COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL
ET SES APPLICATIONS

NICE du 26 au 30 AVRIL 1977

UTILISATION DE TECHNIQUES DE RECONNAISSANCE DE FORMES SUR LE SIGNAL DE PAROLE

Marc BAUDRY, Benoit DUPEYRAT, Xavier RODET

CENTRE D'ETUDES NUCLEAIRES DE SACLAY - Services d'Electronique de Saclay - B.P. n° 2 - 91190 GIF/S/YVETTE (FRANCE)

RESUME

Cet exposé rend compte d'une étude de synthèse et de reconnaissance du signal de parole, faite aux SES du CEA Saclay depuis 1969. L'originalité de cet effort provient de ce que reconnaissance et synthèse sont effectuées sans avoir recours aux techniques d'analyse spectrale et utilisent exclusivement un miniordinateur. Dans l'ordinateur le signal est codé par la seule information des extrema et en synthèse le signal est restitué par interpolation linéaire.

Un premier jeu de programmes permet de synthétiser en temps réel n'importe quelle phrase française. La chaîne littérale frappée sur console est transformée en chaîne phonétique. A partir de celle-ci des règles sur les trois formants et l'intervalle de voisement permettent de restituer un bon signal de parole.

Un autre jeu de programmes permet d'établir directement sur le signal temporel un certain nombre de diagnostics de reconnaissance : détection du voisement et du début du cycle de voisement, segmentation en phonèmes, identification des phonèmes en classes, analyse des composants formantiques, synchrone du cycle de voisement.

Ces algorithmes originaux sont inspirés des techniques informatiques de la Reconnaissance des Formes. Ils permettront d'atteindre de bonnes performances de reconnaissance multilocuteurs en temps réel, ceci en utilisant seulement un miniordinateur.

Un exemple de synthèse sera donné au cours de l'exposé.

SUMMARY

The present paper deals with a research work on speech signal synthesis and recognition. This study has been carried at the Services d'Electronique de Saclay (C.E.A.) since 1969. The main original point consists in that the recognition and synthesis are carried out without spectral analysis techniques and use only a minicomputer. In the minicomputer the input signal is costed from the information of extrema. The synthesis signal is restored by linear interpolation.

A first set of programs performs the real time synthesis of any french sentence. The character string typed at the console is translated into a phonetic string. Hence the application of rules, on the three formants and the pitch interval allow to obtain a very correct speech signal.

Directly on the time signal, another set of programs allow to obtain a number of recognition diagnosis : pitch detection, the beginning of the pitch interval, segmentation into phonemes and identification of the phonemes into classes, synchronous with the pitch analysis of the formants.

These new algorithmes are derived from a "Pattern Recognition" approach. Satisfactory results are obtained in real-time for multi-speaker recognition, only with a minicomputer.

An example of speech synthesis will be presented at the conference.

UTILISATION DE TECHNIQUES DE RECONNAISSANCE DE FORMES
SUR LE SIGNAL DE PAROLE

I - RAPPEL DES DONNEES PHYSIQUES ET PHYSIOLOGIQUES ESSENTIELLES DU SIGNAL DE PAROLE

Nous appelons signal de parole, le signal temporel recueilli aux bornes d'un microphone. Bien que très variable, le signal de parole a des caractéristiques bien particulières.

a) La plupart du temps le signal est dit "voisé". Cette caractéristique correspond au fonctionnement des cordes vocales du locuteur. Le signal a alors un aspect pseudo périodique, particulièrement caractéristique pour les voyelles. La figure 1 en donne une représentation correspondant à la voyelle œ (comme dans cœur ou neuf). Un observateur distingue facilement la pseudo période T de voisement (pitch interval) et le début d'un cycle de voisement D.C.V.

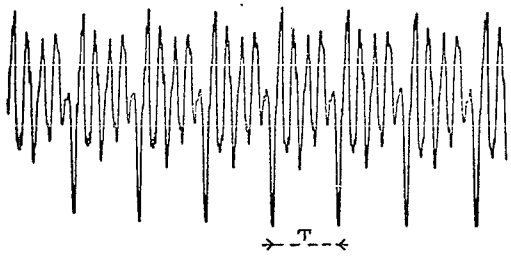

Fig. 1 - Voyelle /œ/

Ce début est caractérisé par une impulsion brève dont le début a toujours le même sens. Le reste du signal est en général d'amplitude moins grande que cette première impulsion. La pseudo période T peut varier suivant l'intonation (la "mélodie" des phonéticiens), de l'ordre de 20 à 100 % pour une même personne et de 100 % d'un homme à une femme. En moyenne T est de l'ordre de 8 ms pour un homme. La figure 4 représente le signal correspondant à la prononciation de "ira". La variation d'un cycle de voisement (C.V.) au suivant est grande maintenant. Il est difficile de considérer le signal comme pseudo périodique.

