction de l'influence sémantique (les éléments auxquels on peut attribuer une signification quelconque sont classés relativement plus haut que les 'nonsense-elements').**

Pour établir l'échelle de valeurs des structures syllabiques des éléments, un troisième type de règles présente les éléments sous forme de construction syllabique et les structures auxquelles les personnes de tests ont eu recours pour les stimuli mal perçus.

---

*60 personnes ont été soumises à ces tests préliminaires dans lesquels elles avaient, pendant une période de 30 minutes, un problème du même caractère à résoudre de manière répétitive ce qui a mis en évidence la diminution de la concentration. Les 15 meilleurs et les 15 plus mauvais ont été éliminés après ce test.

**Il s'est avéré que les structures propres à la langue maternelle sont mieux reconnues que les structures étrangères et qu'elles ont été souvent substituées à ces dernières.

Le plan de nos recherches se proposait de démontrer que les tendances de la capacité d'intelligibilité reflètent le sexe et l'âge des personnes de tests. Les résultats préliminaires, cependant, semblent ne pas occuper une place aussi importante dans le classement du comportement perceptif de ces groupes.

Les auteurs envisagent de continuer leurs études avec des syllabes plus longues, une voix féminine pour l'enregistrement des bandes de tests et l'optimisation tenant compte de la mémorisation et de la vitesse de l'articulation.

Bibliographie

(1) Dreyfus-Graf, Jean A.: Parole codée (phonocode): reconnaissance automatique de langages naturels et artificiels. Revue d'Acoustique 21 (1972), pp. 3-12.

Dreyfus-Graf, Jean A.: Coded speech (phonocodes) and recognition machines. Proc. 8th I.C.A., London 1974.

(2) Dreyfus-Graf, Jean A.: Tests d'intelligibilité de la parole codée (phonocodes). Symposium: Intelligibilité de la parole, Liège 1973.

6èmes JOURNEES D'ETUDE SUR LA PAROLE

Toulouse 28 au 30 Mai 1975

DIALOGUES AVEC UN ROBOT.

QUINTON P, VIVES R, GRESSER JY.

CENTRE NATIONAL D'ETUDES DES TELECOMMUNICATIONS
22301 - LANNION

RESUME

Partant d'une caractérisation générale des machines à dialoguer par la parole, on analyse différentes façons de tirer parti du dialogue, pour l'amélioration de la reconnaissance automatique de la parole ou pour celle de la communication homme-machine. On développe quelques enseignements apportés par une expérience de simulation de dialogue homme-machine.

SUMMARY

Starting from a general characterization of spoken dialog systems, various ways of using the dialog in order to enhance either the recognition scores or man-machine communication are analyzed. Conclusions are drawn from the simulation of a spoken man-machine communication in a real-use situation.

"Dialogues" avec un robot (*)

QUINTON P, VIVES R, GRESSER JY (**)

Introduction

La technologie moderne met l'homme en face de machines de plus en plus complexes, avec lesquelles il lui faut communiquer : il leur fournit une information (ne serait-ce que la simple mise en marche), et elles lui répondent par des voyants, des codes, des bruits....etc.. par lesquels il peut déterminer son action. Les ordinateurs ont une place privilégiée dans ce domaine, car leurs moyens de calcul permettent d'utiliser des supports de plus en plus élaborés pour communiquer.

L'état d'avancement des recherches sur la reconnaissance automatique de la parole incite aujourd'hui à utiliser la parole comme support de communication, ceci afin de se libérer des contraintes inhérentes à l'emploi des capteurs classiques (télétypes, lecteurs de cartes...). Par rapport aux processus classiques, la mise en oeuvre d'une machine à dialoguer présente des particularités qui ne résident pas seulement dans les techniques de captage et d'analyse : il faut s'efforcer de tirer le meilleur parti possible de la souplesse naturelle du dialogue afin d'une part, de compenser les erreurs dues à la reconnaissance, et d'autre part de limiter au maximum les contraintes dont on veut se libérer.

A partir des réalisations actuelles, nous essayons dans une première partie de dégager les orientations du dialogue, et dans la deuxième partie, nous décrirons une expérience de simulation et ses premiers enseignements.

