COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL
ET SES APPLICATIONS

NICE du 16 au 21 JUIN 75

ANALYSE D'UN SIGNAL FORTEMENT STRUCTURE : LE SIGNAL VOCAL

D. DOURS*, R. FACCA*, G. PERENNOU**

Laboratoire C.E.R.F.I.A. Université Paul Sabatier 31077 TOULOUSE CEDEX

RESUME

On peut considérer le signal vocal comme une succession de segments, chacun d'eux ayant une forme permettant de le ranger dans une catégorie phonémique (en principe). Les types de ces segments sont divers et donnent lieu à des sous catégories telles que : voyelles-consonnes, puis encore sonores-sourdes etc...
Si l'on désire reconnaître le signal parlé, c'est-à-dire par exemple, retranscrire en mots écrits le message vocal, il convient :

1° - de procéder à la segmentation du signal en intervalles s'identifiant à un seul phonème ou à une seule catégorie de phonèmes ;
2° - d'identifier la catégorie, puis la sous catégorie du segment etc... ;
3° - de reconnaître le segment.

C'est là la phase analytique préliminaire à partir de laquelle il faudra de nouveau reconstruire, conformément au lexique et à la grammaire, le message initial, par exemple dans le langage écrit. Il se pose à ce niveau des problèmes délicats qui ne sont pas développés dans cet exposé.
Nous nous proposons de décrire le module d'analyse et de reconnaissance du signal vocal mis au point dans le cadre du projet A.R.I.A. (Analyse et Reconnaissance de l'Information Acoustique), conformément aux indications ci-dessus.
Nous développons également une méthode d'analyse des sons voisés, basée sur la détection du signal glottal et l'identification de la réponse impulsionnelle du conduit vocal.
On propose des exemples de signaux vocaux traités par cet analyseur et en ce qui concerne les segments voisés, des comparaisons sont faites avec les résultats obtenus par F.F.T.

* Assistant
** Professeur } à l'Université P. Sabatier - TOULOUSE

SUMMARY

The vocal signal can be considered as a series of segments, each of them having a form allowing for its classification in a phonemic category (that is theoretically). The types of these segments are diverse and can be divided into sub categories such as vowels, consonnants and again into voiced-invoiced,etc.
If we wish to recognize the spoken signal, that is for example, to transcribe into written words the vocal message, we have to :

1° - undertake the segmentation of the signal into intervals corresponding to a single phoneme or a single category of phonemes ;
2° - identify the category, then the sub-category of the segment etc...;
3° - identify the segment.

This is the preliminary analytical sequence from which we shall have to reassemble the initial message according to lexicon and grammar, for instance in the written language. There appears at this point critical problems that are not considered in this article.
We intend to describe the vocal signal analysis and recognition module developped in the A.R.I.A. project (Analyse et Reconnaissance de l'Information Acoustique) according to the above indications.
We are also working on a voiced sound analysis method based on the detection of the glottal signal and the identification of the impulsional response of the vocal tact.
We present examples of vocal signals treated by this analyses and concerning voiced segments, we wake comparisons with the results obtained by F.F.T.

ANALYSE D'UN SIGNAL FORTEMENT STRUCTURE : LE SIGNAL VOCAL

INTRODUCTION

Le signal vocal est destiné à transmettre un message dans une langue déterminée. Toute langue possède des unités minimales, auxquelles toute expression peut être réduite : ce sont les phonèmes. Ces éléments composent les unités supérieures (syllabes, etc) dont est formé un énoncé, le syntagme.

Au niveau physique le signal peut être décomposé en une suite d'intervalles tels que chacun d'eux corresponde à la réalisation d'un phonème. La segmentation en de tels intervalles pose déjà un problème car la transition d'un intervalle au suivant n'est pas nette, des procédures complexes doivent être mises en jeu pour le résoudre. Un deuxième problème est celui de la reconnaissance du signal porté par un intervalle, c'est-à-dire son identification à un phonème. Il est évident que cette reconnaissance sera facilitée si les paramètres extraits lors de l'analyse sont de bonne qualité. Ces deux problèmes étant résolus, on peut associer au signal une suite de symboles phonétiques. Cette suite n'est pas parfaite, car le signal lui-même peut ne pas indiquer clairement un signe déterminé et le système d'analyse et de reconnaissance n'est pas nécessairement optimal. De plus, dans la parole continue, les mots consécutifs ne sont pas physiquement séparés. Ainsi, au niveau de l'interprétation du message correspondant au signal, se posera le problème de la décomposition de la chaîne en une suite de mots du lexique, compte tenu des erreurs possibles au niveau de l'identification des signes et des variantes phonétiques de chaque mot du lexique qu'introduisent les différentes prononciations, les liaisons, etc... Nous ne parlerons pas ici de cet aspect du problème qui se situe au niveau linguistique.

