COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL
ET SES APPLICATIONS

NICE du 26 au 30 AVRIL 1977

METHODES STRUCTURELLES EN RECONNAISSANCE DE LA PAROLE AU NIVEAU ACOUSTIQUE INSPIREES PAR UN MODELE D'OREILLE.

J. CAELEN - M.C. EL JAÏ - G. PERENNOU

CERFIA -- UNIVERSITE PAUL SABATIER - 118 route de Narbonne - 31077 TOULOUSE CEDEX

RESUME

Le développement des ordinateurs et de l'analyse numérique ont permis de proposer des modèles mathématiques raffinés d'oreilles au cours de ces cinq dernières années.

Dans le cadre de cet article nous nous intéressons surtout à l'application de ces modèles à la reconnaissance de la parole au niveau acoustique.

L'intérêt de l'organisation de l'oreille avec ses trois niveaux de filtrage est montré. Puis on donne, dans les grandes lignes, l'organisation d'un système de reconnaissance de la parole continue au niveau acoustique. Des problèmes délicats, tels que la distinction entre consonnes de même catégorie, sont abordés.

SUMMARY

During the last five years the development of computers and numerical analysis have allowed one to propose mathematical sophisticated models of ear.

In this paper we consider especially the application of these models to speech recognition at the acoustic level.

We underline the interest presented by the ear structure with its three filtering levels. Then, we give an outline of the organization of a full speech recognition system at the acoustical level.

Peculiar problems such as separating between consons of the same class are treated.

METHODES STRUCTURELLES EN RECONNAISSANCE DE LA PAROLE AU NIVEAU ACOUSTIQUE INSPIREES PAR UN MODELE D'OREILLE.

INTRODUCTION

L'étude d'un modèle numérique d'oreille ([7].. ...[9]) suggère une approche de la reconnaissance de la parole, au niveau acoustique. On comprend mieux ainsi les performances de l'oreille, loin d'être égalées dans les systèmes de reconnaissance de parole actuellement connus.

Au plan pratique, des résultats de reconnaissance intéressants sont obtenus. Une difficulté subsiste : le temps de calcul. Mais la méthode utilisée, inspirée du comportement humain, permet de ne mettre en œuvre les calculs lourds qu'en cas de nécessité. Comme pour le codage prédictif, un ordinateur spécialisé devrait fournir un dispositif temps réel, permettant de faire progresser considérablement la reconnaissance de la parole.

1. GENERALITES : Le signal vocal - Un modèle d'oreille

1.1. La parole

La parole est le résultat de l'excitation du conduit vocal par l'air expulsé des poumons. Lorsque les cordes vocales modulent cet écoulement, il apparaît alors comme un train d'impulsions appelé fondamental, (dont la forme fait encore l'objet de contreverses) et le son est dit vocalique. Selon la forme du conduit vocal on obtient diverses catégories de sons ou phonèmes (voir tableau 1).

Il est bon d'observer que certains phonèmes sont vocaliques et donc bien caractérisés par les fréquences propres du conduit vocal pendant leur production. Ces fréquences sont appelées formants. Certains phonèmes résultent de bruits de frictions en un point de resserrement du conduit vocal, d'autres sont discontinus : ce sont les plosives.

Enfin les modes de production des sons peuvent s'associer pour donner des sons composites ; par exemple : fricatives voisées .

Les performances d'un bon analyseur de parole sont donc délicates à atteindre puisque les signaux étudiés sont rarement stationnaires, parfois (pseudo) périodiques, parfois discontinus, ou parfois aléatoires.

1.2. Le modèle d'oreille

Il comprend trois niveaux de filtrage :

a) Un filtre récursif du deuxième ordre qui rend compte du fonctionnement de l'oreille externe et moyenne ; il a pour effet de favoriser la bande 500HZ-5000 HZ ; dans l'oreille ce filtrage s'adapte à l'intensité du signal.

b) Un banc de 24 filtres à larges bandes et couplés en série qui rend compte du fonctionnement de la membrane basilaire placée dans la cochlée (figure 2). Ce filtrage a pour effet de rendre disponible un spectre instantané du signal le long de la membrane.

c) Un banc de filtres à bandes étroites, indépendants et variables. Ils possèdent une boucle de contrôle leur permettant de se positionner par ajustement successif sur un mode fréquentiel du signal d'entrée.