(*) Le mot "dialogue" fait penser à SOCRATE, à cette recherche commune du maître et de l'élève, réalité dialectique et médiatrice, visant l'être et le monde dans sa totalité. Il fait penser aussi à GREGOIRE LE GRAND, au jeu rhétorique des questions du dialogue scolastique, aux conversations des philosophes du XVIIè siècle, et, bien sûr, à la philosophie du dialogue de M. BUBER, où la plénitude de la relation d'infériorité réciproque, entre un je et un tu, définit la réalité du dialogue.

Entre un homme et un robot de 1975 il ne peut s'agir de vrai dialogue mais de dialogue purement technique, échanges d'ordres et de questions, visant à établir, peu à peu, une coexistance empirique, entre deux êtres de nature différente. Mais le robot est-il ou n'est-il qu'un pâle reflet de l'ingénieur ?

Au mot "dialogue" on préfère souvent "communication" transmission de messages et de leurs significations, limitant la relation entre deux "produits" au schéma élémentaire de la théorie de l'information. Le mot s'applique sans discrimination aux êtres ou aux machines dans des situations diverses, vues sous l'angle physique, luiguistique ou sociologique. Pour cela il nous parait trop général. Nous préférons l'emploi de "dialogue" dans "dialogue homme-machine", pour exprimer le processus de communication ayant la parole pour support.

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

I - Caractérisation des machines à dialoguer.

Les distributeurs automatiques de boissons se servent des capacités physiques et intellectuelles de l'utilisateur. Il faut lire le mode d'emploi, introduire le bon nombre de pièces dans la machine, sélectionner à l'aide de boutons un type de boisson , ouvrir un casier pour se servir, récupérer la monnaie.... L'utilisateur ne se désaltère qu'après avoir participé plus ou moins activement à un certain nombre de manoeuvres.

Pour rendre performant les systèmes basés sur la communication homme-machine par la voix il est nécessaire de compter sur la compétence de dialogue de l'utilisateur pour

- corriger les mauvaises interprétations de la machine,
- préciser ou répéter les éléments de messages mal reconnus,
- s'adapter à la machine,
- participer à l'adaptation de la machine au locuteur,

machine qui doit en retour posséder une compétence minimale.

I-1 - Schéma général

Dans une caractérisation des machines à dialoguer interviennent un bon nombre de coordonnées d'une machine à reconnaitre.

a) Au niveau du captage, on doit spécifier dans quel bruit ambiant peut avoir lieu le dialogue. La nature du capteur (numéro, téléphone, interphone, visiophone, etc) n'est pas sans incidence sur la forme du dialogue : on dit plus naturellement "allo" dans un combiné téléphonique que dans un micro.

b) Au niveau du décodage interviennent des paramètres comme l'élocution (mots isolés, parole continue, dialogue naturel), le nombre et le type des locuteurs, la nature du langage accepté (souplesse de l'analyse grammaticale, le taux d'erreurs acceptable, l'adaptabilité du langage) ; le nombre de mots du vocabulaire, la possibilité d'employer des synonymes.

c) C'est la sémantique et la pragmatique qui semblent jouer le rôle le plus important dans un dialogue avec un robot. Le niveau de complexité d'un pas du dialogue peut aller de s'assurer de la présence d'un mot ou d'un syntagme dans une phrase, jusqu'à traduire en langage interne une question complexe.

On ne peut bien souvent donner un sens à un message s'il est isolé du contexte où il a été prononcé : il sera parfois nécessaire qu'une machine à dialoguer analyse dans le temps la structure du dialogue pour lever des ambiguités sémantiques (4).

d) A l'encodage les paramètres principaux sont la génération des messages synthétisés, les liens éventuels avec le décodage et la rapidité. On peut prévoir en effet des systèmes synthétisant automatiquement des réparties fonction des questions posées ou plus simplement une synthèse de messages préétablis.

e) A l'émission on retrouve les principales caractéristiques du captage à savoir : les conditions générales de bruit, la nature de l'émetteur, la simultanéité possible avec le captage.

1-2 - Analyse des dialogues des systèmes existants.