Nous nous proposons de décrire le module d'analyse et de reconnaissance du signal vocal (fig. 1) mis au point dans le cadre du projet A.R.I.A.(1) (Analyse et Reconnaissance de l'Information Acoustique), conformément aux indications ci-dessus et de développer la méthode d'analyse des sons voisés.

SEGMENTATION DU SIGNAL VOCAL

La segmentation du signal vocal a pour but :

1° dans les zones de parole, de distinguer les sons voisés (produits par la source glottale) des sons non voisés (dus aux turbulences de l'air dans les constrictions du canal vocal) ; pour les sons voisés on détermine les périodes du fondamental, chaque période est ensuite découpée en deux intervalles T_f et T_0 pendant lesquels la glotte est respectivement fermée et ouverte ; les sons non voisés sont découpés en intervalles successifs de 10 msec, on distingue ensuite les sons fricatifs sourds (signaux aléatoires stationnaires dont l'énergie est dans la bande 2000 Hz-8000 Hz) des sons plosifs (caractérisés par une occlusion engendrant un silence (ou un buzz) dont la durée est de l'ordre de 60 msec, suivie d'une plosion).

2° démarquer les phonèmes entre eux à l'aide des variations de la courbe d'énergie. La fonction $\varphi(t)$ égale à 1 à la frontière de deux phonèmes et 0 ailleurs est alors générée.

Nous allons maintenant préciser les mécanismes de cette segmentation. Le signal s(t) étant préalablement échantillonné, $s(t) \rightarrow (s_1, s_2, \ldots, s_n, \ldots)$ on associe à chaque intervalle $S_k = (s_{(k-1)N+1}, s_{(k-1)N+2}, \ldots, s_{kN})$ les nombres suivants :

$$C_k = \frac{1}{N} \sum_{i=1}^{N} |s_{(k-1)N+i}|, \quad V_k = \frac{1}{2}(C_{k-1} + C_k).$$

Pour un échantillonnage à 15 KHz nous avons pris N=80.

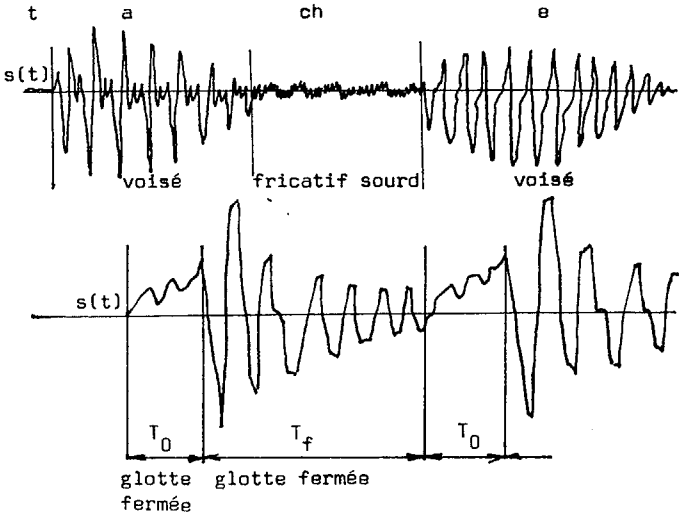
Après fixation d'un seuil $\theta > 0$, la distinction voisé-non voisé s'effectue comme suit :

si $V_k \geq \theta$ S_k est considéré comme voisé et on pose $R(V_k) = F$

Si $V_k < \theta$ S_k est considéré comme non voisé et $R(V_k) = V$.

Segmentation des sons voisés : Des suites de I_L (I_L=6) intervalles S_k consécutifs voisés sont constituées (Notées P(i) dans l'organigramme Fig.2). Un algorithme pour lequel nous renvoyons à [1], est utilisé pour détecter le début des intervalles T_f. Ceci est représenté par la fonction $\pi(t)$, qui vaut 1 à ces instants et zéro ailleurs. A postériori, l'indicateur de voisement $R(V_k)$ peut être rétabli à la valeur V.

