ncy I

U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

DÉCODAGE ACOUSTICO – PHONÉTIQUE EN COMPRÉHENSION AUTOMATIQUE DE LA PAROLE CONTINUE

THÈSE

présentée et soutenue publiquement le **14 juin 1983**

A L'UNIVERSITÉ DE NANCY I

pour l'obtention du diplôme de

DOCTORAT DE 3e CYCLE EN INFORMATIQUE

par

Mohamed LAZREK

Président : ..... *J.P. HATON*
Examinateurs : ..... *D. COULON*
F. LONCHAMP
G. MERCIER
J.M. PIERREL

Université de Nancy I

U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

DÉCODAGE ACOUSTICO – PHONÉTIQUE EN COMPRÉHENSION AUTOMATIQUE DE LA PAROLE CONTINUE

THÈSE

présentée et soutenue publiquement le **14 juin 1983**

A L'UNIVERSITÉ DE NANCY I

pour l'obtention du diplôme de

DOCTORAT DE 3e CYCLE EN INFORMATIQUE

par

Mohamed LAZREK

Président : ..... *J.P. HATON*

Examinateurs : ..... *D. COULON*
F. LONCHAMP
G. MERCIER
J.M. PIERREL

AVANT - PROPOS

Que J.P. HATON, Professeur à l'Université de Nancy 1, veuille trouver ici l'expression de ma profonde reconnaissance pour m'avoir proposé ce travail et pour l'intérêt manifesté tout au long de sa réalisation. Il me fait l'honneur de présider ce jury et je tiens aujourd'hui à lui exprimer ma profonde gratitude pour la clarté de son enseignement et la formation qu'il m'a donnée.

J.M. PIERREL, Maître-assistant à l'Université de Nancy II, m'a fait l'honneur de s'intéresser à ce travail. Je l'en remercie vivement ainsi que d'avoir accepté de participer à ce jury.

Je remercie Monsieur G. MERCIER, Ingénieur au Centre National d'Etudes des Télécommunications à Lannion, qui a bien voulu accepter notre invitation à ce jury.

Tous mes remerciements vont également à Monsieur F. LONCHAMP, Maître-assistant à l'Université de Nancy II, pour les conseils et le soutien qu'il n'a cessé de me prodiguer depuis le début de cette étude et pour sa présence dans ce jury.

Je remercie également Monsieur D. COULON, Professeur à l'Institut National Polytechnique de Lorraine, qui a bien voulu accepter de prendre connaissance de mon travail et de siéger à ce jury.

J'exprime enfin ma sincère gratitude à toutes les personnes qui ont contribué, par leur gentillesse et leur compétence, à la réalisation de cet ouvrage : en particulier Madame N. GAUTIER pour la frappe du manuscrit avec un soin que l'on ne saurait qu'apprécier.

SOMMAIRE

CHAPITRE III RECONNAISSANCE DE LA PAROLE

CHAPITRE IV DESCRIPTION GENERALE DU SYSTEME

CHAPITRE V RECONNAISSANCE DES VOYELLES

INTRODUCTION

La parole est un moyen de communication privilégié pour l'homme, c'est pourquoi l'automatisation de la synthèse et de la compréhension de la parole conduira à de nombreuses applications dans les domaines de la communication homme-machine et de l'aide aux handicapés. D'où le développement important des recherches dans ce domaine.

Parmi les différents niveaux de traitement qui interviennent dans un système de compréhension automatique de la parole, on trouve le décodage acoustico-phonétique, c'est-à-dire le passage de l'onde acoustique de parole à une suite discrète de symboles phonétiques. Ce décodage joue un rôle fondamental dans un tel système. Il se place en effet en frontal entre le signal vocal et les niveaux linguistiques de traitement ; par suite, de la qualité de la transcription phonétique obtenue dépend en bonne partie la performance globale du système de compréhension.

Nous présentons dans cet exposé la partie segmentation et identification des phonèmes dans le système de compréhension automatique de la parole continue, réalisé au CRIN.

Le système décrit ici représente l'actuel niveau phonétique, chargé de fournir un treillis de phonèmes, des systèmes de compréhension de langages MYRTILLE I et MYRTILLE II.

La réalisation d'un tel décodage se heurte à de nombreuses
ficultés de segmentation du message vocal en unités phonétiques él
taires et d'identification de ces unités du fait de la grande vari
de la parole et des phénomènes phonologiques d'intéraction entre p
voisins.

L'onde vocale est un signal très particulier du fait de son
physiologique, qui met en jeu des phénomènes complexes et difficil
modéliser.

Notre système se fonde sur le phonème en tant qu'unité de r
naissance sans pour autant négliger la notion de noyau syllabique,
dans la phase de segmentation que d'identification.

L'analyse acoustique est effectuée par un vocoder à canaux
détecteur de fondamental. Notre système repose sur un modèle mixte
en jeu deux processus complémentaires fonctionnant en parallèle :

- une identification synchrone des spectres à court terme s
  cessifs fournis par le vocoder à canaux, puis identificat
  phonémique par décision majoritaire.
- une segmentation des noyaux vocaliques et identification
  segments.

L'identification des voyelles, des consonnes voisées et des
sonnes sourdes s'effectue séparément sur la base d'indices détermi
partir des données fournies par vocoder.

Les chapitres I et II examinent la production de la parole naturelle et donnent un aperçu sur les phonèmes du français, ainsi que leur classification acoustique et articulatoire.

