TROISIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

Nice 1er au 5 juin 1971

ESSAI DE RECONNAISSANCE DES PHONEMES AU MOYEN D'UNE COCHLEE ARTIFICIELLE*

Pierre ALINAT

THOMSON - C.S.F. Division ASM 06 Cagnes-sur-Mer.

RESUME

On présente une réalisation de cochlée artificielle (organe sensible de l'oreille interne). Les résultats fournis par cette cochlée ont permis de définir un système de reconnaissance des phonèmes. Ce système a été réalisé sommairement et expérimenté. Les voyelles et les consonnes fricatives (voix masculines) sont convenablement reconnues. Les consonnes explosives nécessitent une modification du système de reconnaissance.

SUMMARY

From the results obtained with an artificial cochlea, a phonem recognition system is defined. This system has been devised in a simplified version and tested. Vowels and fricative consonnants (male voice) are suitably recognized. The system needs to be modified for explosive consonnants recognition.

*Etude partiellement financée par la Direction des Recherches et Moyens d'Essais - Paris - FRANCE.

ESSAI DE RECONNAISSANCE DES PHONEMES AU MOYEN D'UNE COCHLEE ARTIFICIELLE

Pierre ALINAT

1.- POSITION DU PROBLEME

Le but de l'étude qui va être décrite est de reconnaître les phonèmes, c'est-à-dire les unités élémentaires d'information vocale. Par exemple le mot "poule" est composé de 3 phonèmes P - OU - L. Il suffit d'en changer un seul pour avoir un mot différent : par exemple "PAUL" composé de P - Ô - L. Dans ce qui suit, on se limite aux phonèmes de la langue française mais les résultats peuvent être étendus aux autres langues.

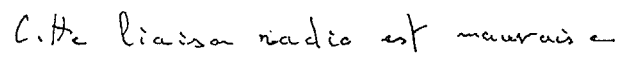
Il n'y a pas de correspondance bi-univoque entre les phonèmes et les signes de l'alphabet. Une liste des phonèmes de la langue française est donnée en annexe 1.

Il est certain que dans une conversation courante entre individus, de nombreux phonèmes ne sont pas réellement prononcés : grâce à cela le débit de parole peut être plus rapide et la fatigue moins grande. Etant donné la grande redondance au niveau mots d'une part, et au niveau phrases et idées d'autre part, l'information n'est pas perturbée. Par contre une reconnaissance se situant au niveau des phonèmes le sera bien évidemment. Toutefois, dans le futur, pour des dispositifs capables de comprendre une conversation, il sera nécessaire de reconnaître tous les phonèmes réellement prononcés, de façon à pouvoir en fonction du vocabulaire et des règles mis en mémoire, reconstituer les mots réellement prononcés.

En attendant, nous sommes obligés d'imposer au locuteur, une prononciation lente et bien articulée, afin que chaque phonème soit effectivement présent.

ESSAI DE RECONNAISSANCE DES PHONEMES
AU MOYEN D'UNE COCHLEE ARTIFICIELLE

On peut, à ce sujet, faire un parallèle avec l'écriture manuscrite dont la reconnaissance pose un problème absolument semblable. Soient les phrases manuscrites ci-dessous :

1 

2 

Les 2 phrases sont également comprises par un être humain (ayant l'habitude de lire et qui sait de quoi il s'agit bien entendu) mais seule la phrase 2 peut être lue par un moyen de reconnaissance des lettres utilisé seul car la phrase 1 comporte de nombreuses lettres mal formées. Toutefois, lorsqu'un être humain lit la phrase 1, il commence par reconnaître les lettres les mieux formées. Il en est de même en parole. La reconnaissance des phonèmes est une étape nécessaire.

2.- PRINCIPES UTILISES

2.1. Le signal parole

Les ondes sonores émises par les organes vocaux d'un individu peuvent être traduites en une tension électrique f(t) au moyen d'un microphone. Le signal f(t) résulte du filtrage d'un signal-excitation e(t) (cordes vocales en friction) par le filtre dépendant du temps constitué par les cavités laryngo-bucco-nasales.

Le signal f(t) peut être considéré comme stationnaire de temps à autre sur des durées de 100 à 200 ms. Mais en général, ces durées doivent être ramenées à 10 ou 20 ms.

C'est un tel signal f(t) que l'on cherche à décomposer en une suite de phonèmes.

2.2. Utilité du prétraitement en reconnaissance des formes

Notre but étant la reconnaissance des phonèmes, nous sommes conduits, comme toujours dans ce genre de problème , à faire subir au signal f(t) un prétraitement avant de résoudre le problème de classification.

Le prétraitement en reconnaissance de formes consiste en un changement de la base au moyen de laquelle on définit le signal. Il n'a échappé à personne que ce prétraitement était très important car c'est de lui que dépend la facilité des opérations de classification.

2.3. Prétraitement adapté au signal parole

Le signal f(t) nous est par définition donné par rapport à une base temps. Si ce signal était périodique, la Transformation de Fourier constituerait le prétraitement adapté. Si ce signal était parfaitement aléatoire sans aucune stationnarité, il vaudrait mieux rester en temps.

Dans notre cas, le signal présente certaines stationnarités et la phase a peu d'importance (tout au moins au point de vue reconnaissance des phonèmes). Un

prétraitement adapté peut être une transformation du type défini par la relation (1) ci-dessous :

$$F'(t, \omega) = \int_{-\infty}^{+\infty} f(\lambda)\, r_{\omega}(t-\lambda)\, d\lambda \qquad (1)$$

suivie d'une détection intégration (pour éliminer la phase) $r_{\omega}(\lambda)$ est la réponse impulsionnelle d'un filtre passe bande Il faut déterminer le $r_{\omega}(\lambda)$ optimum.

