SEPTIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 28 MAI au 2 JUIN 1979

COMPARAISON DU FILTRAGE COCHLEAIRE ET DES METH ODES CLASSIQUES POUR L'ANALYSE DE LA PAROLE CONTINUE

CAELEN J. - EL JAI M.C

LABORATOIRE C.E.R.F.I.A. UNIVERSITE P. SABATIER TOULOUSE

RESUME

Dans l'étude de signaux complexes tels que ceux de la parole, une bonne analyse acoustique est nécessaire si l'on veut obtenir des résultats satisfaisants en reconnaissance. Le filtrage cochléaire, partie d'un modèle d'oreille que nous proposons, permet, grâce à ses filtres couplés non-linéaires, de lever les principaux inconvénients du vocoder à canaux. Il offre des avantages sur le codage prédictif linéaire et notamment :

- pas d'hypothèses sur la forme du signal
- pas de fenêtre temporelle
- échelle fréquentielle adaptée à la parole donc précision du 1er formant améliorée
- adaptation continue des coefficients
- minimisation du bruit de filtrage
- minimisation des distorsions harmoniques sur la structure formantique
- facilité de détection des formants (crêtes)
- pas de procédure spéciale pour le suivi des formants
- image spectrale toujours très nette, même dans les transitions

Ces nombreux avantages font de cette méthode, un outil très intéressant en analyse de signaux complexes et non stationnaires tels que la parole.

SUMMARY

In the study of complex signals as speech, a good acoustical analysis is needed to get good results in speech recognition. The cochlear filtering we purpose, from a model of ear, eliminates, owing to its coupled and non-linear filters, the most disavantages of channel vocoders. In comparison with linear prediction coding, il proposes advantages as :

- no hypothesis upon signal
- frequency scale adapted to speech : best precision of the first formant
- continued adjustement of coefficients
- minimation of filtering noise
- minimation of distortions of harmonic structure in formantic structure
- easy technic in detection of formants (tops)
- no special procedure to track formants
- spectral picture being clear even in the transitions

These many advantages make this method very interesting for the analysis of complex and non-stationnary signals as speech.

COMPARAISON DU FILTRAGE COCHLEAIRE ET DES METHODES CLASSIQUES POUR L'ANALYSE DE LA PAROLE CONTINUE

1. INTRODUCTION :

En reconnaissance automatique de la parole on ne peut négliger l'analyse acoustique qui doit être aussi bonne que possible pour préparer ensuite l'analyse phonétique. L'expérimentation de systèmes complets /9/ montre que cette analyse phonétique, lorsqu'elle est insuffisante, ne peut être compensée par les niveaux d'analyse supérieurs : syntaxique, sémantique etc... On estime la performance minimale de cette analyse à 70% environ, ce qui correspond aux performances humaines. Le problème est donc posé : que faut-il comme analyse acoustique à l'entrée du système de reconnaissance ?

Bien évidemment, cette analyse acoustique dépendra de la structure et des options choisies pour le système complet. Si l'on adopte une analyse phonétique fondée sur les indices et les traits acoustiques des phonéticiens il faut avoir accés à un certain nombre de grandeurs physiques : fréquences et amplitudes des formants transitions, mélodie, intensité, durées etc... et l'on s'aperçoit que la précision temporelle est aussi importante que la précision fréquentielle notamment dans les phases transitoires du signal. Or le signal vocal est très instable, tantôt voisé, tantôt sourd, faible, fort, de structure simple ou complexe. Il faut donc une analyse adaptée.

Traditionnellement on utilise le banc de filtres (Vocoder) ou le codage prédictif linéaire. Dans la recherche de méthodes plus performantes, nous avons étudié et modélisé les fonctions essentielles de l'audition périphérique. Dans la suite des opérations effectuées sur le signal par l'oreille, il en est une particuliérement intéressante : le filtrage de la membrane basilaire (M.B).