Le chapitre III présente le cadre général dans lequel s'insère notre recherche : la reconnaissance automatique de la parole continue. Les problèmes liés au système de reconnaissance phonétique sont évoqués dans la seconde partie de ce chapitre.

La description générale du système constitue le chapitre IV.

Le système procède en trois étapes :

1. Reconnaissance des voyelles (chapitre V)
2. Reconnaissance des consonnes voisées (chapitre VI)
3. Reconnaissance des consonnes sourdes (chapitre VII)

Le chapitre VIII termine sur les résultats expérimentaux.

CHAPITRE I

SYSTEME DE PHONATION

CHAPITRE I

LE SYSTEME DE PHONATION

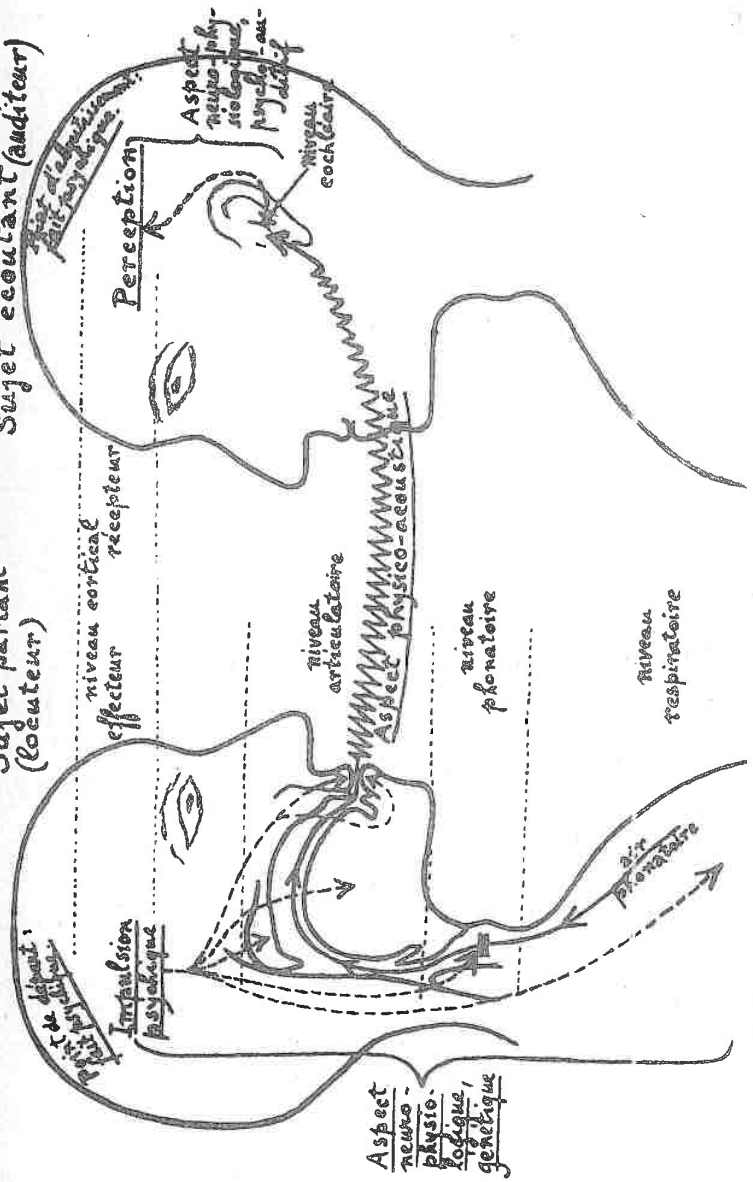
Le but de ce chapitre est de donner quelques notions sur les mécanismes de production de la parole ainsi que quelques éléments de phonétique acoustique que nous avons jugés indispensables pour la suite de l'exposé.

I. PRODUCTION ET PERCEPTION DE LA PAROLE

La parole est le support le plus naturel de communication de la communauté humaine. Le support de la parole est un signal acoustique très complexe par lequel la pensée d'une personne (l'émetteur) est transmise et comprise par une autre personne (le récepteur) (fig. 1). La production de la parole fait intervenir des processus très complexes qui ne sont pas entièrement compris à ce jour.

La pensée de l'homme est assimilée à une source d'informatio qui est codée à travers différents niveaux d'activité dans le cerve pour produire le signal acoustique. Ce signal est transmis par l'int médiaire d'un canal (air ou support technologique) à l'oreille du récepteur. L'oreille décode ce signal et produit une forme que le récepteur peut interpréter.

Dans ce chapitre nous ne décrivons que le système de phonati pour la description du système de l'audition on peut consulter CAELE J. [CAEL-79], RODET X. [RODE-77].



Représentation de la production et de la perception de la parole

II. FONCTIONNEMENT DU SYSTEME DE PHONATION

D'un point de vue physique, la parole est un phénomène vibratoire résultant de trois composantes :

- production de l'énergie.
- excitation du conduit vocal.
- articulation.

Ces facteurs ne sont pas indépendants, ils sont considér séparément dans un simple but de simplification.

II.1. PRODUCTION DE L'ENERGIE

L'énergie nécessaire pour produire un son provient de l contraction de la cage thoracique qui augmente la pression dans poumons. L'air est chassé en direction du larynx qui constitue l'entrée du conduit vocal (fig. 2).