Or, il se trouve justement que l'oreille humaine fait subir au signal f(t) une transformation de ce type.

2.4. L'oreille humaine

On distingue en général l'oreille externe et moyenne d'une part, l'oreille interne d'autre part. L'oreille externe est composée du conduit acoustique, l'oreille moyenne du tympan et des osselets. Elle n'a qu'un rôle de transmission du message acoustique et sa fonction de transfert n'offre aucune difficulté de réalisation.

L'oreille interne joue le rôle d'analyseur spectral. On donne en Annexe 2, les principaux résultats sur ce sujet découvert par BEKESY. Ce qui nous intéresse c'est que l'oreille interne effectue précisément une transformation du type de celle décrite par la relation (1).

La fonction de transfert du filtre élémentaire peut être approchée par :

$$R_{\omega}(p) = \frac{p}{p + 0{,}9\omega} \left[\frac{\omega^2}{p^2 + 2 \times 0{,}3\, \omega p + \omega^2} \right]^2 \qquad (2)$$

Cela détermine une fonction $r_\omega(\lambda)$ adaptée à ce que l'on désire reconnaître (plus exactement c'est le signal f(t) qui a dû s'adapter à l'oreille tout en tenant compte des impératifs dus à l'instrument d'émission).

On va imiter au mieux l'oreille interne au moyen d'une batterie de filtres dont les fréquences centrales ω seront disposées conformément à la figure 3 de l'Annexe 2, et dont les fonctions de transfert seront du type $R_\omega(p)$.

3. - REALISATIONS ET RESULTATS

Les idées ci-avant ont été vérifiées au moyen d'une maquette sommaire dont on va décrire la réalisation. La figure 1 schématise le système tel qu'il a été construit.

3.1. Prétraitement

Le signal f(t) convenablement filtré (100 Hz - 5 000 Hz), subit une emphase pour relever le niveau des hautes fréquences. Un CAG permet de s'affranchir des variations de niveau.

Une portion de cochlée est approchée par une batterie de 60 filtres dont les fréquences centrales sont réparties entre 200 et 5 000 Hz, et les fonctions de transfert conformes à la formule (2). Chaque filtre est suivi d'un détecteur-intégrateur.

Toutes les 4 ms, les 60 sorties sont

ESSAI DE RECONNAISSANCE DES PHONEMES AU MOYEN D'UNE COCHLEE ARTIFICIELLE

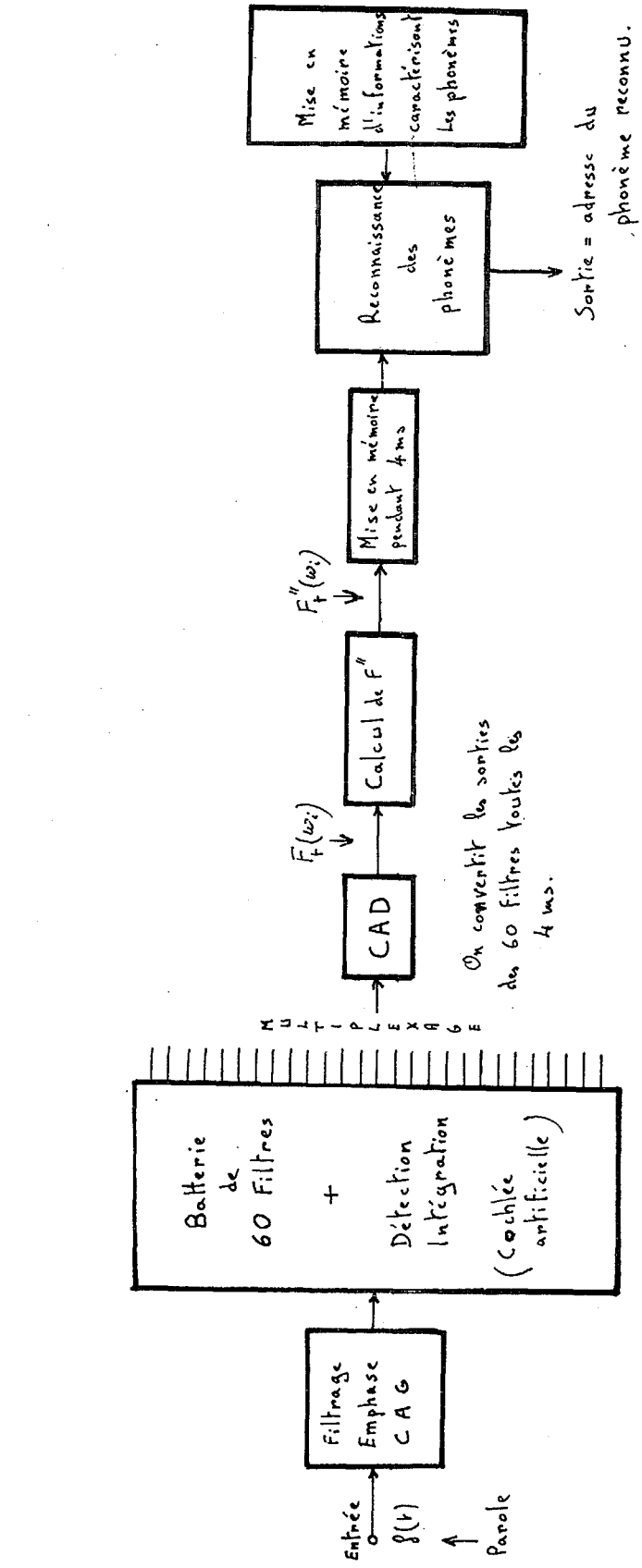


Figure 1

converties numériquement.