Chacun de ces filtres est en correspondance avec un filtre du banc précédent dont il analyse la sortie. Le deuxième banc de filtres rend compte du fonctionnement des cellules nerveuses auditives - d'une manière encore hypothétique il est vrai -

Pour plus de détails sur ce modèle mathématique d'oreille nous renvoyons à ([7], [9], [10]). Nous nous contenterons de donner les avantages qui démarquent le fonctionnement de ce modèle par rapport aux dispositifs classiques du type vocoder (un vocoder, rappelons-le, est constitué essentiellement d'un banc de filtres fixes et indépendants).

1er point :

Les filtres étant à bandes larges, il est possible de suivre les variations instantanées du signal (ce qui est essentiel pour la détection et la reconnaissance de certains phonèmes : plosives, nasales et, à un degré moindre, liquides et semi-voyelles).

2ème point :

Le couplage des filtres permet d'éliminer les oscillations libres des filtres non accordés et les oscillations harmoniques.

Le résultat est que le spectre obtenu apparaît lissé et ne contient que les éléments essentiels (voir figure 4). Ainsi il est possible de faire de la détection de crêtes et de déterminer les bandes de fréquences prépondérantes du signal.

3ème point :

Le deuxième banc de filtres est approprié à une résolution fréquentielle fine puisque les filtres sont adaptables et à bandes étroites. A chaque fois, qu'une telle résolution fine est demandée, il est possible d'y faire appel.

Il est clair qu'un banc de filtres fixes, ne peut cumuler les avantages des points 1 et 3 et qu'il est bien inférieur sur le point 2.

Ce que nous avons dit du signal vocal nous permet donc de comprendre l'écart constaté entre les

METHODES STRUCTURELLES EN RECONNAISSANCE DE LA PAROLE AU NIVEAU ACOUSTIQUE INSPIREES PAR UN MODELE D'OREILLE.

performances de l'oreille humaine et celles des appareils du type vocoder.

2. ANALYSE DU SIGNAL VOCAL

2.1. Les traits pertinents

Les sons de la parole peuvent se classer selon l'aspect du signal. On utilise pour cela des "traits pertinents" (pour utiliser un vocabulaire de phonéticien), articulatoires dans le premier cas, acoustiques dans le second. Ces derniers, préconisés par R. JACOBSON, M. HALLE et G. FANT ([1],[2],[3]) se présentent comme des propriétés du signal. Si à un instant donné le signal possède (resp. ne possède pas) la propriété, on code +1 (resp. -1). Dans les cas intermédiaires, on peut coder 0.

Les traits acoustiques sont évidemment intéressants du point de vue de la reconnaissance. Notons du reste que les études menées par les psychoacousticiens, notamment LANDERCY et RENARD [5],[6] ont mis en évidence de manière indiscutable les liens entre certains de ces traits (opposition grave-aigué, opposition diffus-compact, sur lesquels nous reviendrons) et la perception des sons.

Dans le modèle d'oreille étudié, on peut aussi élaborer des traits acoustiques, dont certains sont directement inspirés par ceux de la phonétique. Bien sûr il faut que ces traits puissent être calculés au moyen d'algorithmes aussi simples que possible.

Quelques paramètres et traits décrivant la dynamique du signal - que nous appellerons indices d'évolution - seront aussi nécessaires.

2.2. Description des paramètres, indices et traits pertinents fournis par le modèle.

2.2.1. Les paramètres

Désignons par $(e_n)_{n\geq 0}$ la suite représentant le signal échantillonné. On choisit une fenêtre temporelle de longueur N. Notre modèle utilise en fait, un échantillonnage à 15 kHZ et une fenêtre de 256 points. Le niveau du bruit est déterminé dans une fenêtre où il n'y a pas de parole par :

$$b = \sum_{i=k}^{K+N-1} e_i^2 .$$

L'énergie du signal par rapport au bruit s'exprime alors, pour l'intervalle [jN+1, j(N+1)], appelé $j^{ème}$ bloc, par :

$$w_j = 10\ \mathrm{Log}_{10}\left[\frac{1}{b}\sum_{i=jN+1}^{j(N+1)} e_i^2\right]$$