Bien qu'il soit difficile de comparer des méthodes isolées du système qui les utilise nous tentons ci-après de les évaluer les unes par rapport aux autres sur le plan de la description acoustique de la parole seul.

2. LE VOCODER A CANAUX :

Il se compose essentiellement d'un banc de filtres (de l'ordre de la dizaine) et d'un détecteur de "pitch". Les filtres couvrent la bande utile de la parole, de 2oo Hz à 1o kHz au maximum et selon les modèles.

3. LE CODAGE PREDICTIF LINEAIRE :

La fonction de transfert du conduit vocal peut être approchée par un modèle linéaire tout pôle, soit :

$$H(z) = \frac{1}{1 + \sum_{i=1}^{M} a_i z^{-i}}$$

et dans le domaine temporel :

$$s_n + \sum_{i=1}^{M} a_i s_{n-i} = e_n$$

où e_n est l'entrée échantillonnée et s_n la sortie, (a_i) $i = 1, 2 \ldots M$ les coefficients de prédiction du filtre qu'il s'agit de déterminer.

posons $\hat{s}_n$ une valeur approchée de s_n avec :

$$\hat{s}_n = -\sum_{i=1}^{M} a_i s_{n-i}$$

donc :

$$e_n = s_n - \hat{s}_n$$

e_n peut alors être interprétée comme l'erreur de prédiction entre le signal s_n et le signal $\hat{s}_n$. Les coefficients (a_i) sont choisis de telle sorte que e_n soit minimum, par exemple au sens des moindres carrés.

si $E = \sum_n e_n^2$ alors l'erreur E est minimum lorsque :

$$\frac{\partial E}{\partial a_k} = 0 \qquad 1 \leqslant k \leqslant M$$

ce qui conduit au système de M équations à M inconnues :

$$\sum_{i=1}^{M} a_i \sum_n s_{n-i} s_{n-k} = -\sum_n s_n s_{n-k} \qquad 1 \leqslant k \leqslant M$$

par la méthode d'autocorrélation il vient :

$$\sum_{i=1}^{M} a_i R_{k-i} = -R_k \qquad 1 \leqslant k \leqslant M$$

avec $R_k = \sum_{n=-\infty}^{+\infty} s_n s_{n-k}$ (notons que $R_{-k} = R_k$)

En pratique le signal est multiplié par une fonction fenêtre w_n, $0 \leqslant n \leqslant N-1$ par exemple la fenêtre de Hamming. Pour revenir au domaine fréquentiel il faut ensuite utiliser la transformée discrète de Fourier. Notons que ceci n'est pas indispensable, mais pour comparer les spectres obtenus avec les autres méthodes c'est nécessaire.

Pour une fréquence d'échantillonnage de 15kHz une fenêtre de 128 points et 14 coefficients de prédiction, la méthode nécessite environ 4000 multiplications ou divisions, la formule générale étant :

$$M.N + 2.N.Log_2N + N - M.(M-1)/2$$

4. LE FILTRAGE COCHLEAIRE :

La fig 1 donne un schéma général de l'audition et nous ne nous intéresserons ici qu'à l'oreille interne, plus précisemment à la membrane basilaire (M.B). La M.B vibre sous l'action d'une onde de compression se propageant dans la rampe vestibulaire. Cette onde est induite dans la périlymphe par le système tympano-ossiculaire, lui-même mis en mouvement par l'onde sonore.