On obtient alors $F_t(\omega_i)$ avec $t = t_o + Kx4$ ms

$i = 1$ à 60.

$F_t(\omega_i)$ représente la déformation de la cochlée à l'instant t.

Il faut maintenant déterminer la position des formants de $F_t(\omega_i)$, c'est-à-dire approximativement les principaux pôles de la fonction de transfert du filtre constitué par la cavité bucco-nasale. Une méthode efficace consiste à calculer la dérivée seconde $F''_t(\omega_i)$ de $F_t(\omega_i)$, et à en chercher les maxima. Il se trouve que $F_t(\omega_i)$ est une courbe connue sous forme échantillonnée. De plus, étant donné que la batterie de filtres n'est pas parfaite, $F_t(\omega_i)$ est entachée d'un bruit. A cause de cela on commence par lisser la courbe en ajoutant des échantillons successifs. Pour diminuer encore l'influence des erreurs, on effectue 2 calculs de F" à partir de 2 courbes lissées différemment, et on multiplie les résultats.

Les équations ci-dessous expriment les opérations réalisées :

on pose $y_n = F_t(\omega_n)$

et on calcule :

$$F''_1 = -(y_{n-3} + 2y_{n-2}) + (y_{n-1} + 4y_n + y_{n+1}) - (2y_{n+2} + y_{n+3})$$

$$F''_2 = -(y_{n-4}+y_{n-3}+y_{n-2}) + 2(y_{n-1}+y_n+y_{n+1}) - (y_{n+2}+y_{n+3}+y_{n+4})$$

ESSAI DE RECONNAISSANCE DES PHONEMES AU MOYEN D'UNE COCHLEE ARTIFICIELLE

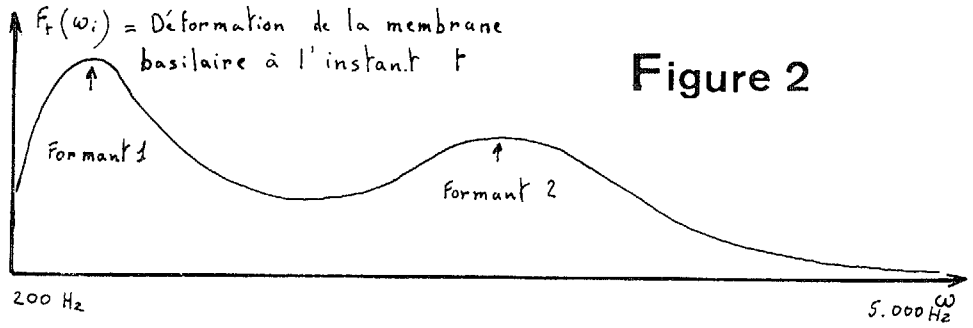
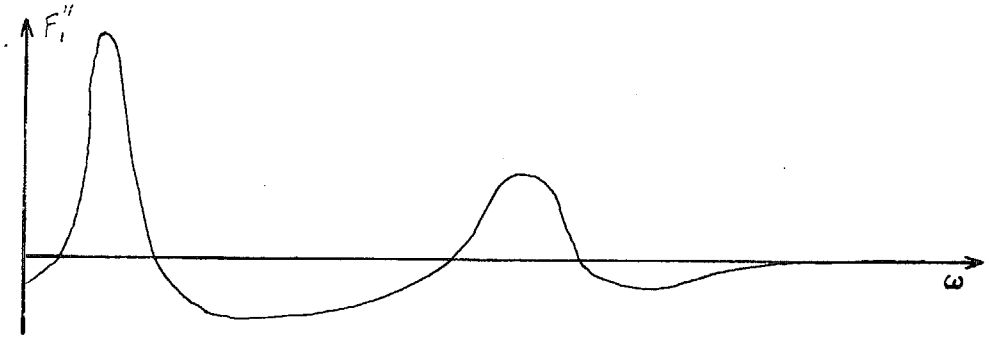
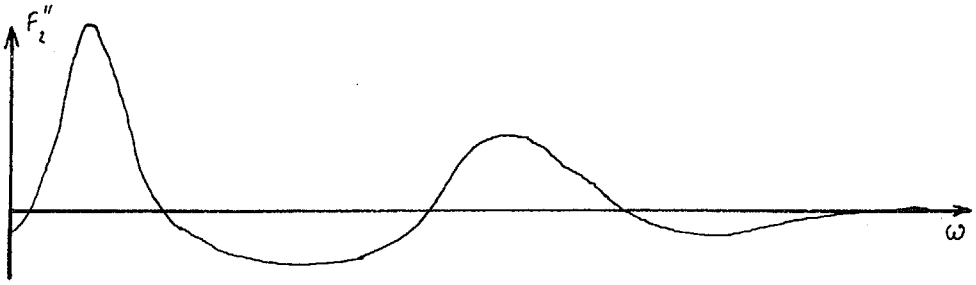
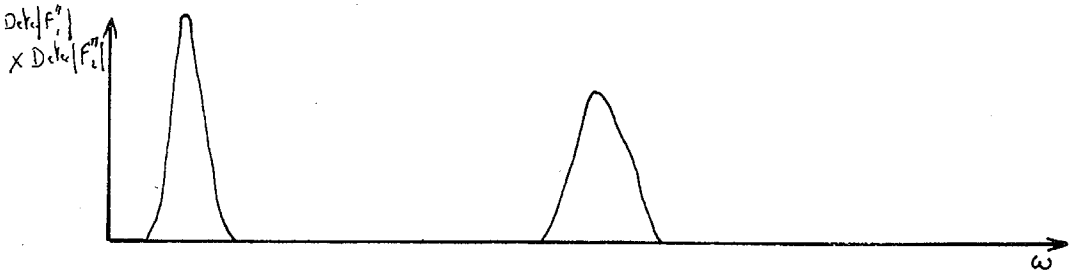
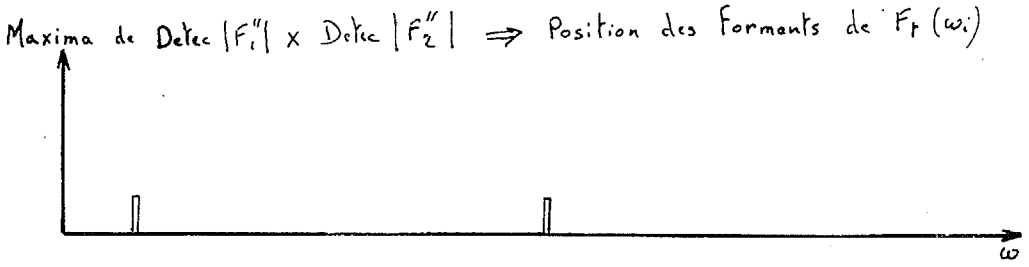