Le spectre du signal, fourni par le premier banc de filtres, est également un paramètre (à valeur dans $\mathbb{R}_+^{24}$) noté $\Omega_j = (\omega_j^1, \omega_j^2, \ldots \omega_j^{24})$, ($\omega_j^k$: sortie de $k^{ème}$ filtre au $j^{ème}$ bloc)

Ce sont les maxima du spectre qui jouent le rôle le plus important. A chacun de ceux-ci sera associé le triplet (ν_j^k, B_j^k, A_j^k) (fréquence - largeur de bande - amplitude du $k^{ième}$ maximum au bloc j). On numérotera $k = 1,2,\ldots,M_j$ ces maxima de la plus petite fréquence (ν_j^1) à la plus grande ($\nu_j^{M_j}$). M_j est donc le nombre de maxima du bloc j.

2.2.2. Les indices d'évolution

a) L'indice de stabilité en énergie

Il caractérise la rapidité de variation d'énergie et se définit par :

$$\delta w_j = \begin{cases} 1 & \text{si } w_{j+1} - w_j > 3\text{dB}, \\ 0 & \text{si } |w_{j+1} - w_j| \leq 3\text{dB}, \\ -1 & \text{si } w_{j+1} - w_j < -3\text{dB}. \end{cases}$$

b) Les indices de stabilité formantique

Ils jouent un rôle important pour démarquer des phonèmes successifs possédant des énergies voisines. Ils sont définis par :

$$\Delta\nu_j = \sum_{i=1}^{M_j} (\inf_k |\nu_j^i - \nu_{j-1}^k|)(\nu_j^i)^{-1}, \quad \delta\nu_j = \begin{cases} +1, & \text{si } \Delta\nu_j > 8 \\ 0, & \text{si } 3 < \Delta\nu_j \leq 8 \\ -1, & \text{sinon} \end{cases}$$

$$\Delta B_j = \left(\sum_{i=1}^{M_i} B_j^i\right) - \left(\sum_{i=1}^{M_{j-1}} B_{j-1}^i\right), \quad \delta B_j = \begin{cases} 1, & \text{si } \Delta B_j > 6 \\ -1, & \text{sinon} \end{cases}$$

2.2.3. Les traits acoustiques

a) Le voisement est fourni par les filtres de plus basse fréquence dans les deux bancs.

Le filtre variable tend à s'accorder sur une fréquence dans la plage 70 Hz - 350 Hz délimité par le filtre fixe. S'il s'accorde, il y a voisement. La fréquence ainsi obtenue est notée F_j° et le trait de voisement V_j prend la valeur 1 - Sinon $V_j = -1$ - (F_j° est la fréquence du fondamental si elle existe ; elle joue un rôle important dans l'étude de l'intonation).

b) La friction Fr_j se définit par :

$$Fr_j = \begin{cases} 1, & \text{si } \nu_j^{M_j} > 4 \text{ kHz et } A_j^{M_j} > 25 \text{ dB}, \\ 0, & \text{si } \nu_j^{M_j} > 4 \text{ kHz et } 10 \leq A_j^{M_j} \leq 25 \text{ dB}, \\ -1, & \text{sinon} \end{cases}$$

Les sons fricatifs sont aléatoires avec une bande spectrale au-dessus de 4000 Hz. Ils sont caractérisés par Fr = 1. Dans le cas de sons composites, la composante fricative fait que l'indice Fr vaut 0 ou 1.

c) L'opposition grave-aigu

Noté GA_j ce trait se calcule si $M_j = 2$. Il vaut alors :

$$GA_j = \begin{cases} +1, & \text{si } A_j^2 > A_j^1 + 3, \\ -1, & \text{si } A_j^1 > A_j^2 + 3, \\ 0, & \text{sinon.} \end{cases}$$

Il joue un rôle important pour divers types de sons et notamment pour caractériser les voyelles d'arrière par rapport aux voyelles d'avant.