Ce sont les propriétés particulières de la M.B qui permettent une première analyse des sons. La M.B joue le rôle de filtres couplés, moyennement sélectifs dont on peut formaliser le comportement par l'équation :

$$\frac{\partial^2 y}{\partial t^2} - c_B^2(x,y)\frac{\partial^2 y}{\partial x^2} + \frac{\omega_c}{Q_c}(x,y)\frac{\partial y}{\partial t} + \omega_c^2(x,y)y = g(x)\frac{\partial s}{\partial t}(t)$$

$$y(o,t) = y(L,t) = y(x,o) = \frac{\partial y}{\partial t}(x,o) = 0$$

avec :

$y(x,t)$	déplacement de la M.B au point x
$c_B(x,y)$	célérité de l'onde de déformation
$Q_c(x,y)$	coefficient de surtension du filtre équivalent au point x
$f_c(x) = \omega_c/2\pi$	fréquence centrale de ce filtre
L	longueur de la M.B (35 mm)
s(t)	signal de sortie de l'oreille moyenne
g(x)	déperdition de l'onde périlymphatique

L'oreille moyenne peut-être simulée par un filtre linéaire, la sortie s(t) est donc celle d'un filtre peu sélectif de fréquence centrale 1500 Hz dont l'entrée est le signal sonore.

La résolution numérique de cette équation avec un schéma aux différences finies implicites sur une grille rectangulaire, conduit à l'algorithme ci-après :

Sur la grille $G = (x_i, t_j)$ $x_i = ih$, $i=1,2,...M$, $t_j = t_o + jT$, $j=0,1...$ on calcule :

POUR i=1 A M FAIRE;
$$F_i = (g_i C_i (s^j - s^{j-1}) + A_i y_i^{j-1} + B_i y_i^{j-2} + F_{i-1})E_i$$
FIN;

POUR i=M à 1 FAIRE;
$$y_i^j = E_i y_{i+1}^j + F_i$$
FIN;

Avec les fonctions auxiliaires :

$$A_i = 2\exp(-\frac{\omega_i}{2Q_i}T)\cos(\omega_i T)C_i$$

$$B_i = -\exp(-\frac{\omega_i}{Q_i}T)C_i$$

$$C_i = (h/c_{Bi}T)^2$$

$$E_i = 1/(2 + C_i + E_{i-1})$$

et les conditions initiales et aux limites :

$$E_o = F_o = y_{M+1}^j = y_o^j = 0$$

Les coefficients sont choisis conformément aux données physiologiques, fig 2. La bande couverte par le modèle est limitée volontairement à 100 Hz - 6500 Hz. La détection du fondamental se fait en comptant les passages par zéro du signal sortant d'un filtre variable (modèle de neurone non décrit ici), filtre placé en aval du filtre basse fréquence du banc précédemment décrit.

Il est facile de voir que le nombre des multiplications peut se réduire à 4M pour chaque point d'échantillonnage. Avec les mêmes fenêtres que pour le codage prédictif et M=24 nous obtenons environ 12000 multiplications pour un bloc de 128 points. Si l'on considère la masse de calculs à effectuer comme parallélisable, chaque filtre ne nécessitant que 4 multiplications, on est ramené à 512 opérations par bloc.

5. COMPARAISON DES METHODES :

5.1 Le vocoder et le filtrage cochléaire :

Le vocoder et le filtrage cochléaire ne se différencient que parce que les filtres de la M.B sont couplés et non indépendants comme dans le vocoder. Ce couplage permet de réduire considérablement le bruit de filtrage (aucun filtre ne peut vibrer isolément) et atténue la structure harmonique au profit de la structure formantique (la réponse impulsionnelle de chaque filtre n'est plus sinusoïdale). Il s'ensuit que les transitions sont beaucoup plus nettes et que la détection des formants se fait par une simple détection des crêtes, puisqu'aucun bruit de filtrage ne vient entacher la mesure. Mais ce gain de performance se fait au prix de calculs plus lourds,irréalisables par voie analogique. Seule la nouvelle génération de microprocesseurs rapides pourra permettre d'atteindre le temps réel.

Nombre d'utilisateurs ont constaté l'insuffisance des vocoders dans l'analyse de la parole (nombre de canaux notamment, sensibilté aux harmoniques etc...), nous pensons que l'alourdissement des calculs justifie le gain de performance surtout en analyse de la parole.

5.2 Le codage prédictif et le filtrage cochléaire :

Nous insisterons davantage sur cette comparaison.