De ces exemples nous retenons que :

- si le signal est quasi périodique (voyelles) l'information se trouve en entier dans un C.V.
- si le signal est variable (transition phonème-phonème) la variation s'effectue sur une dizaine de C.V.

Revenons au cas de voyelles (fig. 1). Isolons un C.V. fig. 2. Rendons le périodique de période T et calculons la série de Fourier.

Fig. 2 - Une période extraite de la voyelle /œ/

Nous obtenons le spectre de raies de la figure 3a. Ce spectre présente trois maxima essentiels. Séparons le en trois bandes de fréquence F_1, F_2, F_3. Revenant à l'espace des temps, nous obtenons les trois signaux C_1, C_2, C_3, correspondant aux bandes F_1, F_2 et F_3 (fig. 3b). Ces signaux sont appelés composantes formantiques ou simplement formants. Ils ont une allure de sinusoïdes amorties. On s'accorde à estimer qu'ils correspondent à la résonance des cavités pharyngée, buccale et nasale. L'impulsion brève initiale au C.V. correspond au claquement des cordes vocales.

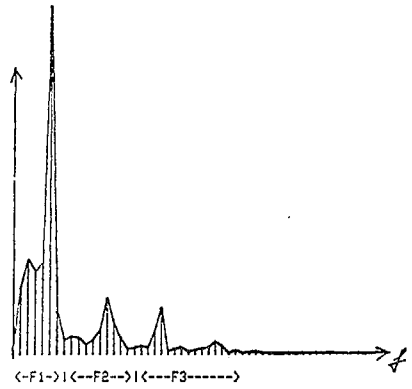

Fig. 3a - Spectre de /œ/

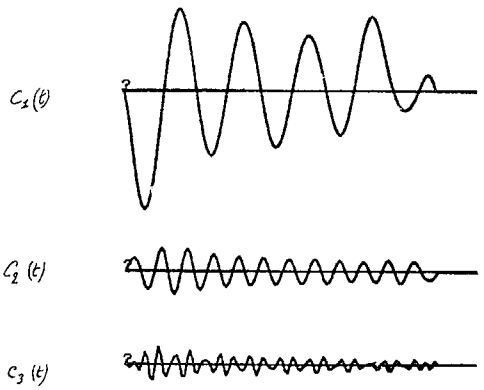

Fig. 3b - Composantes formantiques de /œ/

Lors de la prononciation d'une syllabe telle que "ira", une transition rapide est effectuée de la voyelle quasi périodique i à la voyelle quasi périodique a en passant par la consonne r (Fig. 4)

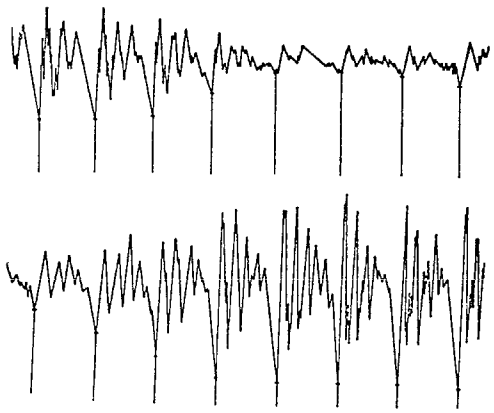

Fig. 4 - Transition ra de ira

En fait le voisement, c'est-à-dire le fonctionnement des cordes vocales, n'est pas interrompu durant la transition. Si le signal est profondément modifié à chaque C.V., une analyse en série de Fourier montre l'existence des trois formants, comme dans le cas quasi périodique précédent.

UTILISATION DE TECHNIQUES DE RECONNAISSANCE DE FORMES SUR LE SIGNAL DE PAROLE

b) Certains sons élémentaires ou phonèmes ne sont pas voisés, c'est-à-dire que les cordes vocales ne fonctionnent pas. L'excitation du conduit vocal est faite par un écoulement d'air turbulent donc bruyant. Il faut d'ailleurs pour cela rétrécir le conduit vocal. Ces phonèmes sont les consonnes fricatives. Ils sont caractérisés par une porteuse bruyante.

c) Enfin certaines consonnes dites plosives sont caractérisées par une occlusion complète ou quasi complète du conduit vocal donc par un silence qui peut être de courte durée. Elles peuvent être voisées ou non.