Segmentation des sons non voisés : On utilise ici une technique de passage à zéro. Si la fréquence est supérieure à 2000 Hz, le segment est classé fricatif, sinon il est assimilé à un silence. La fonction Y(t) qui vaut 1 tous les 2 N échantillons et 0 ailleurs ainsi que la fonction $\delta(t)$ qui vaut +1 (resp -1, resp 0) en fin de silence (resp. en début de silence, resp. ailleurs) sont alors générées.



ANALYSE D'UN SIGNAL FORTEMENT STRUCTURE :
LE SIGNAL VOCAL

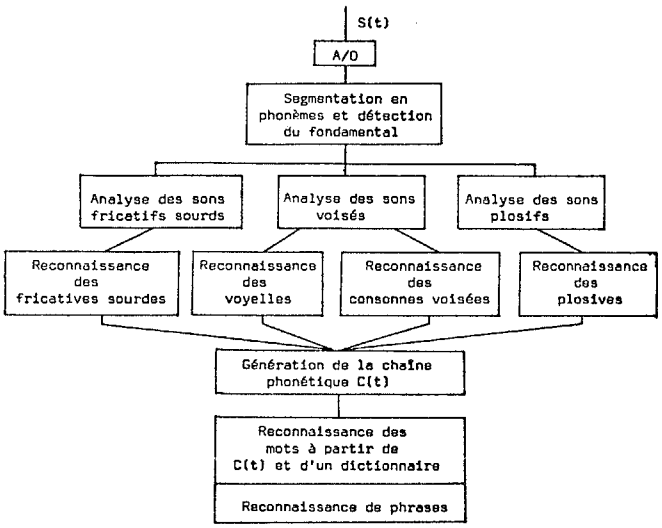


Fig. 1 : Système d'analyse et de reconnaissance du projet A.R.I.A.

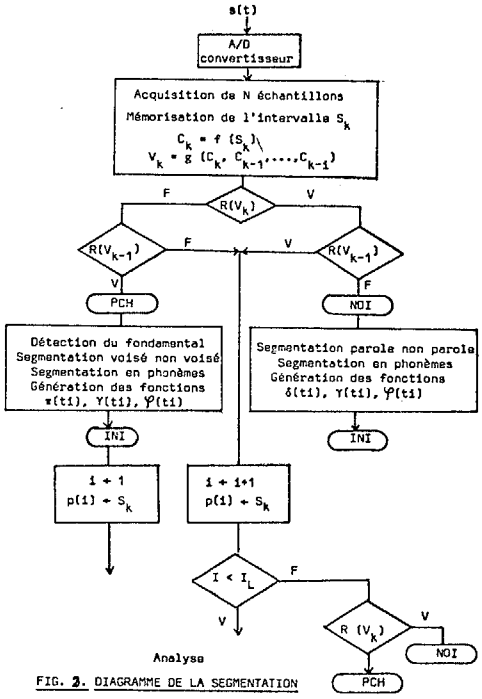


FIG. 2. DIAGRAMME DE LA SEGMENTATION

ANALYSE DES SONS VOISES

Les méthodes classiques d'analyse spectrale, en dépit de leurs avantages, notamment la simplicité de mise en oeuvre et la rapidité d'analyse (vocoder, bancs de filtres (11)) présentent un certain nombre d'inconvénients bien connus lorsqu'il s'agit du signal parlé. En effet, la modulation du spectre par le fondamental, la définition d'un spectre à court terme, le caractère convolutionnel de la production de la parole, font que ces méthodes ne donnent pas toujours les résultats escomptés en ce qui concerne la précision des paramètres obtenus.

Diverses méthodes ont été proposées pour pallier ces inconvénients. Elles sont en général fondées sur l'analyse directe du signal vocal et utilisent toutes le modèle proposé par FANT (2). Dans ce modèle, on considère le signal vocal comme la réponse s(t) d'un système acoustique f(t) [conduit vocal, conduit nasal, radiation buccale,...] à une excitation de type fixé, e (t). Cette excitation est, soit une série d'impulsions quasi-périodiques provenant des cordes vocales dans le cas des sons voisés, soit un bruit blanc dû aux turbulences de l'air dans les constrictions du canal vocal dans le cas des sons non voisés.

Si l'on suppose que pendant une période du fondamental, le système de phonation est linéaire et stationnaire, on a la relation :

$$s(t) = e(t) * f(t) + g(t)$$

dans laquelle :

f(t) est la réponse impulsionnelle du système

g(t) est la réponse libre du système avec conditions initiales.