5.2.1 Inconvénients propres au codage prédictif :

En revenant sur la méthode on peut voir que:

- la modélisation du conduit vocal est linéaire et ne tient pas compte notamment du couplage source-conduit.
- la source est négligée, et même minimisée il s'ensuit un déplacement des formants en fréquence, énergie et bande passante.
- le modèle tout pôle est une approximation parfois néfaste, cas des zéros nasals.
- la fenêtre de pondération perturbe le spectre
- l'échelle linéaire en fréquence obtenue après la DFT n'est pas optimale pour la parole. Le 1 er formant y perd en précision.

5.2.2 Avantages propres au codage prédictif:

- richesse du spectre dans les hautes fréquences qui permet de détecter les formants supérieurs.
- réduction importante de l'information directement utilisable en synthèse

5.2.3 Inconvénients propres au filtrage cochléaire :

En revenant sur la méthode on peut voir ici aussi que :

- volume de calculs plus important mais parallélisable
- impossibilité d'atteindre un optimum autrement que par voie empirique, problème du choix des coefficients.

5.2.4 Avantages propres au filtrage cochléaire :

- répartition adaptée des filtres
- non-linéarités qui permettent d'adapter les coefficients aux variations du signal
- obtention directe du spectre (sans DFT)
- facilité de détection des formants par simple localisation des crêtes.
- bruit de filtrage minimisé
- spectre non distordu ni par les harmoniques ni par les éventuelles hypothèses sur le signal d'entrée, ni par l'utilisation d'une fenêtre de pondération.

6. COMPARAISON DES RESULTATS :

Pour illustrer les affirmations précédentes nous donnons ci-après des analyses de sons effectuées en codage prédictif et par filtrage couplé.

Sur les sonagrammes nous notons une plus grande netteté des transitions pour le filtrage cochléaire. La richesse apparente du sonagramme en codage prédictif est souvent nuisible pour détecter les formants. Des améliorations pour le codage prédictif peuvent être tentées

- nombre de coefficients variables
- intégration sur des cercles plus grands ou plus petits que l'unité
- intégration sur des contours plus sophistiqués.

Malgré de tels essais, la qualité du codage prédictif en analyse de la parole nous semble moins bonne que le filtrage cochléaire.

Nous laissons au lecteur de commenter lui-même les diverses figures.

Sur les coupes spectrales, en parties stables du signal (voyelles) la richesse en hautes fréquences est plus grande pour le codage prédictif. Mais pour qui sait qu'essentiellement les deux premiers formants suffisent en reconnaissance (et perception on est en droit de se demander si une telle richesse est utile. Par contre le premier formant est beaucoup plus imprécis. En parties instables (consonne la sélectivité temporelle n'est pas suffisante et les approximations trop grossières pour une analyse pertinente. Ainsi dans le burts de /k/ fig 6 le caractère "compact" est détruit par le codage prédictif mais respecté par le filtrage cochléaire ([illegible] tration de l'énergie au centre du spectre).

Nous pourrions multiplier les exemples, mais le cadre de cet article nous l'interdit.

7. CONCLUSION :

Les avantages du filtrage cochléaire sur les deux autres méthodes sont indéniables et seront

COMPARAISON DU FILTRAGE COCHLEAIRE ET DES METHODES CLASSIQUES POUR L'ANALYSE DE LA PAROLE CONTINUE

encore plus nets dès qu'il fournira des résultats en temps réel. A ce moment son utilisation en analyse de signaux complexes et non stationnaires comme la parole sera très compétitive vis-à-vis d'autres méthodes. D'ores et déjà, si l'on se contente du temps différé, il offre de grandes possibilité en analyse spectrale et de là en reconnaissance de signaux.