Figure 2









Ces deux opérations correspondent à des filtrages linéaires non récursifs. On pourrait très bien filtrer la courbe différemment en faisant intervenir plus d'échantillons et obtenir ainsi d'autres fonctions F''_n.

On détecte les courbes $F''_i \longrightarrow$ Detect (F''_i) et on réalise le produit $\prod_i$ Detect (F''_i).

On détermine les maxima de la courbe obtenue : leur position peut être considérée comme étant celle des formants de $f_t(\omega_i)$. La figure 2 illustre le processus ci-dessus. Comme nous allons le voir, l'information caractérisant les phonèmes soutenus (voyelles et consonnes fricatives) est liée à la position de ces formants

3.2. Résultats obtenus à partir de $F''(\omega)$

Comme prévu, le prétraitement décrit au paragraphe 3.1. a énormément facilité les opérations de reconnaissance.

Les observations portant sur une vingtaine de locuteurs de sexe masculin ont permis de dégager des règles pour la reconnaissance des voyelles et des consonnes fricatives.

En effet, en étudiant la répartition des positions des formants attachés à ces phonèmes prononcés à l'intérieur de mots divers par des locuteurs variés arti-

culant distinctement, on s'est aperçu qu'il existait des plages de fréquences dans lesquelles ces formants se produisaient presque obligatoirement. La figure 3 représente ces plages au moyen d'une échelle graduée en numéro de filtre. Il semblerait que pour les voix féminines, les plages soient seulement déplacées de 2 ou 3 filtres vers le haut. Nous allons étudier ce que cela représente pour chaque grande classe de phonèmes.