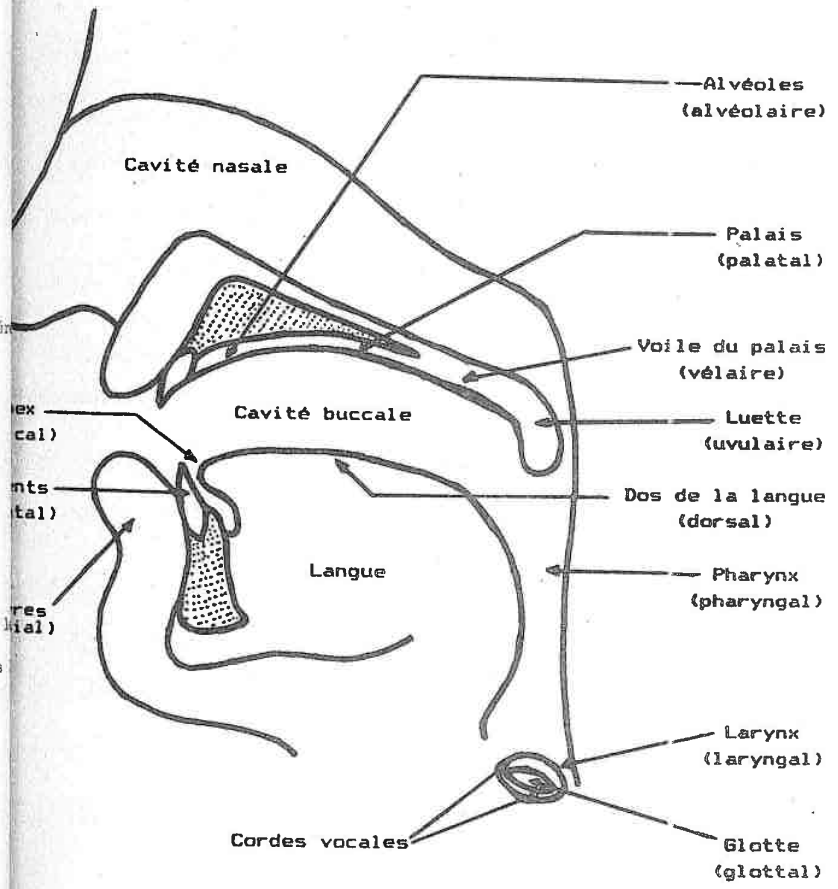


Figure 2 : Schéma de l'appareil phonatoire
d'après H. MELONI [MELO-82]

II.2. SOURCES D'EXCITATION DU CONDUIT VOCAL

L'excitation peut être de deux types : périodique ou bruitée. Ces deux composantes peuvent intervenir simultanément, notamment dans le cas de la production des consonnes sonores.

On distingue, généralement, trois sources d'excitation du conduit vocal :

- la source sonore
- la friction
- l'occlusion.

II.2.1. La source sonore

La source sonore située à l'extrémité du larynx est constituée par les cordes vocales. Elles sont formées d'un muscle et d'un tissu élastique. Elles peuvent fermer totalement le larynx, ou, en s'écartant, déterminer une ouverture triangulaire appelée la glotte. Les vibrations des cordes vocales produisent des impulsions quasi périodiques qui vont exciter le conduit vocal et le conduit nasal. La période des oscillations est déterminée par la masse et la tension des cordes ainsi que la pression de l'air venu des poumons. La fréquence de ces vibrations est désignée sous les noms de pitch,

mélodie ou fréquence fondamentale. Pendant la respiration et la voix chuchotée, l'air passe librement à travers la glotte. De même durant la phonation des sons dits sourds ou non voisés. Au contraire, les sons voisés résultent d'une vibration périodique des cordes vocales. Mais toutes les positions intermédiaires de la glotte depuis la voix chuchotée jusqu'à la voix normale, sont possibles.

Sur la figure 3, est tracée une forme d'onde glottale qui traduit la courbe d'écartement des cordes vocales en fonction du temps, ceci pour un son voisé en parole usuelle.

La pression sous-glottique, la fermeture plus ou moins complète des cordes vocales et l'amplitude de leur mouvement font varier l'intensité du son émis. Les cordes vocales peuvent se tendre, s'allonger ou s'épaissir, ce qui modifie leur réponse élastique et donc leur ouverture. Les limites de variation des sons émis par une même personne sont assez étendues.

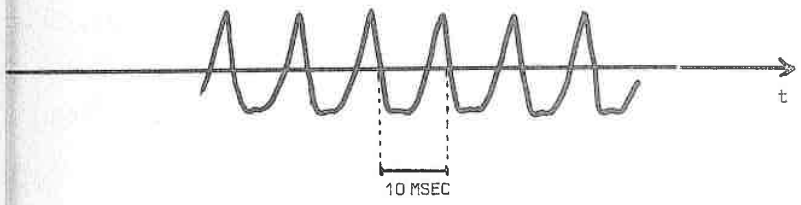


Figure 3. Forme d'onde glottale
(Amplitude en fonction du temps)

II.2.2. La friction

La constriction étroite en un point du conduit vocal consti
une source d'excitation des cavités situées en aval de cette zone.
Ce bruit possède un spectre relativement large et uniforme dont le
contour dépend de la forme de la cavité.

II.2.3. L'occlusion

La troisième source d'excitation est constituée par l'éléva
de la pression dans la zone précédant une occlusion du canal vocal.
L'ouverture brutale du conduit vocal crée une excitation transitoi
dont le spectre décroit inversement avec la fréquence.