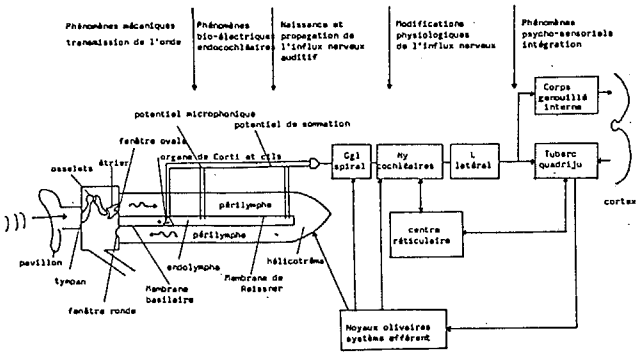

fig 1 : Schéma simplifié de l'oreille

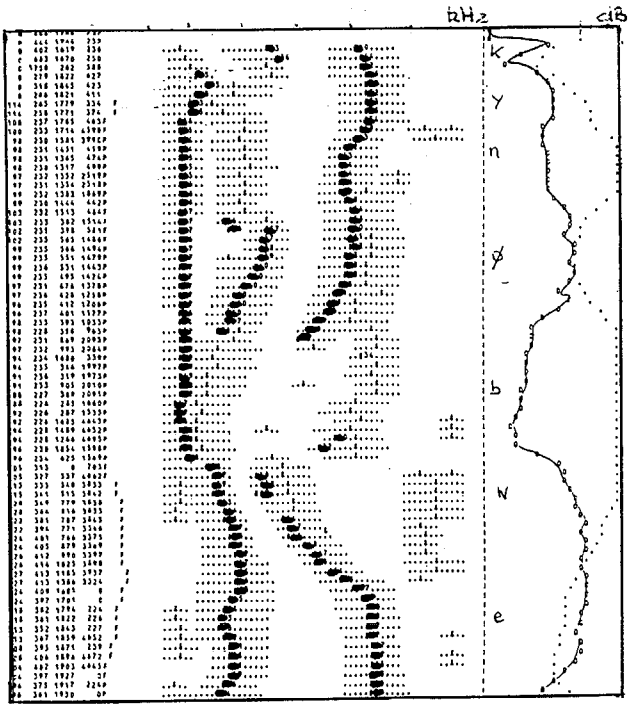

fig 3 : Représentation sonagramme de la phrase "(ave)c une bouée" obtenue par filtrage cochléaire. Les transitions de formants restent nettes et le bruit de filtrage faible.

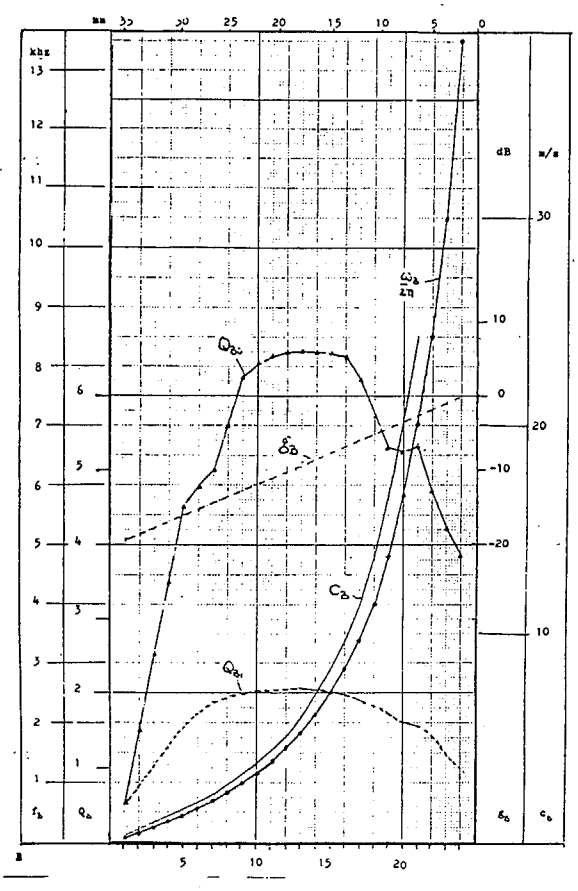

fig 2 : Valeur des coefficients de l'équation de la M.B. en fonction de la distance à l'étrier.
Q_{Bo} valeur du coefficient de surtension au seuil (0 dB)
Q_{Bl} valeur du coefficient de surtension limite (130 dB)