Les plus simples des dialogues avec un robot ont été élaborés à l'aide de machines à reconnaître des mots isolés (5), (9), (10), (11). L'introduction du dialogue est réalisé en faisant synthétiser un message correspondant à ce que la machine reconnait.

Même sur les exemples les plus simples, les expériences de dialogues avec un robot ont montré que le dialogue en lui-même pouvait être une source de connaissance très performant. Une équipe de l'université de Stanford (5) a construit un système capable de s'adapter au locuteur. Grâce à un dialogue du style,

Calculateur : what is 7 minus 3 ?
Elève : 4
Calculateur : yup, that's right.
Calculateur : 6 over 3 is what ?
Elève : 2
Calculateur : Did you say 0 ?
Elève : no
Calculateur : Sorry my mistake - please say 2.
Elève : 2
Calculateur : Thank you !

La machine peut ajuster ses paramètres acoustiques pour le nouveau locuteur. On peut encore simplement tirer parti du dialogue en donnant la possibilité à l'utilisateur de corriger une mauvaise interprétation de la machine (6). Dans notre système saisie de données C.A.O. par la parole (8), on peut trouver la séquence suivante :

Utilisateur : je commence par la liste des connecteurs.
Calculateur : liste des connecteurs.
Utilisateur : entrée numéro six
Calculateur : entrée numéro dix (pour "six" la machine reconnait "dix")
Utilisateur : non, entrée numéro six
Calculateur : entrée numéro six.

Pour "six" la machine retrouve "dix" comme réponse la plus probable mais on tire parti du dialogue en donnant la seconde interprétation possible qui se trouve être "six".

Dans les systèmes de consultation de fichiers, la plupart des équipes (1), (2), (3), (4), (7), (8) ont choisi la parole continue. Si les dialogues paraissent plus naturels les machines peuvent devenir très complexes.

Les équipes de S.D.C. (3) ou de B.B.N (1) développent des systèmes où le dialogue travaille sur une base de données figée : renseignements concernant les flottes sous-marines de certains pays pour S.D.C., base de données géologique et minéralogique pour le système Lunar de B.B.N.

Les questions sont du type

"Total quantity where type equals nuclear and country equals USA" ou "What is the average concentration of rubidium in high-alkali rocks ?".

D'autres équipes (2), (4), (7), élaborent des systèmes plus sophistiqués où le dialogue peut modifier certaines parties, comme en (5), de la base de données. Le jeu d'échec de C.M.U. ou le système d'assistance en réparation de petits organes électromécaniques du S.R.I. nécessitent une analyse profonde de la structure du dialogue pour lever certaines ambiguités.

1-3 - Spécificité du dialogue

Y a-t'il dans les systèmes à dialoguer des éléments qui ne figurent ni dans les systèmes de communication sur le mode écrit, ni dans les systèmes de reconnaissance automatique de la parole ?

La situation de dialogue sur un élément d'information est supplémentaire pour chacun des interlocuteurs, humain ou machine, sur l'autre. Il y a différents moyens de l'incorporer au reconnaisseur classique.

Le modèle du monde de l'automate peut contenir une représentation du locuteur, dans sa forme la plus simple satisfait ou non satisfait (8,9).

Tout ou partie d'un niveau d'analyse ("pragmatique", "thématique"...) s'ajoute aux niveaux usuels (acoustique , phonétique , lexical, syntaxique, sémantique) (1,3,8) le second procédé est complémentaire ou supplémentaire du premier. L'analyse peut s'effectuer en profondeur, reconnaissance et dialogue sont intimement

mêlés (1,3). Elle peut apparaitre nettement à la surface des échanges, par la mise en ordre des questions et réponses, pour l'ensemble du dialogue (9, 15, 8 actuel) ou pour certaines phases (4).

L'expérience décrite en 2, présente un dialogue mis en forme globalement.

2 - Une expérience de simulation de dialogue

On n'a que peu de renseignements sur les réactions d'une personne qui dialogue avec une machine. C'est pourtant une donnée à priori que l'on ne peut pas modifier, et dont il faut tenir le plus grand compte si l'on veut que le dialogue soit possible. Afin de mettre en évidence les problèmes qui peuvent se poser, nous avons défini un dialogue simple que nous avons simulé.