d) La compacité

Ce trait, noté C_j, traduit le fait qu'il y a (ou non) une seule voûte d'énergie se définit par :

$$C_j = \begin{cases} 1, & \text{si } \bigcap_{k=1}^{M_j} B_j^k \neq \emptyset \\ -1, & \text{sinon.} \end{cases}$$

Il caractérise certains /a/, /ɔ/ ainsi que les nasales correspondantes /ã/ et /ɔ̃/. Il permet aussi de distinguer entre les plosives.

e) La plosion et l'occlusion

Pour définir ce trait, on introduit le paramètre auxiliaire $u_j = \sum_{i=1}^{\omega} \omega_j^i$ qui cumule l'énergie dans la bande 2,5 kHz - 6 kHz. On pose alors :

$$P_j = \begin{cases} 1, & \text{si } w_j < 5 \text{ dB et } u_j - u_{j-1} > 10 \text{ dB}, \\ -1, & \text{sinon} \end{cases}$$

$$\mathcal{O}_j = \begin{cases} 1, & \text{si } w_j < 2{,}5 \text{ dB}, \\ 0, & \text{si } 2{,}5 \leq w_j \leq 5 \text{ dB}, \\ -1, & \text{si } w_j > 5 \text{ dB} \end{cases}$$

3. RECONNAISSANCE

La reconnaissance d'un phonème exige, le plus souvent, que l'on tienne compte de la manière dont il se déroule dans le temps. C'est particulièrement vrai dans le cas de plosives (non initiales) puisqu'elles sont caractérisées par une implosion et une explosion encadrant une occlusion. C'est également vrai pour des voyelles qui comportent une première phase ascendante en énergie, une phase centrale stable, puis une phase de décroissance en énergie. Ce schéma les distingue des consonnes dont les spectres, par ailleurs, présentent souvent des similitudes avec ceux des voyelles.

Toutes les indications pour aborder le problème se trouvent rassemblées dans l'ensemble des paramètres, indices et traits acoustiques.

Nous distinguerons trois niveaux de reconnaissance. Au premier niveau, nous utilisons uniquement des traits acoustiques et nous segmentons la parole en phonèmes successifs.

Au deuxième niveau, nous introduisons les paramètres du premier banc de filtres et au troisième, des traitements plus lourds en cas de nécessité.

3.1. Premier niveau de reconnaissance : segmentation et préreconnaissance.

Nous distinguerons les catégories phonétiques suivantes : voyelles (V), consonnes vocaliques (CV), consonnes sourdes (CFr), consonnes fricatives voisées (CFrV), plosives sourdes (CP) et plosives voisées (CPV).

1ère étape : Convenons qu'un bloc j est stable si $\delta w_{j+1} = 0$. On constitue les intervalles maximaux en blocs stables.

2ème étape (indépendante de la première):

Chaque bloc stable est classé selon la table suivante :

	voisement V	occlusion O	friction Fr
V ou CV	+1	-1	-1
CFr	-1		+1 ou 0
CFrV	+1	0 ou -1	+1 ou 0
CP ou silence	-1	+1	-1
CPV	+1	+1 ou 0	-1

3ème étape :

Si la catégorie est V ou CV on attribue la catégorie V quand le début est croissant ($\delta w = +1$) et la fin décroissante ($\delta w = -1$).

Dans le cas où plusieurs intervalles fricatifs sont consécutifs, ils sont rassemblés en un seul CFr si l'un des intervalles est CFr et possède au moins 3 blocs. Sinon l'intervalle est classé CFrV.

On procède de même pour les occlusions.

Enfin, si deux intervalles successifs classés V ont une différence notable d'énergie (> 3 dB), la plus faible est rétablie CV liquide ou semi-voyelle. On trouvera une illustration de tout ceci dans la figure 2.

3.2. Deuxième niveau de reconnaissance : le spectre cochléaire et les traits acoustiques.

a) Cas des voyelles

Pour un bloc de la première moitié de la voyelle on classe selon les tableaux 2,3,4. Si la décision est ambiguë et compacte, du premier et du deuxième formant sinon.

S'il n'y a pas d'ambiguïté, la voyelle est définitivement identifiée, sauf en ce qui concerne la nasalisation.

b) Cas des consonnes

Il pose beaucoup de problèmes. Dans les systèmes de reconnaissance de la parole, on n'affine guère les catégories déjà créées au première niveau.