Le modèle que nous venons de décrire suggère diverses méthodes d'analyse ayant toutes pour but l'obtention d'un ensemble de paramètres caractérisant le signal vocal. Ces paramètres peuvent être obtenus :

- par identification de la fonction de transfert (transformée en Z) du système de phonation sur chaque période du fondamental ;
- par identification de la réponse impulsionnelle du système de phonation sur chaque période du fondamental.

Les méthodes de prédiction linéaire (codage prédictif) sont basées sur l'identification de la fonction de transfert. Il existe deux types de méthodes :

- les méthodes de covariance parmi lesquelles on peut citer :
  - la méthode d'ATAL-HANAUER (3), (4)
  - le filtrage optimal de Kalman de GUEGUEN et CARAYANNIS (5)
- les méthodes d'autocorrélation avec principalement :
  - le filtrage digital inverse de MARKEL (6)
  - la méthode de ITAKURA-SAITO (7), (8).

Ces méthodes qui ne diffèrent que par le processus de résolution adopté, permettent si on le désire, de déterminer les pôles de la fonction de transfert à partir des paramètres obtenus.

La méthode d'analyse que nous développons (1),(11) est une méthode directe en synchronisme avec le fondamental. Elle s'apparente à celle proposée par E.N. PINSON (10) en ce sens qu'elle consiste à identifier la réponse impulsionnelle du système de phonation sur chaque période du fondamental, dans des intervalles de temps pendant lesquels le signal glottal est négligeable.

ANALYSE D'UN SIGNAL FORTEMENT STRUCTURE :

'E SIGNAL VOCAL

1. MODULE D'ANALYSE DES SONS VOISES.

Partant du modèle de FANT, il est facile de vérifier que sous les hypothèses classiques (canal vocal assimilé à une série de résonnateurs en cascade), la réponse libre du système est de la forme :

$$\hat{f}_N(P,t) = a_0 + \sum_{i=1}^{N} e^{-\pi B_i t} (a_i \sin 2\pi F_i t + b_i \cos 2\pi F_i t)$$

avec F_i : fréquence du $i^{ème}$ formant

B_i : caractérise l'amortissement du $i^{ème}$ formant

a_i et b_i caractérise l'amplitude de la fréquence F_i

N est le nombre de formants. Il dépend du modèle choisi (1 à 4)

P est le vecteur paramètre tel que :

$$P = (P_\beta, P_\xi)$$

avec : $P_\beta = (a_0, (a_i, b_i))\ i = 1,N$

$P_\xi = (B_i, F_i)\ i = 1,N$

1.1. Méthode générale d'approximation

Soit $s(t) = f(t) + b(t)$ le signal observé pendant que la glotte est fermée (intervalle T_f), dans lequel $b(t)$ est un bruit stationnaire indépendant du signal et $f(t)$ le signal à approcher.

Le problème consiste à identifier le vecteur P qui minimise un critère déterminé. Nous avons choisi l'erreur quadratique moyenne :

$$E(P) = \int_{T_f} (s(t) - \hat{f}_N(P,t))^2 \, dt.$$

Il est facile de vérifier que E(P) est une forme quadratique par rapport aux composantes de P_β.

En posant $\underset{P}{\text{grad}}\, E(P) = (\underset{P_\beta}{\text{grad}}\, E(P), \underset{P_\xi}{\text{grad}}\, E(P))$, on résoud le problème de la façon suivante :

1° - pour P_ξ fixé, on cherche P_β qui minimise E(P) ;

comme $\underset{P_\beta}{\text{grad}}\, E(P)$ est linéaire en P_β, il suffit pour cela de résoudre un système linéaire.

2° - En fixant P_β à la valeur trouvée au 1°, on cherche P_ξ qui minimise E(P) ; nous avons choisi une méthode du gradient à pas séparés.

A partir de P_ξ ainsi obtenu, on itère le processus tant que E(P) décroit. La difficulté de la méthode réside dans le choix de la valeur initiale P_{ξ_0}. En effet, elle doit être choisie soigneusement, car E(P) n'est pas une fonction convexe. Pour l'obtenir, on peut initialiser le processus à partir de différentes valeurs P_{ξ_i} appartenant à l'ensemble des valeurs possibles Ω. On ne conserve alors que celle relative à la meilleure erreur quadratique. Théoriquement on ne peut affirmer avec certitude que le minimum global soit atteint. Pourtant en pratique si l'on prend pour ensemble Ω, tous les P_{ξ_i} correspondant aux différents phonèmes de la langue, on obtient de bons résultats.

L'inconvénient de cette méthode est qu'elle nécessite des temps de calcul prohibitifs vu le très grand nombre de réalisations possibles des phonèmes.