TABLEAU DES PHONEMES DU FRANCAIS

Voyelles orales

/i/	I	*i* dans *il*, habit, dîner
/e/	}	*é* dans th*é*, d*é*
/ɛ/	{	*è*, *ai*, *ê* dans être, d*ai*s, procès
/y/	U	*u* dans *u*ser, t*u*, sur
/ /	}	*eu* dans f*eu*, j*eu*, p*eu*
/œ/	E	*eu* dans c*œu*r, p*eu*r, n*eu*f
/ə/	}	*e* dans pr*e*mier, l*e*
/ɑ/	}	*a* dans *a*voir, Paris, patte
/a/	A	*a*, *â* dans *â*ne, p*â*te, m*â*t
/o/	O	*o*, *au* dans d*o*s, chev*aux*, hôte
/ɔ/	+	*o* dans r*o*be, *o*r
/u/	W	*ou* dans *ou*vrir, c*ou*vert, l*ou*p

Voyelles nasales

/ɛ̃/	} 1	*in* dans *in*térêt, p*ain*, s*ein*
/œ̃/		*un* dans al*un*, parf*um*
/ɑ̃/	*	*an*, *en* dans *en*trée, bl*anc*
/ɔ̃/	/	*on* dans *on*dée, b*on*, h*on*te

Semi-voyelles

/j/	Y	*y* + voyelle dans *y*eux
/y/		*u* + voyelle dans h*u*ile, l*ui*
/w/	=	*ou* + voyelle dans *ou*i, L*ou*is
/ɲ/		*gn* dans a*gn*eau, bai*gn*er

Consonnes fricatives

/s/	S	*s*, *c* dans *s*ouffler, hélas, cha*s*se, pla*c*e
/z/	Z	*z* ou *s* dans *z*one, rai*s*on, ga*z*
/ʃ/	X	*ch* dans *ch*eval, mâ*ch*er
/ʒ/	J	*j* ou *g* dans *j*ambe, âgé, pa*g*e
/f/	F	*f* dans *f*ou, a*ff*reux, chef
/v/	V	*v* dans *v*ite, ou*v*rir

Consonnes nasales

/n/	N	*n* dans *n*ourrir, fa*n*al, dolme*n*
/m/	M	*m* dans *m*aison, a*m*ener, blê*m*e

Consonnes liquides

/l/	L	*l* dans *l*arge, mo*ll*esse, ma*l*
/r/	R	*r* dans *r*ude, ma*r*i, ouv*r*ir

Consonnes plosives

/p/	P	*p* dans *p*as, dé*p*asser, ca*p*
/b/	B	*b* dans *b*eau, a*b*îmer, clu*b*
/t/	T	*t* dans *t*u, é*t*aler, lu*tt*e
/d/	D	*d* dans *d*ur, bro*d*er, ble*d*
/k/	K	*k*, *c*, *qu* dans *c*aste, *k*épi, *qu*e
/g/	G	*g* dans *g*are, va*g*ue, zig-za*g*

REMARQUES GENERALES

1) On peut estimer que deux systèmes de phonation fonctionnent "en parallèle". Les cordes vocales excitent les cavités du conduit vocal pour produire les "voyelles" de façon continue (à l'exception des fricatives) et d'autre part, par des actions rapides la bouche module le son produit, il en résulte les consonnes. Nous constatons que les transitions voyelle-consonne-voyelle montrent une variation rapide des formants, en amplitude et fréquence.

2) La durée T d'un C.V. peut varier à l'intérieur d'une phase. Une telle variation dite mélodie est un aspect essentiel d'une parole naturelle. En plus de cette variation lente existe une variation rapide, déterminée par chaque phonème, la micromélodie. Son absence donne à la voix un aspect artificiel.

II - DESCRIPTION DE L'EQUIPEMENT ET DU CODAGE ADOPTE

Synthèse et reconnaissance de la parole s'effectuent sur un miniordinateur MULTI 20 doté d'une mémoire rapide de 32 K octets. Le seul matériel annexe est un échantillonneur-codeur. Cet équipement échantillonne le signal du microphone à la fréquence de 10 kHz et sur 64 niveaux. Ensuite seule est conservée l'information des extrêma de cet échantillonnage. Un extremum ei est codé par le couple c_i : (a_i, t_i), a_i étant l'amplitude, t_i l'intervalle de temps le séparant de l'extremum suivant.