Dans le cadre restreint de cet article, il n'est guère possible de s'étendre sur ce problème. Cependant les tableaux 5,6,7 montrent que des possibilités d'affinage de la décision existent. Sur la figure 2 on a indiqué les phonèmes tels qu'ils sont reconnus ainsi.

Par exemple, la première plosive est reconnue par utilisation du tableau 5 comme étant un /p/ ou un /t/ car ν_2 = 1600 Hz, V=1 et C = -1. La troisième plosive est identifiée à un /k/ car ν_2= 800 Hz V = -1, C = 1.

3.3. Troisième niveau de reconnaissance : affinage du spectre et dynamique des phonèmes.

A ce niveau, on essaie de lever les ambiguïtés restantes.

a) Dans le cas des voyelles non encore reconnues, après affinage du spectre, on obtient le premier, et si nécessaire, le deuxième formant (ceci est fait par le deuxième banc de filtres. Sur la figure 3, ces formants ressortent par application du codage prédictif). En utilisant le tableau 7 (on pourra à ce sujet consulter [8]) on obtient la classe de la voyelle.

b) La nasalité peut affecter les voyelles /ɛ/,/ɔ/, /a/, /œ/ (également /e/ à cause de la confusion possible avec /ɛ/). Détecter ce trait de nasalité est un problème largement ouvert. Cependant on peut conclure qu'une des voyelles précédentes est nasalisée (et donc /ɛ/ → /$\tilde{\varepsilon}$/, /ɔ/ → /$\tilde{ɔ}$/, /a/ → /$\tilde{a}$/, /œ/ → /$\tilde{œ}$/ lorsque :

- la voyelle est longue (> 5 blocs), décroissante en énergie avec élargissement de la bande passante.
- ou lorsque apparaît un formant nasal caractéristique au voisinage de 250 Hz.

c) Les plosives doivent comporter une phase finale explosive (P = +1) sinon l'occlusion est un silence. Deux types d'information doivent intervenir à ce niveau :

- la courte friction qui peut accompagner la plosion (voir tableau 8),
- la durée de la zone instable d'explosion souvent longue pour le /k/.

Une partie des ambiguïtés résiduelles peuvent se lever à ce niveau.

d) Dans le cas d'une consonne située entre, deux voyelles ou entre une voyelle et un silence, on peut lever certaines ambiguïtés en utilisant le test suivant :

si consonne longue (≥ 5 blocs) et frontière (s) discontinue (s) ($\delta\nu$ = +1) alors consonne nasale.
Ainsi sont détectées les deux premières nasales de la figure 2, le tableau 6 permettant ensuite d'achever la décision.

N.B. L'ensemble des traitements qui peuvent s'effectuer à ce niveau est en fait très large. On pourrait par exemple, remarquer que fricatives et plosives voisées sont allongeantes par rapport aux consonnes sourdes correspondantes ; en fait nous considérons que ce troisième niveau de reconnaissance est un domaine de recherche très largement ouvert.

CONCLUSION

La reconnaissance de la parole pose un problème fondamental d'intelligence artificielle dans un contexte linguistique.

Il est maintenant clair que des systèmes monolithiques, si séduisants soient-ils, n'apporteront que des solutions partielles à un problème d'une telle complexité. Au niveau acoustique, nous pouvons comprendre les performances de l'oreille et la souplesse de son organisation. Dès à présent, les modèles mathématiques d'oreille, et particulièrement celui utilisé ici, permettent de présenter des performances d'analyse et de reconnaissance du signal vocal auxquelles ne peuvent prétendre les systèmes simples basés sur un banc de filtres fixes.

Ajoutant à cela l'accord avec les observations de physiologie et de la psychoacoustique, nous ne pouvons que souhaiter l'approfondissement des études dans ce domaine et la réalisation de dispositifs, non plus simulés sur ordinateurs mais réalisés pour le temps réel.