On peut également obtenir le paramètre P_{ξ_0} par une méthode d'approximation séparée des formants qui ne nécessite que des temps de calcul beaucoup plus courts.

1.2. Méthode d'approximation séparée des formants.

On peut montrer (10) que, même si le signal vocal comporte 3 ou 4 formants, ce qui est le cas usuellement, les tentatives d'identification avec un modèle à un seul formant $\hat{f}_1(P,t)$ permettent d'obtenir successivement les différents formants, en tant que minimum de l'erreur quadratique E(P). Différentes méthodes pratiques peuvent être mises en oeuvre. La méthode dite des points de convergence est celle que nous avons retenue (1). L'avantage est que le nombre de paramètres de $\hat{f}_1(P,t)$ étant restreint, les calculs en sont allégés.

1.3. Principe d'utilisation des différentes méthodes.

La méthode d'approximation séparée des formants donne des résultats suffisamment précis pour être exploitée seule. Cependant elle nécessite un temps de calcul relativement long, c'est pourquoi on ne l'utilise que comme processus d'initialisation.

Processus d'initialisation. On appelle ainsi toute la procédure concernant la recherche séparée des formants. Elle comprend l'algorithme d'approximation d'un formant ainsi que toute la logique de détection des points de convergence. Elle fournit le paramètre P_{ξ_0} directement expoitable par le processus de suivie.

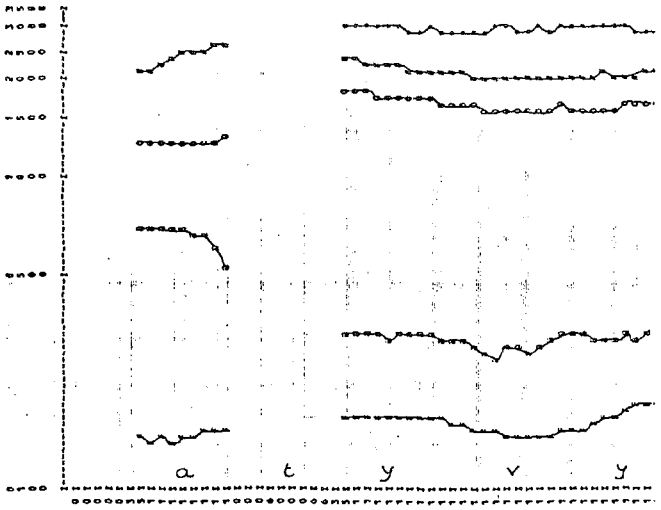
Processus de suivie. C'est la procédure qui, partant de P_{ξ_0} permet d'obtenir le vecteur P qui minimise E(P).

Mode d'enchaînement. Dans le cas où le segment analysé est le début d'une zone voisée, on ne connaît absolument rien sur le signal, on passe obligatoirement par le processus d'initialisation qui fournit P_{ξ_0} au processus de suivie pour affiner les résultats. Si ce n'est pas le début d'une zone voisée, on suppose que le signal évolue peu d'une période à l'autre. Le paramètre P_ξ obtenu sur la période précédente est alors fourni au processus de suivie. Si l'algorithme converge, c'est que l'hypothèse faite sur la stationnarité du signal est vérifiée. S'il ne converge pas c'est que le signal évolue rapidement, on passe alors au processus d'initialisation.

ANALYSE D'UN SIGNAL FORTEMENT STRUCTURE :
LE SIGNAL VOCAL

RESULTATS ET COMPARAISON

Les résultats de l'analyse sont donnés d'une part sous forme de "sonagrammes", d'autre part sous forme de tableau de chiffres. La figure suivante représente le sonagramme de "As-tu-vu". On peut remarquer la grande précision avec laquelle les fréquences des formants sont obtenues, dans un cas pourtant difficile, les 2ème, 3ème et 4ème formants étant très proches les uns des autres.



Nous avons comparé nos résultats avec ceux obtenus par F.F.T. Il ressort de cette comparaison une concordance des résultats évidente (fig. 9 et 10) dans le cas où la F.F.T. donne de bons résultats.

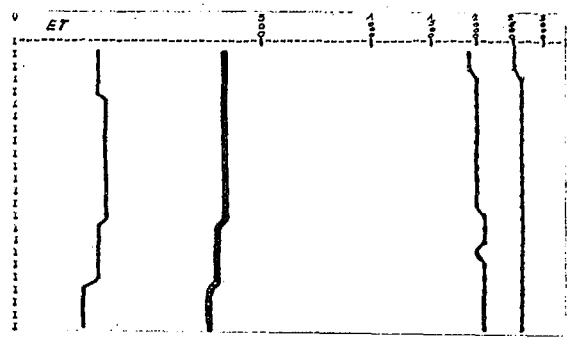


Fig : 9 ANALYSE TEMPORELLE. LOCUTEUR MASCULIN (DD) FONDAMENTAL (125 Hz).