Un signal de parole est donc mémorisé dans l'ordinateur par une suite de couples c_i. Il est restitué par interpolation linéaire et envoyé sur un haut parleur.

Recueillir ainsi la parole, la mémoriser et la restituer est un ensemble d'opérations qui ne demande pas d'autre équipement que le MULTI 20, mis à part bien entendu l'échantillonneur-codeur. L'expérience montre que la parole ainsi reproduite est de qualité tout-à-fait suffisante.

En pratique deux mots de 8 bits sont nécessaires pour enregistrer un extremum e_i. Mettant à part les fricatives qui comportent beaucoup de hautes fréquences, on peut estimer la cadence d'information nécessaire à 5 K octets par seconde environ.

III - SYNTHESE DE PAROLE PAR REGLES

1) Un premier programme permet de transformer une chaîne litterale en chaîne phonetique.

La procédure est la suivante :

Un mot entier est recherché dans un dictionnaire d'exceptions, comportant des mots entiers ou des racines. Si le mot n'est pas trouvé le programme analyse chaque lettre selon des règles de transcription phonétique précédemment établies. Le mot étant transcrit, la liaison est traitée.

L'établissement des règles et des exceptions est faite en explorant au dictionnaire de 65 000 mots. Par exemple la transcription de X nécessite 10 règles. De plus 15 mots le comportant doivent être mis dans le dictionnaire d'exceptions.

Le programme de transcription litteral ou phonétique nécessite environ 10 K octets.

2) Un second programme permet, à partir de la chaîne phonétique :

a) d'établir la mélodie et la micromélodie, c'est-à-dire la variation de T,

b) de faire varier à chaque C.V. les trois formants significatifs suivant des règles préétablies.

Cette variation s'effectue par des anamorphoses en amplitudes et en temps pour chaque composante formantique, $c'_i(t) = \alpha\ c_i\ (\beta\ t)$.

Suivant que $\alpha <$ ou > 1 l'amplitude diminue ou augmente, suivant que $\beta <$ ou > 1 il en est de même pour la fréquence (Fig. 5). Chaque C.V. c'(i) est obtenu comme la somme des trois composantes après anamorphose (Fig. 6)

Si l'intervalle du C.V. choisi est plus court que l'intervalle T d'origine, le motif est tronqué en conséquence, sinon il est complété par une portion nulle.

Une voix de femme diffère d'une voix d'homme par une division de T par 2 environ, alors que les fréquences des formants varient peu, environ 17 %. En effet si la fréquence des cordes vocales augmente

beaucoup, les dimensions du conduit vocal varient peu. A partir d'un même ensemble de séquences triples il est donc possible d'obtenir une voix de femme ou une voix d'homme.

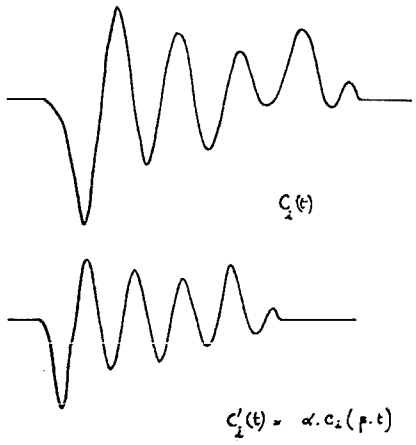

Fig. 5 - Anamorphose en amplitude et temps

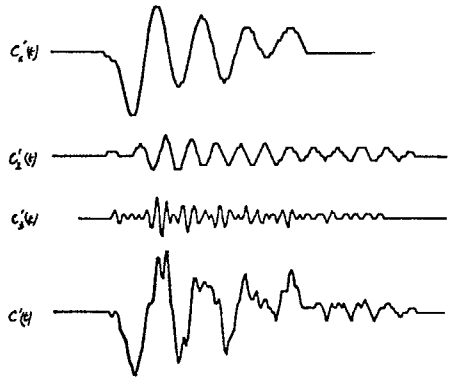

Fig. 6 - Calcul d'un C.V.
$C'(t) = C'_1(t) + C'_2(t) + C'_3(t)$

Il faut noter que cette transformation facile à expliquer et à faire dans l'espace temps, serait beaucoup plus difficile à faire dans l'espace des fréquences.