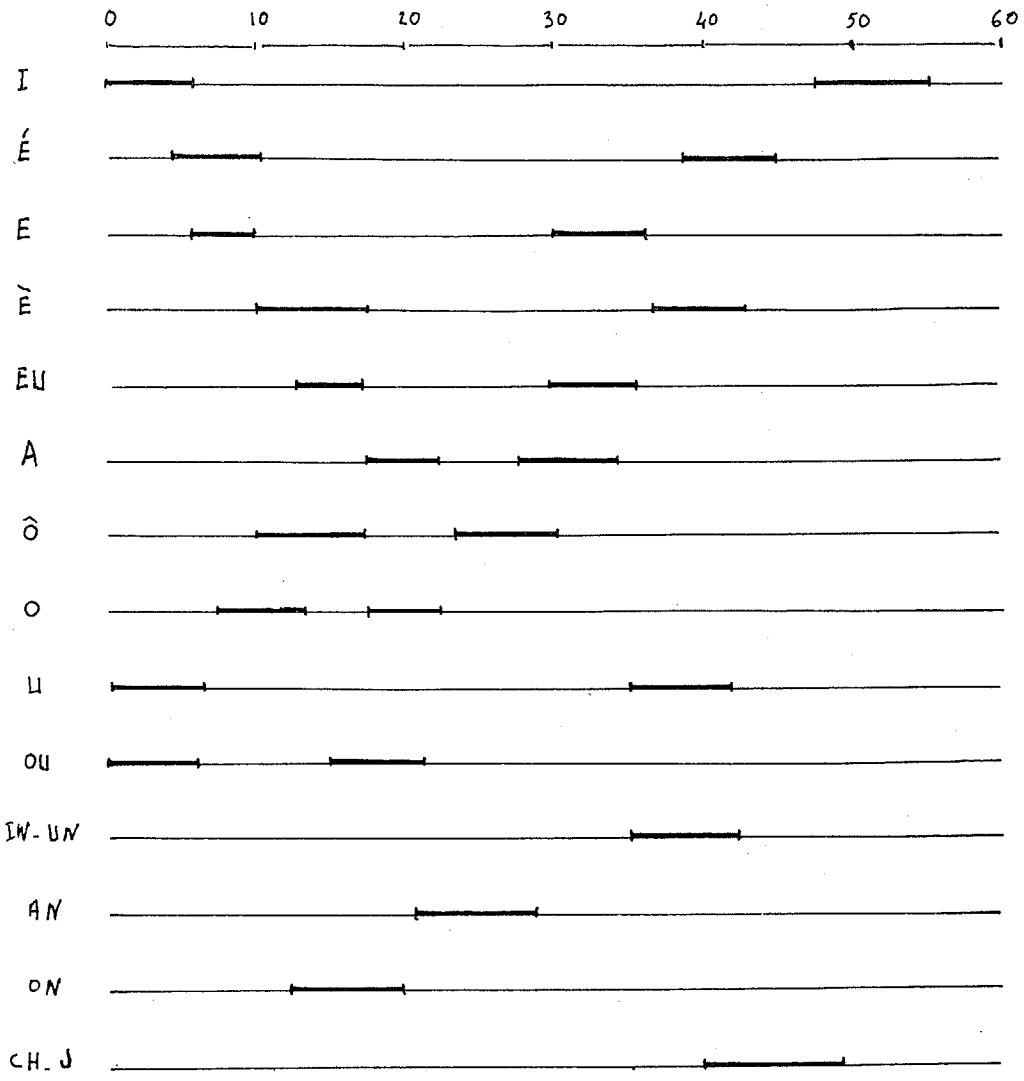


Figure 3

3.2.1. Voyelles simples [I] [É] [E] [È] [EU] [A] [Ô] [O] [U] [OU]

On voit sur la figure 3 que pour une voyelle bien articulée , la position des deux premiers formants, détermine sans ambiguïté la nature de la voyelle. Il est remarquable que les plages sont assez régulièrement espacées, et que leur longueur est grossièrement constante ce qui permet de penser que la répartition des fréquences centrales des filtres est bonne.

Les amplitudes respectives des formants (observées sur $F''(\omega)$) doivent seulement satisfaire l'inégalité :

$$\text{Ampl (form. 1)} > K \text{ Ampl (form. 2)} \qquad (3)$$

Le facteur K qui est proche de 1 n'a pas été exactement déterminé.

Il peut arriver que les harmoniques 2 ou 3 du signal d'excitation, créent dans la zone du premier formant, un formant parasite mais l'amplitude du "parasite" est toujours inférieure à celle du premier formant qui peut de ce fait, être distingué.

3.2.2. Voyelles nasales [IN, UN] [AN] [ON]

Les voyelles nasales ont aussi en général, deux formants mais l'inégalité (3) est inversée : c'est le critère supplémentaire qui sert à les caractériser.

3.2.3. Consonnes fricatives [CH, J] [S, Z] [F, V]

Le cas des consonnes fricatives est très semblable à celui des voyelles. Il n'y a plus toutefois qu'un seul maximum à localiser au lieu de 2.

La nature de l'excitation (voisée ou non) permet de distinguer les sourdes [CH, S, F] des sonores [J, Z et V]. La position du formant pour [S, Z] et [F, V] est située au-delà de la borne supérieure de la cochlée réalisée. Toutefois, quelques mesures ont été faites en relisant une bande magnétique à la moitié et au quart de la vitesse d'inscription.

3.2.4. Consonnes explosives [P, B] [T, D] [K, G] [M] [N]

latérale [L]

roulante [R]

Pour ces consonnes dont les critères de reconnaissance font intervenir le temps, il n'est pas possible actuellement de fournir des résultats. Il est probable qu'il faudra faire subir aux fonctions $F_t(\omega_i)$ ou $F''_t(\omega_i)$ une dérivation d'ordre impaire par rapport au temps en employant des méthodes analogues à celles du paragraphe 3.1.

Il vaudrait mieux toutefois, dans ce cas là, réaliser les filtrages par des méthodes récursives au lieu de non récursives.

3.3. Reconnaissance en temps réel des voyelles simples et de (CH, J)

On a construit un système de reconnaissance simple permettant de vérifier en temps réel sur un grand nombre d'individus, la validité des critères mis en évidence au paragraphe 3.2.

Ce système ne tient aucun compte des amplitudes relatives des formants : il ne peut donc distinguer les nasales.

La figure 4 indique le principe utilisé pour la reconnaissance. On inscrit dans une première ligne à mémoire magnétostrictive pour chaque phonème que l'on désire reconnaître, les plages dans lesquelles doit se produire au moins un formant.

On inscrit de même dans une seconde ligne les plages dans lesquelles ne doivent pas se produire de formant.