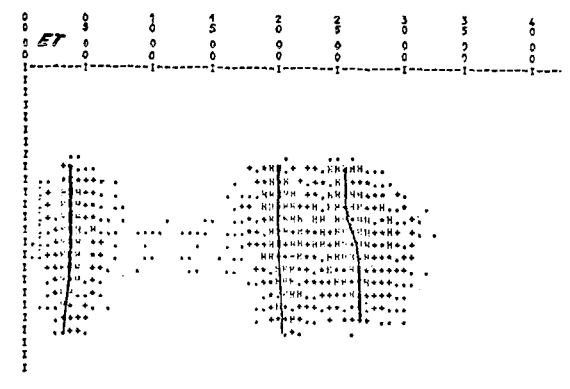


Fig : 10 ANALYSE FREQUENTIELLE. LOCUTEUR MASCULIN (DD) FONDAMENTAL 125 Hz. FENETRE 15 ms.

On peut noter également qu'il est difficile de distinguer le 2ème formant du 1er formant dans l'analyse par F.F.T. (figure 16), alors que ceux-ci apparaissent clairement dans l'analyse temporelle (figure 15). Ce phénomène se reproduit chaque fois que deux formants sont proches, que ce soient F_1 et F_2 dans les voyelles d'arrière ou F_2 et F_3 pour la voyelle |y|.

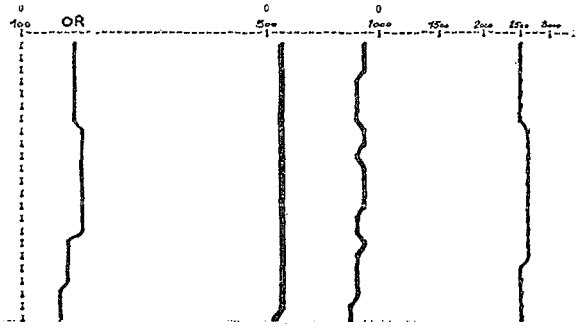


Fig :15 ANALYSE TEMPORELLE, LOCUTEUR MASCULIN (RF) FONDAMENTAL 130 Hz.

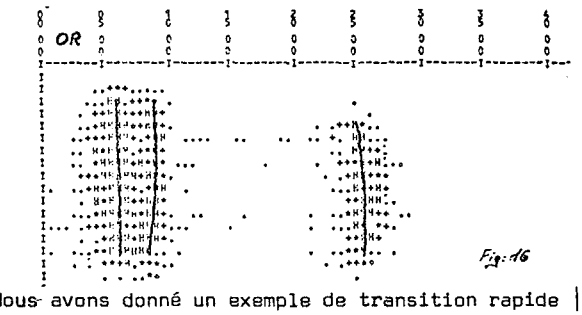


Fig:16

Nous avons donné un exemple de transition rapide |u|, |i| pour montrer l'efficacité de la méthode temporelle dans ces cas-là (figures 17 et 18).

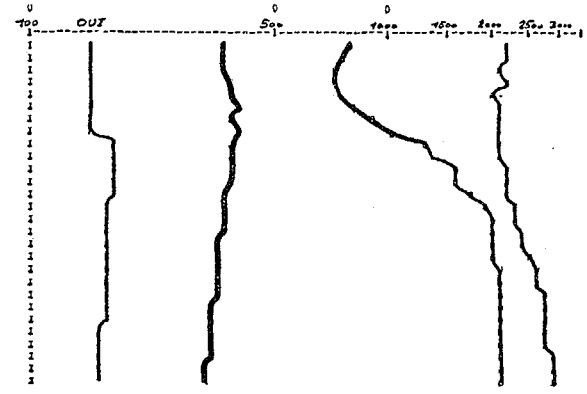


Fig : 17 ANALYSE TEMPORELLE - LOCUTEUR MASCULIN (D.D) FONDAMENTAL 125 Hz.