ERRATA :

1. P.1., col.2, 14e ligne. Supprimer (voir figure 3).
2. P.2., col.1, 5e ligne. Lire : les sons de la parole peuvent se classer selon leur mode de production ou selon l'aspect du signal.
3. P.4, col.1, 5e ligne. Lire : si la décision est ambigüe et la voyelle compacte et aigüe on appelle une recherche fine du premier formant, sinon, celle du premier et du deuxième formant.

METHODES STRUCTURELLES EN RECONNAISSANCE DE LA PAROLE AU NIVEAU ACOUSTIQUE INSPIREES PAR UN MODELE D'OREILLE.

REFERENCES

(1) JAKOBSON R. and HALLE M. (1956)
Fundamentals of language Mouton the Hague

(2) DELATTRE P. (1966)
Studies in french and comparative phonetics
Mouton the Hague

(3) FANT G. (1961)
Sound spectrography
Proc. 4 th Int. Congr. Phon. Sci. HELSINKI

(4) TROUBETZKOY N.S. (1964)
Principes de Phonologie
Klincksieck Paris

(5) LANDERCY A. - RENARD R. (1974)
Perception des voyelles françaises filtrées
Revue de Phonétique Appliquée Mons

(6) LANDERCY A. - RENARD R. (1975)
Champ fréquentiel et reconnaissance de voyelles françaises
Revue de Phonétique Appliquée - Mons

(7) CAELEN J. (1974)
Un modèle mathématique de cochlée
Application à l'analyse de la parole.
Thèse Docteur Ingénieur - Toulouse

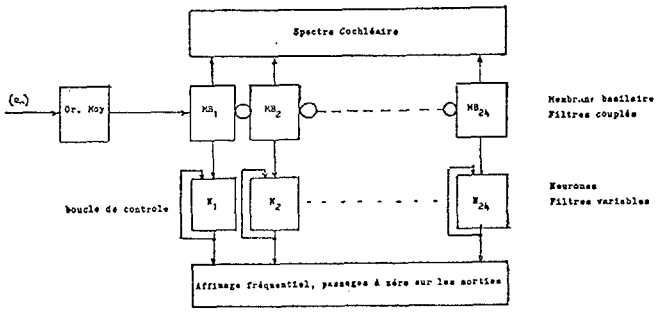
(8) DOURS D.- FACCA R. - PERENNOU G. (1976)
Analyse temporelle du signal vocal comparée à l'analyse fréquentielle du point de vue de la reconnaissance
Rapport SESORI

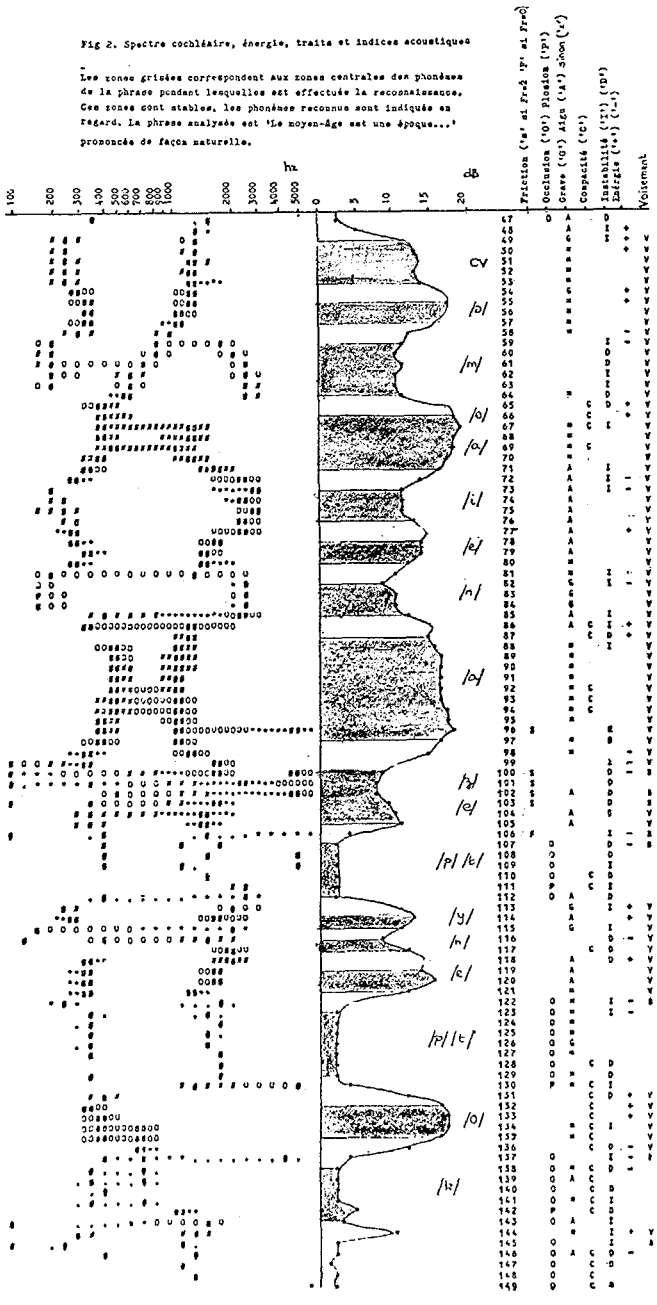
(9) CAELEN J. (1977)
Etude de la fonction de filtrage de l'oreille par un modèle mathématique
Revue d'Acoustique.