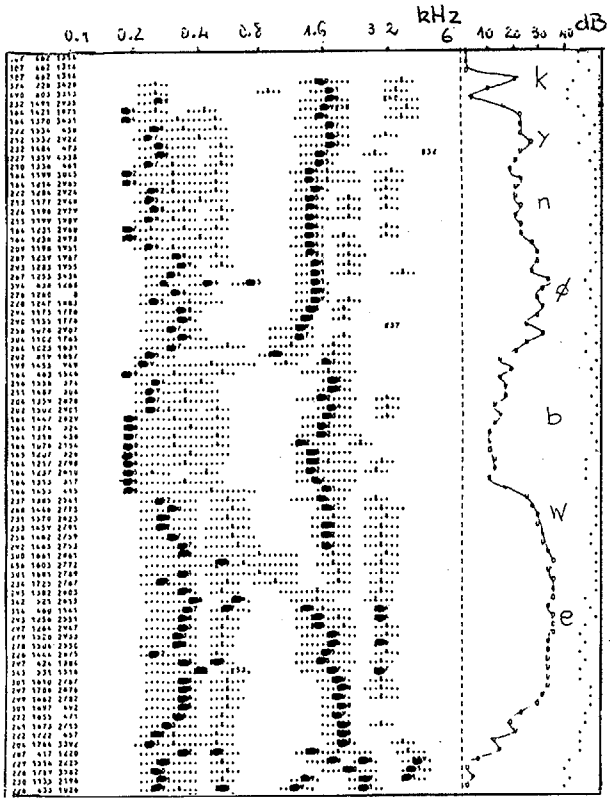

fig 4 : Même représentation que la figure précédente, obtenue par codage prédictif. Les transitions de formants sont irrégulières, la précision du 1er formant insuffisante.