II.3. CONDUIT VOCAL - ARTICULATION DES SONS

L'articulation définit la forme du conduit vocal que l'on pe
considérer comme un conduit de section et de longueur variables, l
à son origine par les cordes vocales et à son extrémité par les lèv
(figure 2). La cavité nasale constitue une dérivation supplémentair
dont l'ouverture est commandée par le voile du palais. Sa longueur
jusqu'aux lèvres est d'environ 17 cm pour un adulte.

Les articulations constituées par la langue, le palais, les joues et les lèvres, modifient la forme du conduit qui fonctionne comme un filtre renforçant ou atténuant les harmoniques du spectre du signal d'excitation. Les maxima spectraux sont les formants, ils sont au nombre de quatre, allant jusqu'à 5 000 Hz (figure 4).

L'inertie des organes de l'articulation conduit à réaliser une unité phonétique en trois phases : tension musculaire visant à atteindre la position requise, stabilité des articulateurs durant la tenue du son et retour à la position neutre. Les phonèmes sont définis à partir de la phase stable, mais dans le discours continu cette position n'est jamais réellement atteinte.

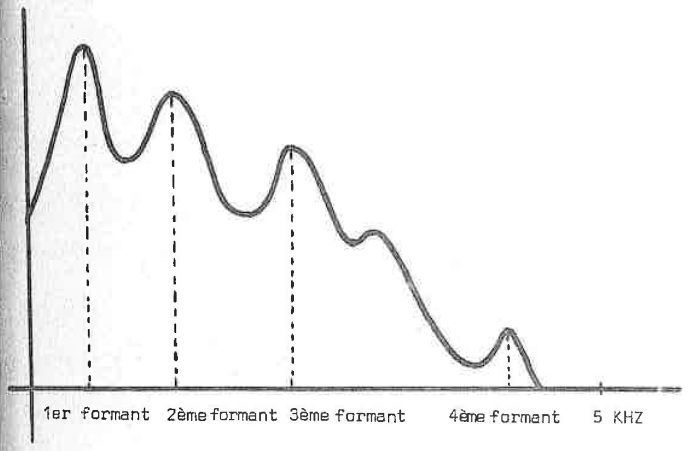


Figure 4 : Position des formants

Les phonéticiens distinguent le lieu d'articulation, où le conduit vocal est le plus resserré, il peut être au niveau des lèvr des dents, des alvéoles, du palais dur ou du velum. L'organe d'arti lation, qui provoque le resserrement du conduit vocal ou lieu d'art culation, peut être la lèvre inférieure ou la langue dans laquelle distingue la pointe, le dos et le dessous (figure 2). Ces points d'articulation sont étudiés en détail au paragraphe V du chapitre suivant.

CHAPITRE II

ETUDE PHONETIQUE

CHAPITRE II

ETUDE PHONETIQUE

I. LES SONS DU LANGAGE

Les sons du langage sont dûs à des vibrations complexes périodiques ou non. La différence entre les sons voisés et les sons non voisés réside dans la vibration des cordes vocales. Dans le premier cas, les cordes vocales sont rapprochées et en vibrant elles modulent le débit d'air phonatoire, dans l'autre, elles sont largement ouvertes et ne vibrent pas.

Pour l'audition, l'homme est apte à reconnaître des formes, c'est-à-dire à classer des signaux complexes en catégories. De même, l'appareil phonatoire est capable de produire les signaux appartenant à telle ou telle catégorie.

Voyons à présent les différents sons constituant le langage.

II. LES PHONEMES DU FRANCAIS

On détermine couramment les phonèmes d'une langue suivant la méthode des paires minimales. On recherche la plus petite variation dans un segment de niveau supérieur, le mot en général, qui le change de catégorie. Par exemple on reconnaît que les mots "père" et "mère" ou "port" et "mort" diffèrent seulement par le phonème initial. Puis

on distingue d'autres phonèmes en opposant d'autres paires : "père" et "port" ou "mère" et "mort". De proche en proche on établit ainsi l'inventaire complet des phonèmes d'une langue. Il existe cependant, entre certains éléments sonores, des différences parfaitement perceptibles qui toutefois n'en font pas des phonèmes différents. On parle alors de "variantes" d'un phonème ou allophones. En français, par exemple, le /r/ "grasseyé" ou "parisien" se distingue du /r/ "roulé". Les deux /r/ sont des variantes d'un même phonème.

Un phonème est un segment phonique qui obéit aux caractéristiques suivantes :

- il a une fonction distinctive,

- il ne peut être décomposé en segments successifs ayant cette propriété,

- il n'est défini que par les éléments qui ont une valeur distinctive.

Cette définition exclut les variantes articulatoires de certain d'entre eux que l'on doit au contexte ou bien à des caractéristiques régionales ou individuelles.

Le nombre de phonèmes est limité, ce sont des unités abstraites de classification. Mais les variantes, ou réalisations concrètes des phonèmes, sont théoriquement illimitées.

Le système phonétique du français comporte 16 voyelles, 17 consonnes et 3 semi-voyelles (figure 5).

LISTE DES PHONEMES DU FRANÇAIS.