(10) CAELEN J. (1976)
Etude des filtres du 2e ordre variables et couplés - Application à l'analyse de signaux non stationnaires.
Note CERFIA TOULOUSE

(11) DOURS A. - FACCA R. - PERENNOU G.
Analyse d'un signal fortement structuré : le signal vocal.
Colloque GRETSI - Nice Juin 1975

Fig 1 Schéma général du modèle d'oreille





METHODES STRUCTURELLES EN RECONNAISSANCE DE LA PAROLE AU NIVEAU ACOUSTIQUE INSPIREES PAR UN MODELE D'OREILLE.

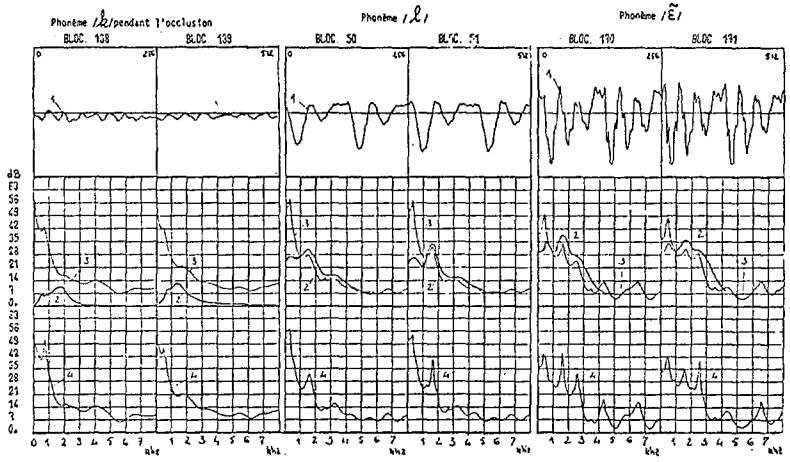


FIGURE 3 - 1 : signal parole ; 2 : spectre cochléaire ; 3 et 4 : méthode de Markel sur les cercles unité et de rayon 1.05

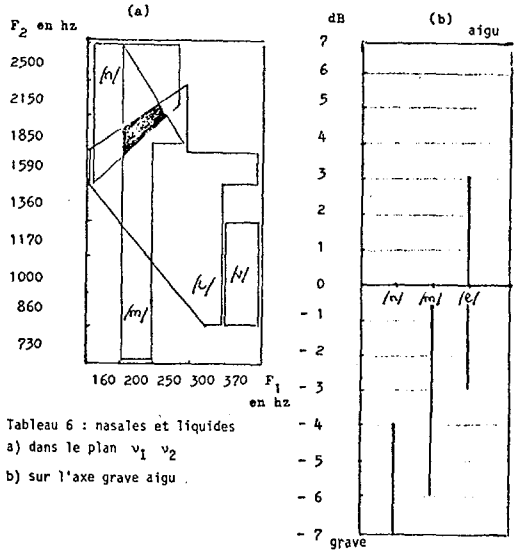


Tableau 6 : nasales et liquides
a) dans le plan ν_1 ν_2
b) sur l'axe grave aigu

TABLEAU 1 : Les phonèmes du français : classement articulatoire.