Ce second programme nécessite environ 20 K octets, y compris la mémoire des échantillons des séquences c_i des 30 phonèmes.

3) Une phrase tapée sur la console de l'ordinateur est prononcée après une seconde environ par le haut-parleur. Sans avoir le naturel d'une phrase prononcée par un être humain et restituée comme nous l'avons indiqué au paragraphe II, le résultat émis est très compréhensible. Comparé aux synthèses produites par circuits résonnants, notre synthèse semble plus claire en ce qui concerne les consonnes, spécialement les plosives. Elle a l'intérêt de ne pas nécessiter d'équipement spécialisé, mais seulement un miniordinateur.

IV - ALGORITHMES DE RECONNAISSANCE PHONETIQUES

Les programmes actuellement au point permettent de

- décider entre silence et parole,
- déterminer le caractère voisé ou non,
- détecter le début du cycle de voisement DCV
- segmenter une parole continue en phonèmes,
- déterminer les composantes formantiques.

Ceci en temps réel, grâce à des techniques informatiques, directement inspirées de la Reconnaissance des Formes.

1) Silence ou parole

Le système est destiné à fonctionner dans l'ambiance bruyante d'une salle d'ordinateurs. Il est donc nécessaire de savoir distinguer entre une parole et un silence ou plutôt une absence de parole, qui peut être bruyante.

Soit a_b l'amplitude moyenne du bruit de fond.

- Nous décidons que nous avons à faire à une parole si pendant 50 ms, il n'y a pas une suite de a_i tels que $|a_i| < a_b$ et $\Sigma\, t_i > 10$ ms.

- La fin d'un son est détecté si la suite de a_i est telle que $|a_i| < a_b$ et $\Sigma\, t_i > 500$ ms.

- Enfin les zones de silence des plosives non voisées sont détectées par la présence d'une suite de a_i tels que $|a_i| < a_b$ et 30 ms $< \Sigma\, t_i < 500$ ms.

Le rapport signal saturé au bruit de fond est environ de 30 db. Une dynamique plus importante demanderait un codage plus précis en amplitude, donc plus d'octets par seconde.

2) Détection du voisement

A partir du codage choisi, il est aisé d'obtenir l'histogramme H_1, des intervalles t_i séparant deux passages par zéro de la dérivée du signal, et également l'histogramme H_2 des intervalles de temps séparant deux passages par zéro du signal.

Les seuls sons non voisés sont les fricatives s, ʃ, f et le début des plosives p, t, k. Ce dernier est caractérisé par un silence, dit d'ailleurs silence occlusif, par allusion à la fermeture du conduit vocal. Un examen comparé de H_1 et H_2 permet de prendre une décision avec sécurité.

3) Début du cycle de voisement

Le début d'un DCV a toujours le même signe, le plus souvent il est de grande amplitude.

Supposons connue la position d'un DCV ainsi que la valeur approchée de T, deux données qui sont en général faciles à obtenir dans une voyelle.

Même dans les transitions rapides, nous sommes assurés que la variation de T ne dépasse pas 15 %. Donc à partir d'un DCV nous cherchons le suivant dans la fenêtre $T_{inf} = 0{,}87\ T$ et $T_{sup} = 1{,}20\ T$.

Trois cas se présentent :

a) sons u, ɔ̃ et parfois a, ɑ̃, ə, o. Il existe un seul extremum e_i dans la fenêtre, ou les autres sont inférieurs à e_{max}. Alors le DCV est e_{max}.

b) sons i, e, ɛ, y, œ̃. L'existence d'un deuxième formant important peut entrainer que le DCV ne correspond pas à e_{max}. Le DCV est l'extremum le plus à gauche de e_{max} tel que $a_i > \frac{3}{4}\, a_{max}$ et situé dans la fenêtre en temps. Cette règle heuristique n'est justifiée que par l'expérience.

c) sons a, ɑ̃, ə, o. Ces phonèmes ont un premier formant situé entre 400 et 1 000 Hz et un deuxième formant faible. Il se peut que la deuxième oscillation du cycle soit d'amplitude plus forte que la première. La valeur de l_{max} est toujours telle que $l_{max} < 1{,}2$ ms, mais la montée du DCV est toujours plus grande. Ce critère permet d'éliminer e_{max} et de revenir en arrière pour choisir un DCV correct.