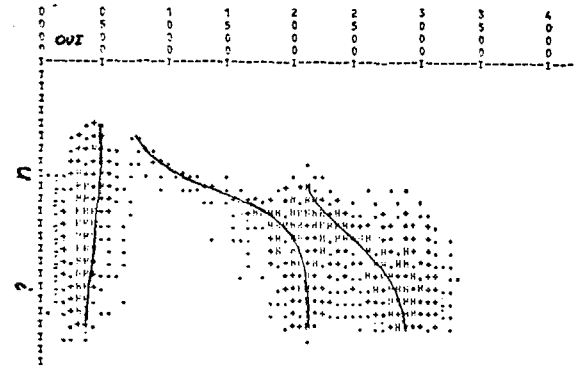
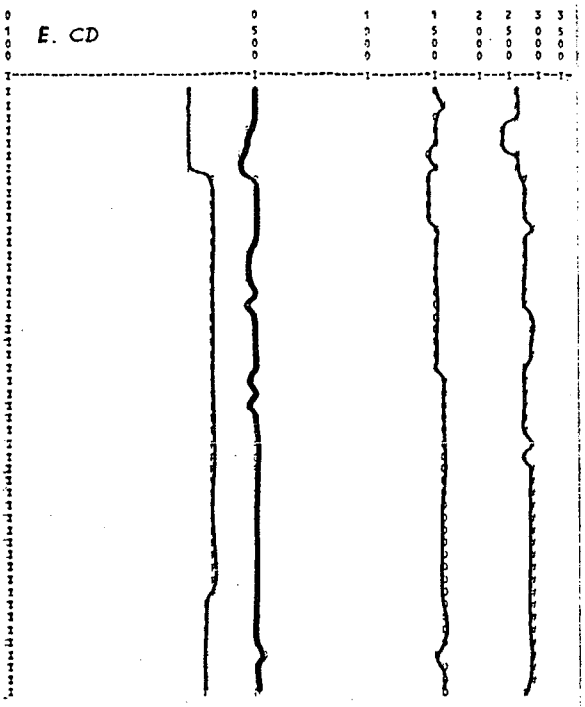


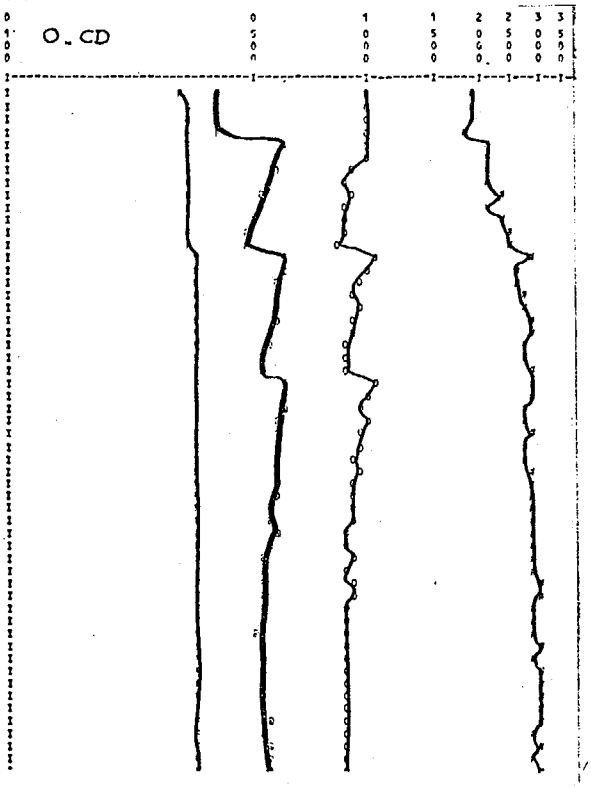
Fig :18 ANALYSE FREQUENTIELLE. LOCUTEUR MASCULIN(D.D). FONDAMENTAL 125 Hz. FENETRE 15 ms.

Un autre intérêt de la méthode temporelle réside dans la possibilité de traiter aussi bien des voix d'hommes que des voix de femmes ou d'enfants. Les fig. 19 et 20 représentent l'analyse temporelle d'une voix d'enfant dont le fondamental est de 350 Hz. La fig. 21 représente les mêmes sons analysés par F.F.T. Dans cette analyse, il est très difficile de séparer les formants. Une des raisons essentielles en est la modulation du spectre par le fondamental.