COMPARAISON DU FILTRAGE COCHLEAIRE ET DES METHODES CLASSIQUES

POUR L'ANALYSE DE LA PAROLE CONTINUE

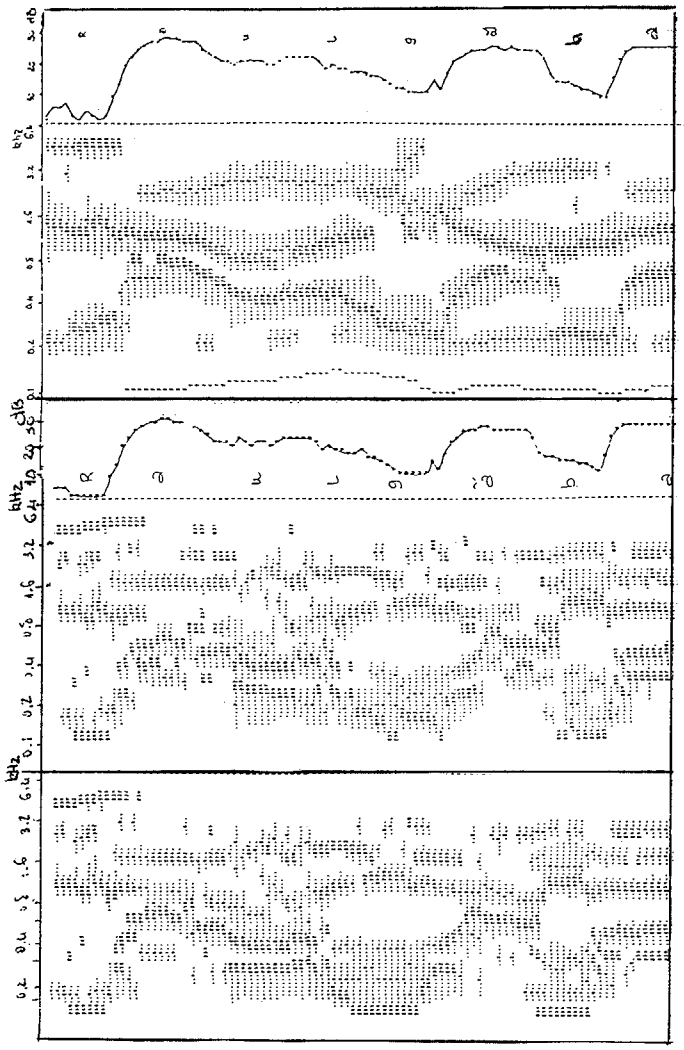

fig 5 : Analyse de la phrase "Raoul gambade" par filtrage cochléaire en haut et codage prédictif sur le cercle de rayon 1 au milieu et sur le cercle de rayon 1.03 en bas. Malgré une légère amélioration sur le cercle de rayon 1.03, le codage prédictif donne des résultats beaucoup plus incertains qu'en filtrage couplé.

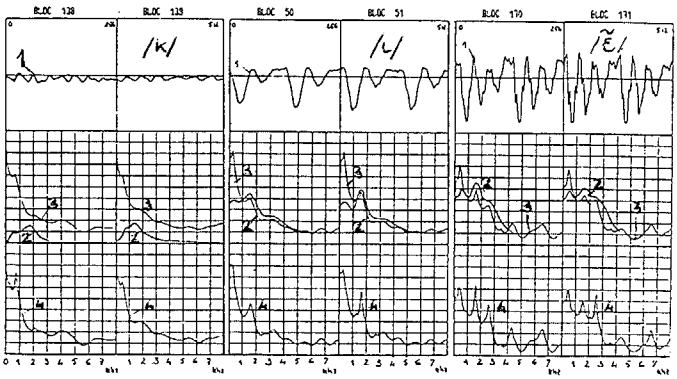

fig 6 : Coupes spectrales de segments stables prélevés dans des sons de parole naturelle.
1. signal, 2. filtrage cochléaire, 3. codage prédictif rayon 1, 4. codage prédictif rayon .95
/k/ reste compact par l'analyse 2.

8. BIBLIOGRAPHIE :

/1/ ATAL B.S, HANAUER S.L 1971
Speech analysis and synthesis by linear prediction of the speech wave J.A.S.A vol 50 n° 2 637-655

/2/ CAELEN J. 1974
Un modèle mathématique de cochlée et son application à l'analyse du signal vocal Thèse D.I Toulouse

/3/ CAELEN J. 1977
Etude de la fonction de filtrage de l'oreille à partir d'un modèle mathématique Rev Acoustique Vol 10 n° 42 226-234

/4/ FANT G. 1967
Sound, features and perception 6° Int Congress of Phon Sc. Prague

/5/ MAKHOUL J. 1975
Linear prediction : a tutorial review Proc I.E.E.E vol 63 n°4

/6/ MAKHOUL J. 1975
Spectral linear prediction : properties and applications I.E.E.E Trans ASSP- 23

/7/ MARKEL J.D., GRAY A.H. 1976
Linear prediction of speech Springer-Verlag Berlin

/8/ NILSSON H.G., MOLLER A.R. 1977
Linear and nonlinear models of the basilar membrane motion. Biol Cyb 27 Springer-Verlag Berlin

/9/ REDDY R. 1976
Speech recognition by machine : a review Proc I.E.E.E 64 501-530

/10/RHODE W.S, ROBLES L. 1974
Linear and nonlinear models of the basilar membrane motion. Biol Cyb 27 Springer Verlag. Berlin

/11/SCHROEDER M.R 1975
Models of hearing. Proc I.E.E.E 63 n°9