Phonème (1)	Mot clef	représentation machine	nature
a	plat	A (2)	Voyelles
ɑ	mât	AA (2)	
i	il	I	
y	nu	Y	
ɔ	bol	O	
o	eau	OO	
ə	le	EE	
ɛ	lait	AI	
e	blé	E	
ø	peu	EU (2)	
œ	heure	OE (2)	
u	ou	OU	
ã	an	AN	Voyelles nasales
ɔ̃	on	ON	
ɛ̃	lin	IN (2)	
œ̃	brun	OEN (2)	
v	vie	V	Fricatives sonores
z	zéro	Z	
ʒ	je	ZZ	
f	feu	F	Fricatives sourdes
s	son	S	
ʃ	chat	SS	
b	bon	B	Occlusives sonores
d	dans	D	
g	gare	G	
p	pas	P	Occlusives sourdes
t	tas	T	
K	coût	K	

Phonème	Mot clef	Représentation machine	nature
m	ma	M	Consonnes nasales
n	nous	N	
ɲ	agneau	GN	
ŋ	camping	NG	
R	rue	R	Liquides
l	lent	L	
w	voir	W	Semi-voyelles
j	bailler	J	
ɥ	huit	HU	

(1) Symbolisme d'après l' IPA (International Phonetic Association)
(2) Phonèmes pratiquement indistinguables

Figure 5

III. LES VOYELLES

III.1. CLASSIFICATION

On distingue habituellement un système minimal des voyelles du français, celui qu'il faut réaliser pour se faire comprendre, et un système maximal qui permet de parler correctement (figure 6), (les différents traits utilisés sur la figure 6 sont décrits au paragraphe suivant). Généralement, les systèmes de reconnaissance automatique de la parole se contentent du système minimal.

Les voyelles sont essentiellement caractérisées par la position de leurs deux premiers formants : ce qui conduit à une représentation schématique dans le plan F_1, F_2 (figure 7). Dans une autre classification on dispose, sur un axe le degré d'ouverture, sur l'autre le caractère antérieur ou postérieur de l'articulation (figure 8).

III.2. ETUDE DES VOYELLES

Les voyelles sont étudiées à partir de l'examen des films radiocinématographiques de l'appareil phonatoire. Elles sont constituées par le maintien d'une même position de la partie de la langue qui produit l'aperture vocalique. Aucun changement, durant ce laps de temps, n'intervient dans la distance entre la langue et la voute palatale au lieu d'articulation. Par contre, les autres organes continuent leurs mouvements.

La durée des voyelles varie avec la vitesse d'élocution et avec la nature de la syllabe (accentuée ou non).

On définit la vitesse d'élocution par le nombre de phonèmes prononcés par seconde. Pour une élocution assez rapide (8 phonèmes p seconde) la durée d'une tenue vocalique varie entre 20 et 80 ms pour la syllabe non accentuée, et de 80 à 160 ms en position accentuée.

Système maximal du français

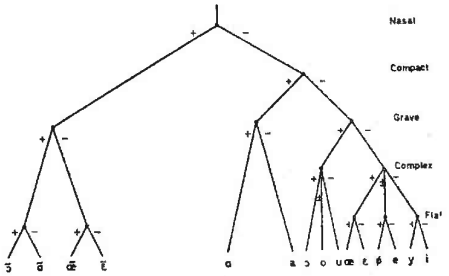


Schéma des traits distinctifs qui servent à opposer les voyelles françaises les unes aux autres. + signifie que le groupe en question possède la marque positive indiquée à droite (voir par exemple le pr mier groupement en + nasales et - nasales, les dernières = les voyel orales).

Système minimal du français

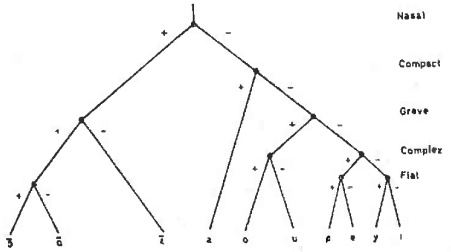
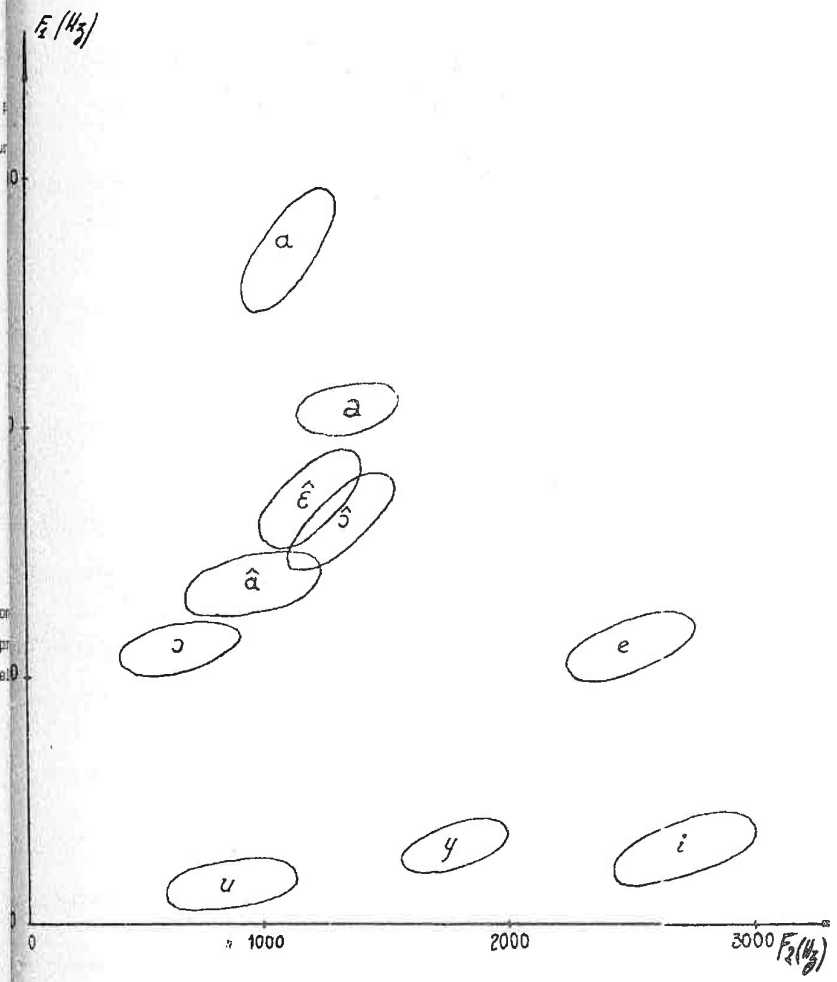