Les figures 7 et 8 permettent de constater que le programme détectant les DCV fonctionne de façon satisfaisante dans les transitions voyelle-consonne ou consonne-voyelle.

UTILISATION DE TECHNIQUES DE RECONNAISSANCE DE FORMES
SUR LE SIGNAL DE PAROLE

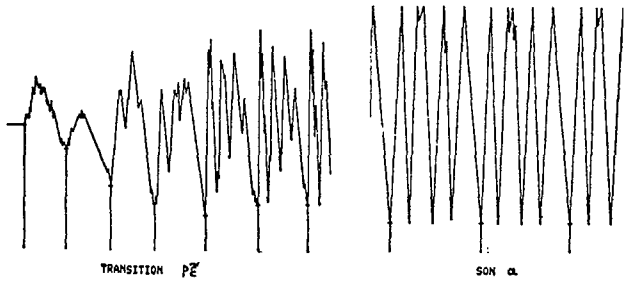

Figure 7

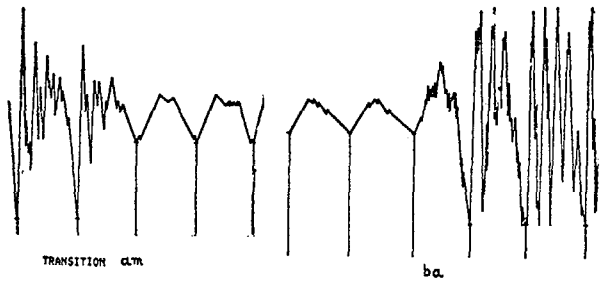

Figure 8

Le DCV est noté sur ces figures par un segment vertical accompagné d'un tiret horizontal.

Les règles de détection ont été trouvées et vérifiées sur un grand nombre de signaux réels de parole. Ils traduisent l'expérience d'un expérimentateur qui déciderait l'emplacement d'un DCV. Il est clair que rien ne permet d'affirmer à priori que les règles sont simples. Enfin il est intéressant de noter que la décision dépend du contexte. Trouver les DCV dans une consonne est une opération aléatoire si on ne connait pas la suite précédente et suivante des DCV des voyelles qui l'encadrent.

4) Segmentation en phonèmes

a) silence ou parole

Les segmentations début et fin de son ont été vues en 1). Un silence occlusif annonce une plosive p, t, k. Actuellement le programme de segmentation sépare p de t, k par l'information du début du signal qui est de signe contraire à celui des DCV.

b) présence de voisement

H_1 et H_2 sont construits sur une durée de un à deux T. Leur examen permet de décider soit la présence de voisement et de localiser son apparition ou sa disparition avec précision sinon de séparer les trois fricatives bruyantes s, ʃ, f. De plus la variation d'amplitude des DCV, la variation de T et de H_1 et H_2 permettent de segmenter par programme la parole continue en phonèmes. Le programme de segmentation fournit de plus une préclassification des phonèmes en cinq classes : voyelles - plosives voisées b,d,g - nasales m,n - liquides l,r - semi voyelles ou fricatives j,z,v.

Il est connu depuis longtemps que l'information des passages par zéro du signal ou de la dérivée c'est-à-dire H_2 ou H_1 permet de faire une séparation entre les voyelles par exemple. Il est important de remarquer que c'est l'utilisation de plusieurs paramètres simultanés qui nous permet d'obtenir les résultats précédents par programme logique. Pour aller plus loin une information précise sur les trois formants est nécessaire.

5) Détecteur de formants synchrone avec le CV

Ayant détecté les débuts de cycle de voisement les CV sont isolés. Un CV est défini par la suite de couples c_i définissant les extrema e_i.

Un"cycle de voisement" de durée T se décompose en une somme de signaux élémentaires appelés formants. Chacun des formants est approximativement une sinusoïde amortie. La connaissance de l'amplitude et de la fréquence de ces formants est essentielle à la reconnaissance des voyelles et des consonnes voisées. Habituellement cette information est obtenue par des techniques d'analyse spectrale, par exemple la Transformée de Fourier Rapide (TFR).