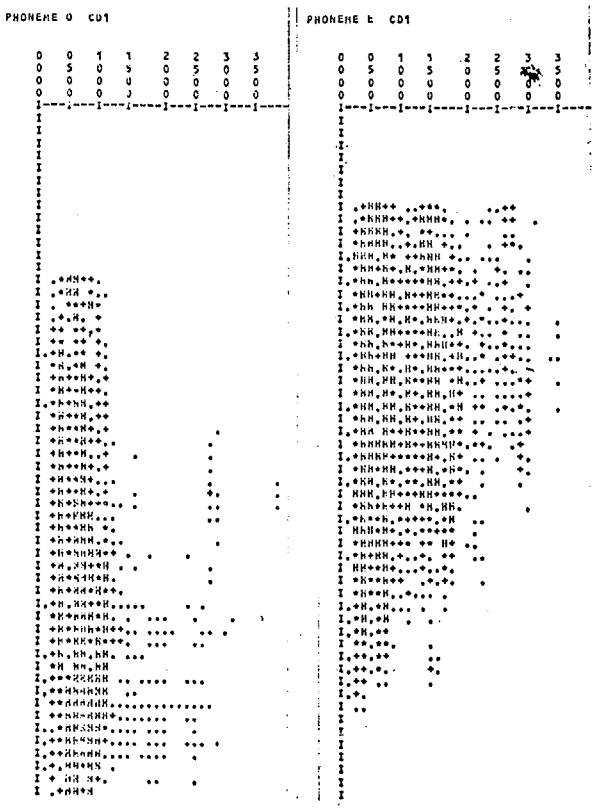
ANALYSE TEMPORELLE D'UNE VOIX D'ENFANT. FONDAMENTAL 350 Hz. PHONEME ϕ. ECHELLE LOGARITHMIQUE

Fig : 19



ANALYSE TEMPORELLE D'UNE VOIX D'ENFANT. FONDAMENTAL 350 Hz. PHONEME O. ECHELLE LOGARITHMIQ

Fig : 20



ANALYSE SPECTRALE (FENETRE 8 ms) D'UNE VOIX D'ENFANT. FONDAMENTAL 350 Hz. PHONEME O ET ϕ. ECHELLE LINEAIRE.

Fig : 21

CONCLUSION

Le système d'analyse et de reconnaissance envisagé dans le projet A.R.I.A. a été conçu dans l'optique d'un système hiérarchisé de reconnaissance de phrases construites à partir d'un vocabulaire non limité et prononcées de façon naturelle par un ensemble de locuteurs quelconques.

Dans une optique aussi générale, il semble que la reconnaissance analytique soit la seule envisageable. Elle impose une bonne segmentation et une analyse très fine mais elle ne nécessite que la mise en mémoire d'un nombre d'éléments phonétiques limités, et de ce fait, n'introduit que peu de limitation sur le vocabulaire utilisé.

ANALYSE D'UN SIGNAL FORTEMENT STRUCTURE :
LE SIGNAL VOCAL

BIBLIOGRAPHIE

(1) 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'U.P.S. Toulouse 1974

(2) FANT G.
Acoustic Theory of speech Production
Mouton the Hague. Paris 1970

(3) ATAL B.S. HANAUER S.L.
Speech Analysis and synthesis by linear prediction of the speech Wave.
JASA vol. 50 pp. 637. 655. 1971

(4) ATAL B.S. SCHROEDER
Adaptative predictive Coding of speech signal
Bell Syst. Tech J.49 pp 1973-1986-1970

(5) GUEGUEN C.J. CARAYANNIS G.
Analyse de la parole par filtrage optimal de Kalman
4ème Journées d'étude sur la parole
Bruxelles 1973

(6) MARKEL J.D.
Digital inverse filtering, a new tool for formant trajectory estimation
IEEE Trans-Audio-Electroacoust.
vol. AU20 pp. 129-137 June 1972

(7) ITAKURA F. SAITO S.
An analysis synthesis thelephony based on maximum likewood method.
Proc. Int. Congr. Acoust. C-5.5. Tokyo 1968

(8) ITAKURA F. SAITO S.
A statistical method for estimation of speech spectral density and formant frequencies.
Electronics and Communications in Japan
Vol. 53 A n° 1 - 1970

(9)DOURS D. FACCA R. PERENNOU G.
Analyse temporelle du signal vocal, comparée à l'analyse fréquentielle classique du point de vue de la reconnaissance.
5ème Journées d'études sur la parole.
Orsay mai 1974

(10) PINSON E.N.
Pitch synchronous Time-domain estimation of formant frequencies and bandwiths.
JASA vol 53 pp. 1265-1273 1963

(11) ALINAT P.
Essai de reconnaissance des phonèmes au moyen d'une cochlée artificielle.
3ème colloque sur le traitement du signal et ses applications.
Nice 1971