ESSAI DE RECONNAISSANCE DES PHONEMES
AU MOYEN D'UNE COCHLEE ARTIFICIELLE

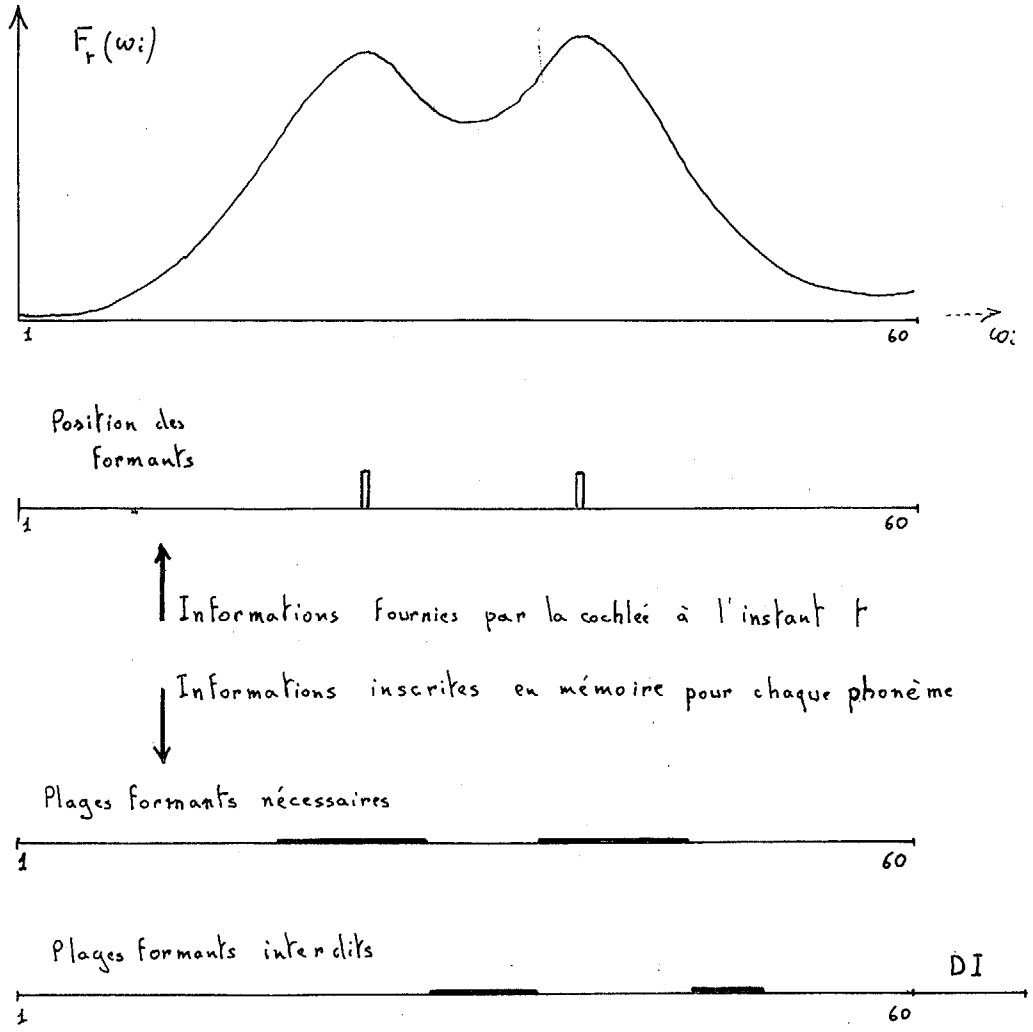


Figure 4

Toutes les 4 ms, les positions des formants effectivement présents à la sortie de la batterie de filtres, sont comparées avec les plages inscrites par l'opérateur. Si elles satisfont aux conditions nécessaires et suffisantes correspondant à un phonème particulier et ce, pendant un temps supérieur à une durée imposée (DI) inscrite également dans la seconde ligne par l'opérateur, on

décide que ce phonème a été prononcé.

Ce système donne des résultats corrects pour toutes les voyelles non nasales et les fricatives [CH, J]. Seuls sont gênants les parasites dûs au bruit sur $F_t(\omega_i)$ et les phonèmes non reconnus (consonnes explosives et voyelles nasales). Les autres fricatives pourraient être reconnues en prolongeant la batterie de filtres vers les hautes fréquences pour pouvoir détecter le formant de [S, Z] (5 500 Hz environ) et de [F, V] (à 7 000 Hz environ). les photographies de l'Annexe 3 donnent quelques exemples de reconnaissances réalisées.

4. - CONCLUSION

D'après les résultats obtenus, on peut penser que le principe utilisé dans cette étude est valable. Il faudrait toutefois étendre la méthode aux consonnes explosives et aux voix féminines. Il faudrait également étudier les possibilités d'amélioration de la cochlée artificielle en utilisant des techniques digitales ou acoustiques.

La reconnaissance des phonèmes soutenus par le système des plages, est simple et efficace lorsque le locuteur articule convenablement. Toutefois lorsqu'on fera intervenir pour la reconnaissance de paroles mal articulées le vocabulaire, la grammaire et les idées, il deviendra peut-être souhaitable de faire intervenir des probabilités de présence de phonèmes.

A partir de système de ce type, il sera possible de transmettre des messages oraux vers des machines. Il serait en effet commode que le moyen de communication que

les hommes emploient le plus souvent entre eux, puisse être utilisé également pour commander les différentes machines dont ils disposent. La seule condition serait une prononciation lente et bien articulée.

ESSAI DE RECONNAISSANCE DE PHONEMES
AU MOYEN D'UNE COCHLEE ARTIFICIELLE

A N N E X E 1

Liste des phonèmes de la langue française

Les principaux phonèmes de la langue française sont donnés ci-dessous. Ils sont répartis en différentes classes et un exemple est donné pour chacun. Le signe correspondant de l'alphabet phonétique international est donné lorsqu'il y a lieu.

1. Voyelles simples

[I]	lit
[E]	tasser
[E]	relire
[E]	mère
[EU]	beurre
[Ô]	Bordeaux
[O]	Bordeaux
[U]	du
[OU]	doux

2. Voyelles nasales

[IN]	pain
[UN]	un
[AN]	plan

ESSAI DE RECONNAISSANCE DES PHONEMES
AU MOYEN D'UNE COCHLEE ARTIFICIELLE

[ON] blond

3. Consonnes fricatives

Sourdes

[CH]	chat	[ʃ]
[S]	son	[S]
[F]	feu	[F]

Sonores

[J]	Jean	[ʒ]
[Z]	zéro	[Z]
[V]	Var	[V]

4. Liquides

[L] long

[R] rond

5. Explosives

Sourdes

[P]	pas	[p]
[T]	ta	[t]
[K]	ccq	[K]

Sonores

[B]	bon	[b]
[D]	don	[d]
[G]	gare	[g]

Nasales

[M]	mon	[m]
[N]	non	[n]

6. Diphtongues

Il s'agit du son obtenu en passant de façon continue d'une voyelle à une autre voyelle : exemple oeil ([EU] ⟶ [I]). Cet ensemble de deux voyelles est parfois considéré comme un phonème unique.