Schéma du vocalisme réduit qui résulte de la disparition des oppositi faibles dans le système. C'est le système minimal qu'il faut réalise pour se faire comprendre en français.

Figure 6 - D'après Malmberg [MALM-72]



Représentation des voyelles françaises dans le plan (1° formant, 2° formant)

Figure 7 - D'après Haton [HATO-74-b]

	palatales		vélaires
	non-arrondies	arrondies	
FERMEES	i	y	u
MI-FERMEES	e	ø	o
MI-OUVERTES	ɛ	œ	ɔ
OUVERTES	a		ɑ

Figure 8 : D'après Malmberg [MALM-72]

III.3. DUREES DES VOYELLES

En français, la durée n'est pas un trait distinctif. Une voyel s'allonge ou se raccourcit en fonction de nombreux facteurs tels que type de consonne qui suit, la position dans la phrase, la fonction du mot où la voyelle figure, la nature de la voyelle. Par exemple, la durée croît avec l'ouverture de la voyelle. Les nasales sont relativement plus longues que leurs correspondantes orales.

III.4. FORMANTS DES VOYELLES

L'étude des formants des voyelles nécessite non pas l'observation des fréquences précises mais des régions du plan F_1, F_2 (figure 7 susceptibles de se recouvrir partiellement. D'après LIENARD [LIEN-73] il est plus exact de définir les fréquences des formants à un coefficient d'affinité près.

Plusieurs facteurs sont responsables de ces variations :

i) La fréquence d'un formant n'est pas indépendante de celle du fondamental, et même de l'intensité sonore. En général elle croît proportionnellement au fondamental mais avec un facteur faible.

ii) Pour un même individu, à une même fréquence fondamentale, la dispersion des formants est déjà importante car l'articulation n'est jamais répétée de façon identique.

iii) Enfin la fonction de transfert dépend des dimensions du conduit vocal et varie donc d'un individu à l'autre.

III.5. VOYELLES NASALES

Le français connaît quatre voyelles nasales /ɔ̃/, /ɛ̃/, /ɑ̃/, /œ̃/ dont les positions articulatoires sont proches de celles des voyelles orales /ɔ/, /ɛ/, /a/ et /œ/ respectivement. Dans la parole courante, l'ouverture du velum, qui caractérise la nasalisation, est relativement lente par rapport aux autres mouvements.

L'ouverture retardée du velum semble avoir une conséquence importante sur le son produit : les voyelles nasales sont moins stables que les voyelles orales, à cause de la nasalisation croissante (fig. 9).

IV. LES CONSONNES

A partir du pitch, on regroupe les consonnes en deux classes : les consonnes voisées et les consonnes non voisées.

Dans la classe des consonnes voisées on distingue quatre groupes :

- les fricatives
- les plosives
- les nasales
- les liquides

et les consonnes non voisées en deux groupes :

- les plosives
- les fricatives.

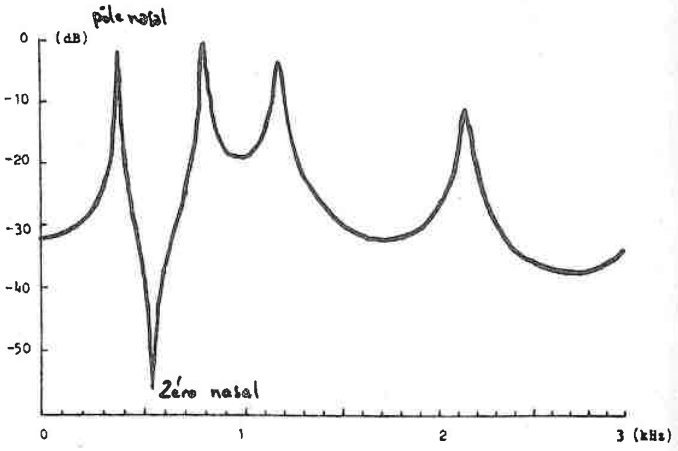


Figure 9 : Spectre de la voyelle nasale /ã/

IV.1. FRICATIVES NON VOISEES

Elles sont caractérisées par une zone bruitée dont le spectre présente une forme générale typique, mais variable suivant le contexte phonétique. La fricative sourde /ʃ/ présente des variations importantes du spectre dans les basses fréquences en fonction du contexte.