Nous utilisons un algorithme permettant d'obtenir directement ces renseignements sur un cycle de voisement. Il est fondé sur l'idée simple suivante :

Soit un signal s composé d'une somme de deux signaux s_1 et s_2 de frésuence pure f_1 et f_2 et d'amplitudes a_1 et a_2 (figure 9). ($f_2 > f_1$)

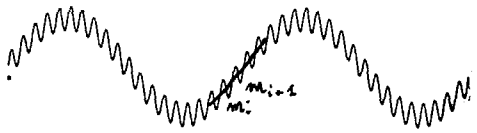

Figure 9

Soit e_i les extrêma de s
Soit m_i le milieu du segment e_i, e_{i+1}
Soit $f_2/f_1 \leq a_2/a_1 \leq (f_2/f_1)^2$ (1)

On montre que si la condition (1) n'est pas vérifiée, l'ensemble des m_i est un échantillonnage du signal s_1. En particulier si $a_2/a_1 < f_2/f_1$ la condition (1) n'est pas remplie. Par la suite nous serons en général dans le cas où $a_2 < a_1$. Remarquons qu'un point m_i d'échantillonnage de s_1 se trouve entre deux extrêmas de s_2. Pour que cet échantillonnage de s_1 soit utilisable il faut d'après la condition de Shannon que le nombre depoints d'échantillonnage soit au moins de 2 f_1 par unité de temps, d'où la condition :

$$f_2 \geq 2f_1 \qquad (2)$$

en fait plus importante que (1) qui en général est vérifiée en pratique.

DESCRIPTION DE L'ALGORITHME

Soit e_i le nom de l'extrêmum de rang i du signal s. Il est défini par l'amplitude a_i et l'intervalle de temps t_i qui le sépare de e_{i-1} (fig. 10).

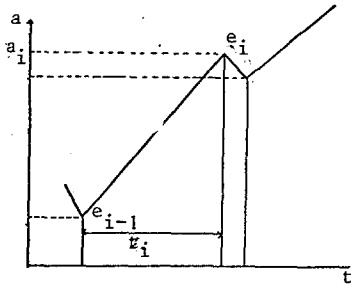

Fig. 10

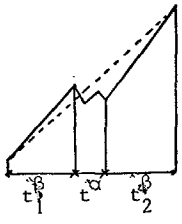

Fig. 11 - cas n impair

UTILISATION DE TECHNIQUES DE RECONNAISSANCE DE FORMES SUR LE SIGNAL DE PAROLE

L'algorithme se compose de deux parties :

1) Filtrage de la composante de fréquence la plus élevée

a) calcul de l'histogramme H des t_i sur le cycle de voisement de durée T.

b) calcul de s_t, seuil correspondant au premier minimum suivant le premier maximum de H.

c) calcul de D, valeur moyenne de la durée des $t_i \leq s_t$. Le formant de fréquence la plus élevée a pour fréquence $1/(2D)$.

d) calcul de A, valeur moyenne de la valeur absolue des différences d'amplitudes successives des segments (e_{i-1}, e_i) tels que $t_i \leq s_t$.

e) construction de s moins le formant : Si $t_i > s_t$ mêmes échantillons, sinon le nouvel échantillon est le milieu du segment e_{i-1}, e_i et e_i est supprimé.

2) Reconnaissance des extrêma du signal filtré

a) calcul de tous les extrêma. Appelons segment α les segments tels que $ti \leq s_t$, segments β les autres

b) soit n le nombre de segments α consécutifs. Soit t^α la somme des t_i correspondants. Soit t_1^β et t_2^β les segments β les entourant.

Si n est impair, remplacer t_1^β par $t_1^\beta + t_\alpha + t_2^\beta$ et supprimer t_2^β.(figure 11)
Si n est pair, remplacer t_1^β par $t_1^\beta + t^\alpha/2$ et t_2^β par $t_2^\beta + t^\alpha/2$ (figure 12).

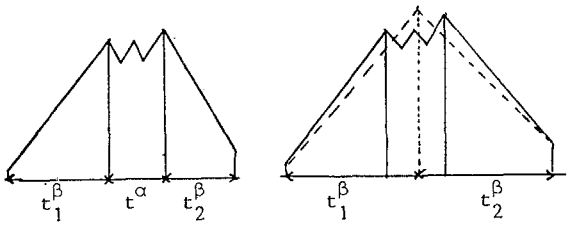

Fig. 12 - cas n pair

A la suite de ces opérations les segments t_i sont de durée supérieure à s_t. Le signal ainsi obtenu représente approximativement s moins le formant de fréquence plus élevée. Tant que le signal élémentaire de fréquence la plus basse n'est pas atteint, le traitement est recommencé en partant de 1).