ESSAI DE RECONNAISSANCE DES PHONEMES
AU MOYEN D'UNE COCHLEE ARTIFICIELLE

A N N E X E 2

L'oreille humaine

La partie de l'oreille qui nous intéresse est la cochlée schématisée au point de vue mécanique par le dessin ci-après :

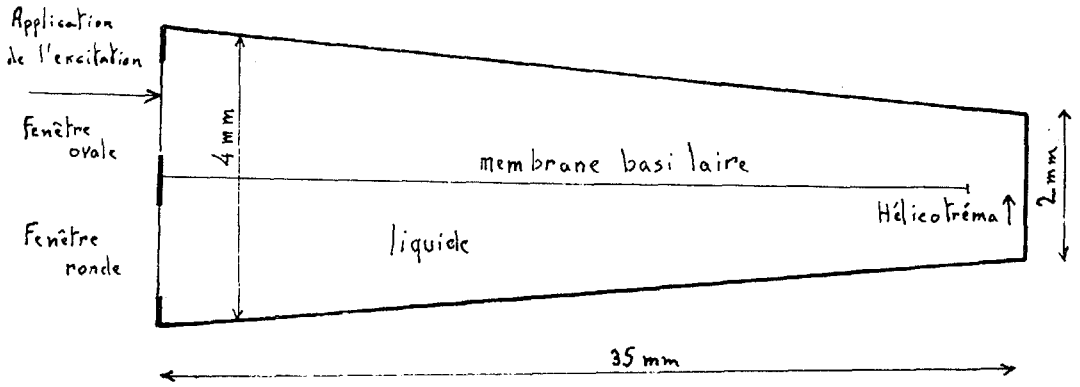


Figure 1

Les terminaisons nerveuses sont disposées tout au long de la membrane basilaire. BEKESY [1] a montré, qu'excitée par une onde sinusoïdale de fréquence F_0 la membrane basilaire était parcourue d'ondes progressives selon le schéma de la figure ci-dessous.

ESSAI DE RECONNAISSANCE DES PHONEMES
AU MOYEN D'UNE COCHLEE ARTIFICIELLE

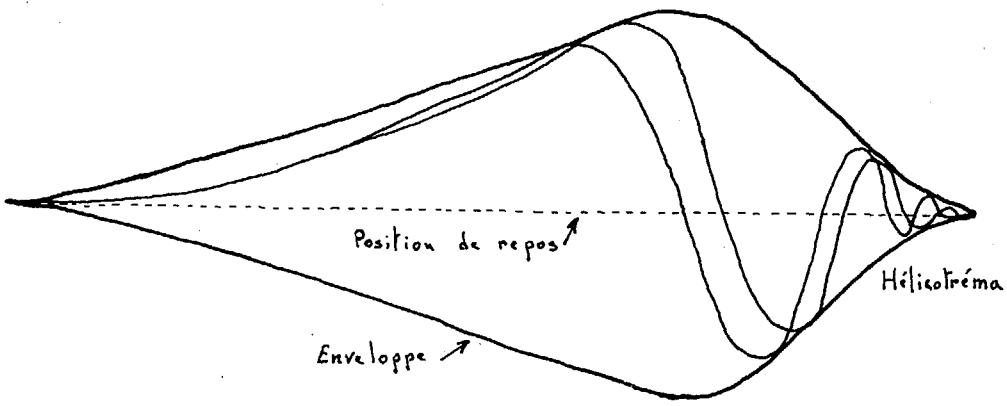


Figure 2

La position du maximum de l'enveloppe ne dépend que de la fréquence et on peut ainsi graduer en fréquence, la membrane basilaire.

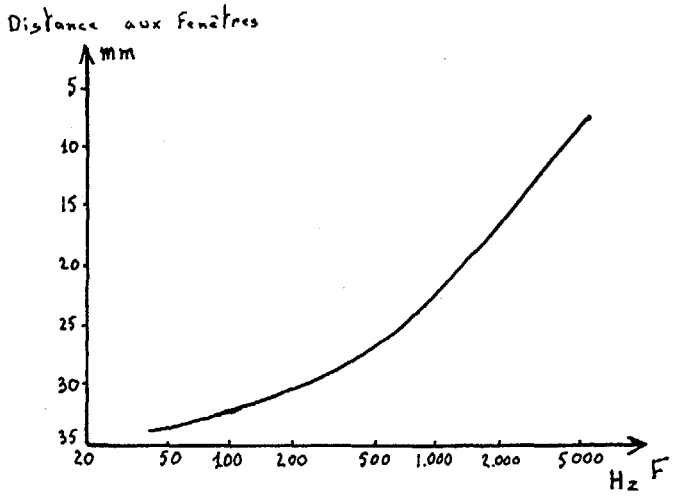


Figure 3

Lorsque la cochlée est excitée par un signal f(t) l'amplitude des déplacements de l'enveloppe au point

correspondant à la fréquence $\frac{\omega}{2\pi}$ est égale (à un facteur multiplicatif près) à l'amplitude du signal sortant d'un filtre de fonction de transfert approchée par la formule ci-dessous : référence [6]

$$R_\omega(p) = \frac{p}{p+\omega}\left[\frac{\omega}{(p+\frac{\omega}{2})^2+\omega^2}\right]^2\left(\frac{2\,000\,\pi\omega}{\omega+2\,000\,\pi}\right)e^{-\frac{3}{4}\frac{\pi}{\omega}p}$$