De façon générale, l'amplitude maximale du signal est toujours plus faible que dans la voyelle et décroissante de /ʃ/ à /p/.

IV.2. FRICATIVES VOISEES

Nous pouvons dire schématiquement que les fricatives voisées sont produites par la superposition d'une excitation de voisement au bruit de la fricative sourde de même articulation. Plus précisément, cette tenue coarticulée réalise la continuité des trajets des formants.

IV.3. PLOSIVES NON VOISEES

La partie stable d'une plosive sourde est un silence (occlusion du conduit vocal) suivi d'un bruit d'explosion (relachement de l'occlusion) qui dépend de la consonne et de la voyelle adjacente (figure 10).

L'étude des spectres dans les transitions [ALIN-72] nous montre que :

- /p/ est caractérisé par un spectre ayant des basses fréquences importantes bien plus rapidement que /t/ et /k/.

- /t/ est caractérisé par un spectre plat et situé assez h
en fréquence, par opposition avec /k/ qui présente un maximum uniq
au dessus du deuxième formant.

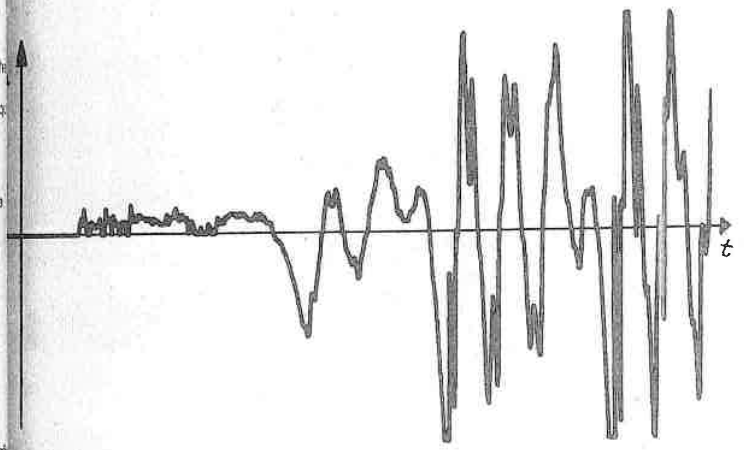
Les caractéristiques de l'explosion et la durée du silence
occlusif dépendent beaucoup de la position de la consonne dans la
phrase (initiale ou finale par exemple).

IV.4. PLOSIVES VOISEES

Les plosives sonores sont /b/, /d/, /g/. Leurs spectres pr
tent un formant très bas. Dans les transitions de début et de fin d
voyelle, les sens de variations des fréquences formantiques sont se
blement les mêmes que dans le cas des plosives sourdes de même arti
lation.

IV.5. NASALES

Lors de la prononciation de la consonne nasale /m/ le conduit
buccal est complètement obturé par les lèvres tandis que pour /n/ o
/y/ il est obturé par la langue, et dans les deux cas le vélum est
ouvert ajoutant le couplage des cavités nasales.
Le signal est quasi-stationnaire.



Bruit d'explosion après un silence occlusif pour /a t a/.

(AMPLITUDE EN FONCTION DU TEMPS)

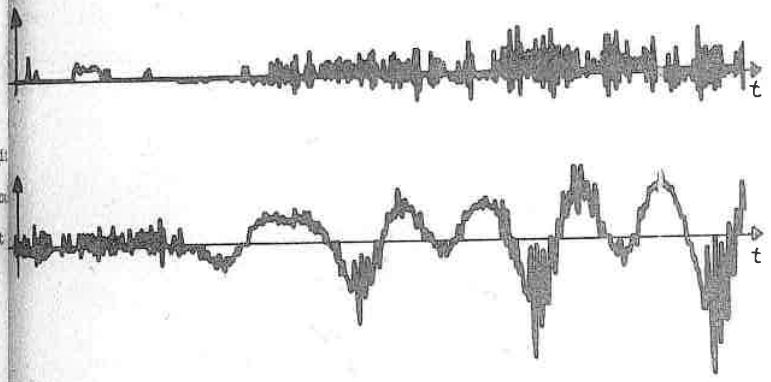


Figure 10 : Bruit d'explosion après un silence occlusif pour /i t i/

IV.6. LIQUIDES

i) La consonne /r/ :

Elle peut présenter des réalisations très variées (figure 11). La coarticulation intervient toujours de façon essentielle. En gén il n'y a pas de tenue acoustique sur le signal mais une diminution, souvent assez régulière, de l'amplitude et une sorte de perturbatio signal à chaque pseudo-période, du point de vue de sa forme et de sa

ii) la consonne /l/ :

Du point de vue articulatoire, le /l/ est une latérale, c'est à-dire l'air passe de manière dissymétrique de part et d'autre de l langue dont la pointe appliquée au palais constitue un obstacle cen- tral. Sur le plan spectral, le /l/ est fluide, c'est-à-dire qu'il n' pas de structure vraiment propre et ne présente pas de stabilité car téristique : il est très dépendant du contexte, comme on le constat dans les occurrences /ili/, /ulu/ et /ala/ (figure 12).

L'amplitude du signal est forte en général, parfois presque égale à celle des voyelles, ce point sera détaillé lors de l'étude de la courbe d'énergie pour la détection des noyaux vocaliques dans le chapitre V.