EXEMPLE D'APPLICATION

L'algorithme a été essayé dans les conditions suivantes :

- fréquence d'échantillonnage ..... 10 kHz
- nombre de niveaux d'amplitude 64

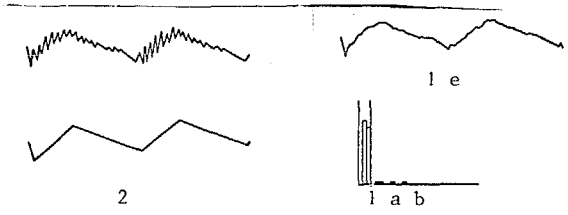

Figure 13

Les figures 13 et 14 représentent les étapes 1 a b e, 2 de l'algorithme pour les voyelles I et E. On peut également représenter le résultat sous la forme d'un sonagramme :

Abscisse : t - ordonnée log f et intensité en noircissement. Les figures 15, 16, 17 et 18 donnent quelques exemples.

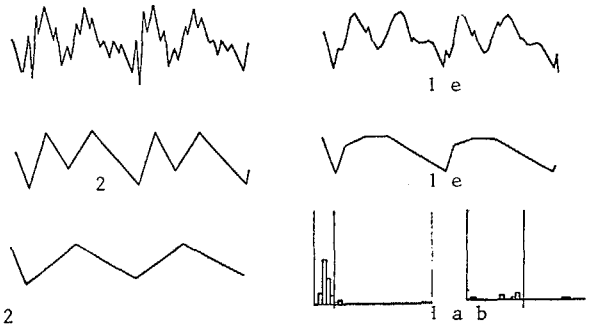

Figure 14

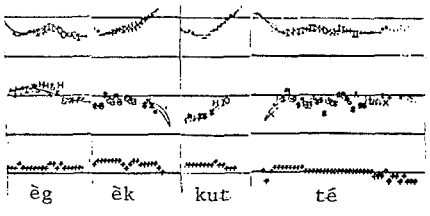

Figure 15

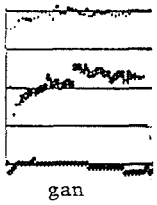

Figure 16

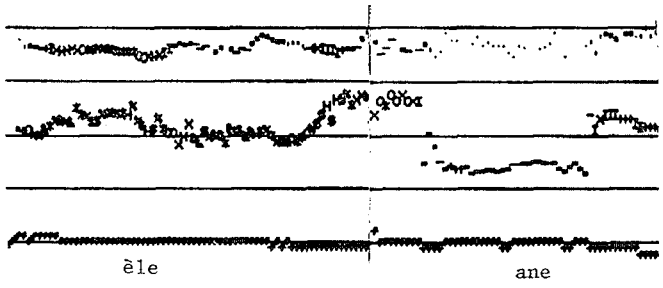

Figure 17

Figure 18

Ce programme est utilisé pour trouver les formants des voyelles et suivre leur évolution dans les transitions voisées du signal. Les valeurs des formants observées pour les voyelles sont en accord avec les valeurs trouvées dans d'autres méthodes. Dans les transitions, la variation des formants est également en accord avec ce que l'on observe dans d'autres méthodes.

En conclusion, l'algorithme décrit permet d'obtenir en temps réel sur un miniordinateur l'amplitude et la fréquence des formants du signal de parole voisée et ceci pour chaque cycle de voisement. Comparés aux techniques d'analyse spectrale, les avantages de notre méthode tiennent des faits suivants : indépendance aux fréquences analysées, à la durée T du cycle de voisement, utilisation des seules informations (a_i, t_i) des extrêma e_i. Rappelons que la TFR nécessite des échantillons équidistants et donc une cadence d'informations beaucoup plus élevée. D'autre part le spectre obtenu dépend de façon critique de l'intervalle de temps choisi et des techniques d'apodisation.

La simplicité de mise en oeuvre de l'algorithme en fait un outil de diagnostic très utile en reconnaissance de parole.

- BAUDRY M., DUPEYRAT B.,"Analyse du signal vocal" - Détection du fondamental et recherche des formants - 7èmes journées du GALF - Nancy 19-21 mai 1976
- RODET X., SANTAMARINA C., "Synthèse sur un miniordinateur du signal vocal dans sa représentation amplitude temps" - 6° journées du GALF. Toulouse 18-20 mai 1976