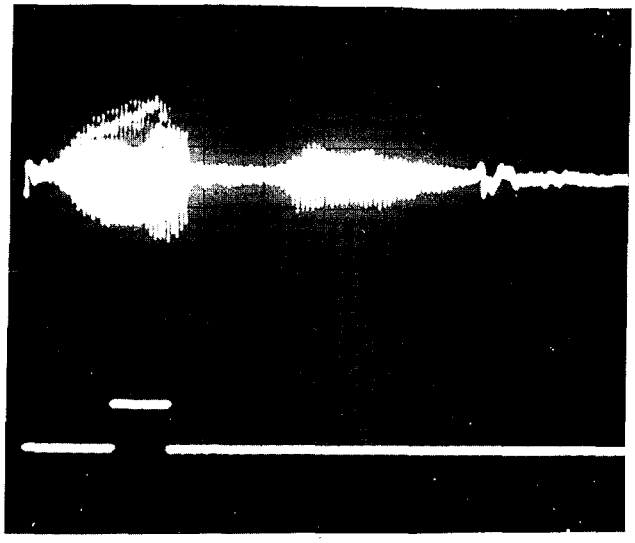
On voit qu'il s'agit d'un filtre passe-bande peu surtendu. Des mesures plus récentes utilisant des niveaux d'excitation plus faibles permettent de penser que les surtensions sont plus élevées.

ESSAI DE RECONNAISSANCE DES PHONEMES AU MOYEN D'UNE COCHLEE ARTIFICIELLE

A N N E X E 3

Résultats expérimentaux

PRO DUIT



Signal à l'entrée

Reconnaissance de O

Temps

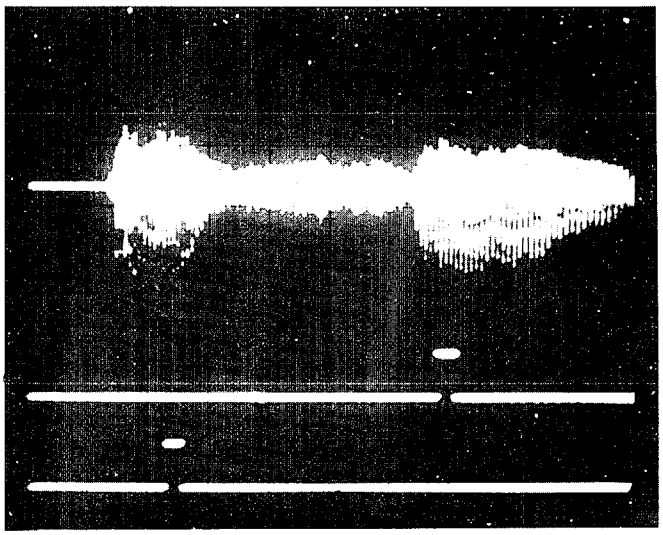
100 ms/carreau.

ESSAI DE RECONNAISSANCE DES PHONEMES
AU MOYEN D'UNE COCHLEE ARTIFICIELLE

A N N E X E 3

(Suite)

TA SS ER

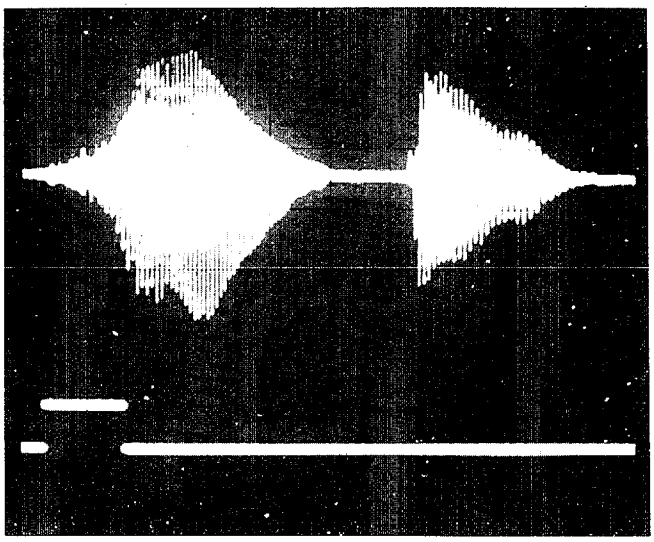


Entrée

Reconnaissance de É

Reconnaissance de A

J E T ER



Entrée

Reconnaissance de J

ESSAI DE RECONNAISSANCE DES PHONEMES
AU MOYEN D'UNE COCHLEE ARTIFICIELLE

B I B L I O G R A P H I E

[1] VON BEKESY

"Experiment in hearing" Mc Graw Hill Book.

[2] FANT G.

"Acoustic Theory of Speech production"
Mouton and Co. - 1960.

[3] S.J. CAMPANELLA et D. PHYFE

"Application of a Model of the Analog Ear to Speech signal Analysis"
IEEE Trans. Audio and Electro Acoustics - Vol. AU 16 n° 1 - Mars 1968.

[4] T.B. MARTIN et J.J. TALAVAGE

"Application of Neural Logic ot Speech Analysis and Recognition"
Bionics Symposium 1963 - Vol. 2.

[5] LIBERMAN - INGERMANN - LISKER - DELATTRE - COOPER

"Minimal Rules for Synthetizing Speech"
JASA - Vol. 31 n° 11 - Novembre 1959.

[6] J.L. FLANAGAN

"Speech Analysis, Synthesis and Perception"
New York : Academic Press Inc.