		d'arrière		d'avant			
				labiales		non labiales	
		ouvertes	fermées	ouvertes	fermées	ouvertes	fermées
voyelles	orales	a (0)	o u (ou)	a œ	ø y (eu) (u)	ɛ (è)	e(é) i
	nasales	ɑ̃ (an) (on)		œ̃ (eun)		ɛ̃ (in)	
semi-voyelles			w(oui)		ɥ (lui)		j(ill)

			labial	dental	palobal
Consonnes	plosives	sourdes	p	t	k
		sonores	b	d	g
	fricatives	sourdes	f	s	ʃ (ch)
		sonores	v	z	ʒ (j,)
	nasales		m	n	ɲ (agneau)
	liquides		l	ʁ	

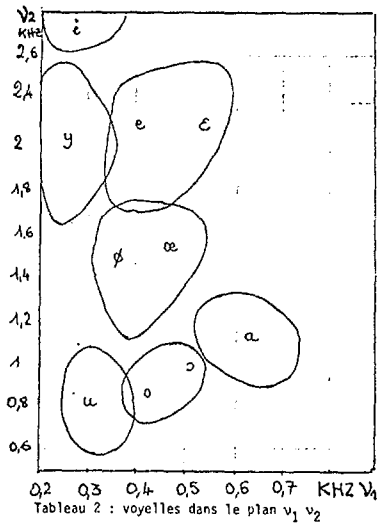


Tableau 2 : voyelles dans le plan ν_1 ν_2

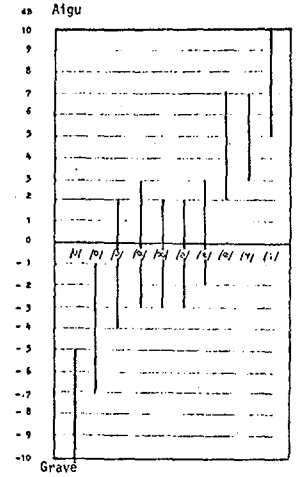


Tableau 3 - Voyelles sur l'axe "grave-aigu"

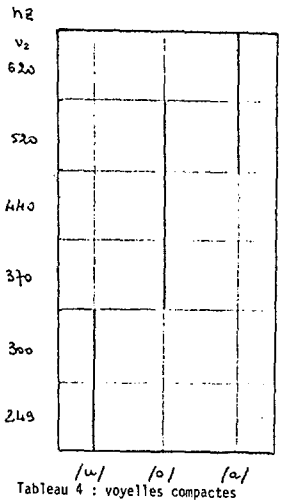


Tableau 4 : voyelles compactes

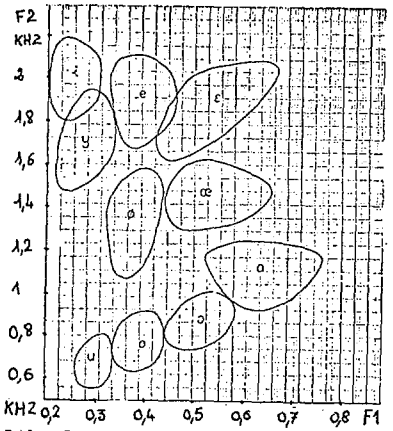
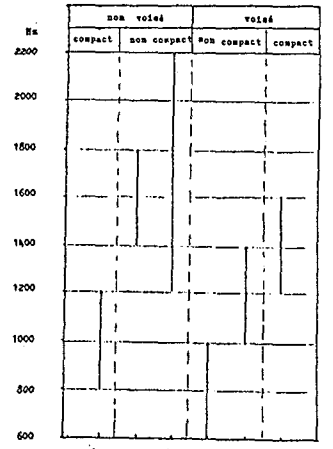


Tableau 7 : voyelles dans le plan F_1-F_2

Tableau 5

Représentation des plosives sur l'axe "fréquentiel"



Ne sont représentées que les occurences de phonèmes ayant une énergie dans la bande 600 - 2200 hz

Tableau 8

Représentation des consonnes sur l'axe "fricatif"