2-1 - L'expérience

Ce dialogue permet de demander le numéro de poste, le grade ou le numéro de bureau d'une personne du Groupement C.E.I. du CNET (comprenant environ 100 personnes). L'expérience se passe de la manière suivante :

- On appelle par téléphone quelqu'un à qui l'on demande de bien vouloir essayer un prototype de centre de renseignement, en précisant bien que la formulation de la question est libre.

- Le téléphone est ensuite branché sur un vocodeur, fonctionnant en mode local. La voix du simulateur est ainsi déformée avant de parvenir à l'interlocuteur (Voir figure)

- Le dialogue s'engage : le simulateur a pour rôle de respecter au mieux l'organigramme du dialogue défini au préalable, notamment en ce qui concerne les critères d'interprétation sémantique des questions.

L'expérience a été faite avec 8 interlocuteurs différents. Au total, le cycle élémentaire de l'organigramme du dialogue, correspondant à la recherche d'un renseignement, a été parcouru une trentaine de fois. Un exemple de dialogue est donné en annexe.

2-2 - Les résultats

2-2-1 - Validité de l'expérience

Toutes les personnes qui ont participé à l'expérience ont réellement cru que le dialogue était automatisé. Leurs

réactions étaient donc spontanées, et on peut considérer que les enregistrements sont significatifs de ce qui se passerait avec une véritable machine à dialoguer. Il faudrait bien sûr refaire l'expérience avec un échantillon représentatif de la population pour en avoir la certitude.

2-2-2 - Nature du dialogue

Comme on pouvait s'y attendre, la formulation des questions est quasi-standard ; il n'y a pas beaucoup de façon de demander un renseignement de ce genre. Les critères sémantiques que nous avions choisi étaient volontairement très simples : il s'agissait de reconnaître quelques mots clés dans la question (numéro, grade, bureau, monsieur, madame, etc...). En supposant que la reconnaissance de ces mots se passe correctement, ce qui est imaginable dans l'état actuel de la reconnaissance, le dialogue aurait été orienté correctement dans tous les cas.

2-2-3 - Les erreurs et les malentendus

Les causes des malentendus sont très diverses, et il est intéressant de constater que les quelques dialogues nous ont fourni à peu près tous les types de malentendus que l'on peut à priori imaginer. On peut ainsi avoir une idée de la réaction de l'interlocuteur face à un évènement accidentel dans le dialogue : question mal formulée par la machine, ou incompréhensible, etc... Inversement, on peut essayer de détecter dans la réponse de l'interlocuteur un trait indiquant que sa réaction n'est pas "normale". Bien qu'il semble prématuré de s'attaquer à un problème aussi complexe, cela suggère que des études soient entreprises pour aboutir à des dialogues acceptant ces malentendus et réagissant de manière variée en fonction de leur nature. A défaut, un "modèle des erreurs" peut faciliter l'évaluation des performances d'un prototype.

Conclusion

Les obstacles auquels se heurte la réalisation d'une machine de reconnaissance amènent à essayer d'utiliser toutes les ressources du dialogue, notamment en acceptant certaines erreurs. Il est assez surprenant de constater que cela a été peu développé dans les systèmes conversationnels par télétype qui tolèrent rarement des fautes, en dehors des cours assistés par ordinateur.

Ceci demande l'approfondissement de la structure du dialogue, de l'influence du contexte sur le sens d'une phrase, de l'interdépendance à la fois syntaxique et sémantique entre une question et sa réponse. En attendant, il n'est possible d'aboutir à des échanges cohérents, que sur des sujets très limités.

BIBLIOGRAPHIE :

(1) WOODS W.A.(BBN) : Motivation and Overview of BBN Speechlis, An Experimental Prototype for Speech Understanding Research. Proc.IEEE Symp. Speech Recognition, CMU, (April 1974).

(2) Baker J.K.(CMU) The Dragon System - An overview - Proc.IEEE Symp. Speech Recognition,CMU (April 1974).

(3) Ritea H.B. (SDC): A voice-controlled Data Management System, Proc. IEEE Symp. Speech Recognition, CMU, (April 1974).

(4) Deutsch B.G.(SRI) : The Structure of task oriented dialogs - Proc. IEEE Symp. Speech Recognition,CMU, (April 1974).

(5) Danforth D.G., Rogosa D.R., Suppes P. (Stanford U.) Speaking of learning models, Proc.IEEE Symp. Speech Recognition, CMU,(April 1974).

(6) Holden A.D.C.,Stasbourger E., Price L.(U.Washington et al.) A computer programming system using continuous speech input- Proc.IEEE Symp.Speech Recognition,CMU(April 1974).

(7) Lesser V.P., Fenneil R.D., Erman L.D., Reddy D.R.(CMU) Organization of the Hearsay Speech Understanding System, Proc. IEEE Symp. Speech Recognition, CMU(April 1974).

(8) Mercier G., Buisson L., Vivès R., Genin J. Premier compte rendu semestriel relatif au contrat concernant deux langages parlés pour un dialogue intelligent entre l'homme et la machine.

(9) Gresser J.Y., Mercier G., Un exemple de reconnaissance automatique de la parole, Commutation Electronique n° 32, janvier 1971, N°32, p 48-63.

(10) MartinT.B.(T.T.I.) 1973 annual report.

(11) Vicens P. : Aspects of speech recognition by computer, Stanford A.I. Project, Memo AI-85, CS 127 (1969).

(12) Winograd, T. Procedures as a Representation for Data in a Computer Program for understanding Natural Language. Thesis. (1970)

(14) Weizenbaum, J. Contextual Understanding by Computer ,in Recognizing Patterns, Kolers P.A and Eden M editors.

(15) + Gresser et al. Rec. autom. de langage...N.T. CEI/CSI/32 CNET, juin 13 - 58 p, 17 réf.

ANNEXE

EXEMPLE DE DIALOGUE SIMULE.

Centre de renseignements : Centre de renseignements, je vous écoute.

Interlocuteur : Allo ! Euh ! je voudrais connaître le numéro de téléphone de Madame X.

CDR : De quel département ?

I : CEI.

CDR : CEI ?

La machine répète de manière interrogative ce qu'elle a compris.

I : .... (silence).

L'intonation n'était sans doute pas convaincante.

CDR : CEI ?

I : Oui.

Recherche...

CDR : Le numéro de téléphone de Madame X est le 2555. C'est ce que vous vouliez savoir ?

I : Euh ! non..., j'aurais voulu aussi connaître son grade.

La question de la machine est mal formulée : elle a été confondue avec c'est tout ce que vous vouliez savoir. Du même coup, la réponse de l'interlocuteur est incompréhensible pour la machine.

CDR : Pardon ?

I : J'aurais voulu savoir son grade.

CDR : Désolé, je n'ai pas compris. Je vous écoute.

La machine attendait désespérément une réponse de type oui-non... Retour au début de l'organigramme.

I : Euh ! quel est le grade de Madame X ?

L'interlocuteur rectifie la forme de sa question, pensant avec juste raison que la machine n'a pas prévu les adjectifs possessifs... Sous cette forme, la question est compréhensible par la machine.

etc...

CONFIGURATION UTILISEE PENDANT LE DIALOGUE

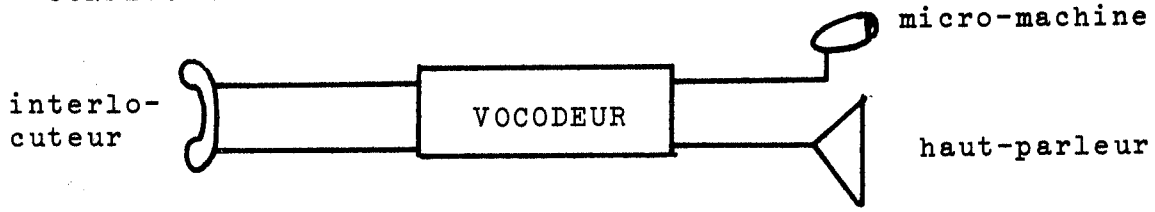


TABLE DES MATIERES
DU VOLUME I

THEME IIIB - SORTIE PARLEE

THEME IIIC - COMMANDE DES PROGRAMMES