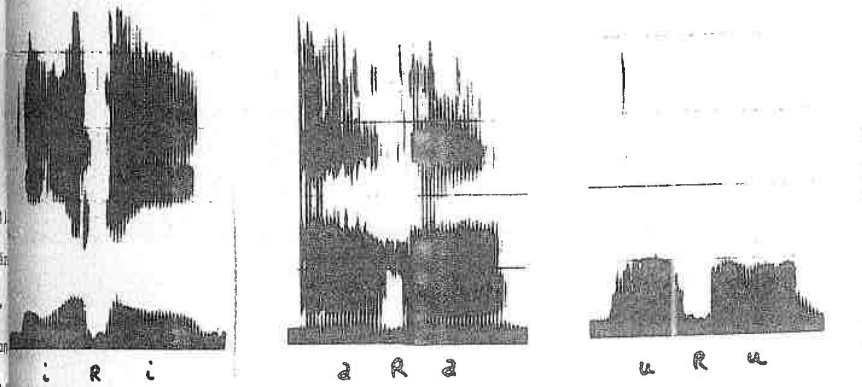


Figure 11 : La consonne /R/ en contexte.

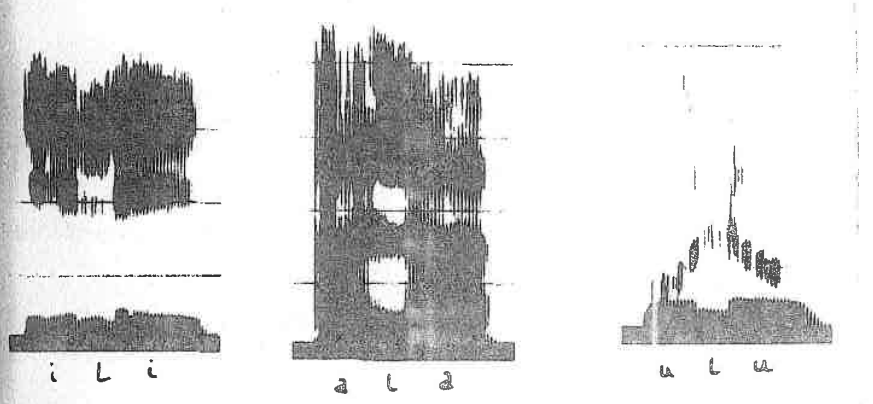


Figure 12 : La consonne /L/ en contexte.

Les sonagrammes ont été réalisés par F. LONCHAMP

V. CLASSIFICATION PHONETIQUE

Les critères permettant de définir les phonèmes sont propres
type d'approche sous lequel on les considère (articulatoire, percep
acoustique). Si à l'évidence la définition acoustique est primordia
pour la reconnaissance automatique, l'aspect articulatoire ne doit
être négligé à cause du lien étroit qui unit ces classifications
[SHIR-83].

V.1. CLASSIFICATION ACOUSTIQUE

La grande variabilité des spectres caractéristiques d'un phon
rend délicat le problème de la désignation de formes distinctives.
Cependant JAKOBSON [JAKO-63] propose un système binaire de traits di
tinctifs fondé sur la corrélation plus ou moins directe de la matièr
acoustique d'un phonème avec ses propriétés articulatoires. On peut
distinguer deux types de traits :

- Les traits de tonalités.
- Les traits de sonorités.

V.1.1. Traits de tonalités

Les traits de tonalités mesurent la quantité de l'énergie dan
les extrêma du spectre. Ces traits sont :

Grave / aigu : concentration de l'énergie dans les basses (ou hautes) fréquences.

Bémolisé / non-bémolisé : abaissement (ou non) des composants de haute fréquence.

Diésé / non-diésé : les phonèmes diésés s'opposent aux phonèmes non-diésés par un déplacement vers le haut ou un renforcement de certains de leurs composants de haute fréquance.

V.1.2. Traits de sonorité

Les traits de sonorités mesurent la quantité et la concentration de l'énergie dans le spectre. Ces traits sont :

Consonantique / non-consonantique : énergie totale réduite (ou élevé)

Compact / Diffus : concentration d'énergie plus élevée (ou plus réduite) dans une région relativement étroite et centrale du spectre, accompagnée d'un accroissement (ou diminution) de la quantité totale d'énergie.

Voisé / non-voisé : présence (ou absence) d'une excitation périodique de basse fréquence.

Vocalique / non-vocalique : présence (ou absence) d'ure structure de formant nettement définie.

Discontinu / continu : silence (dans les bandes de fréquences supérieures à F₀) suivi et/ou précédé d'une diffusion de l'énergie sur une large bande de fréquence (sous forme d'une explosion ou d'une transition rapide des formants vocaliques).

	a	ɑ	ɔ	ɛ	œ	ø	e	o	i	y	u	ɑ̃	ɔ̃	ɛ̃
Nasal	−	−	−	−	−	−	−	−	−	−	−	+	+	+
Compact	+	+	−	−	−	−	−	−	−	−	−	0	0	0
Grave	−	+	+	−	−	+	−	−	−	−	+	+	+	−
Complex	0	0	+	+	+	+/−	+/−	+/−	−	−	−	0	0	0
Flat	0	0	0	−	+	0	−	+	−	+	0	−	+	−

Traits Acoustiques du Système Vocalique du Francais (d'après B. MALMBERG)

	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	ɲ	r	l	w
Nasal	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	−	−	−
Vocalique	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	+/−	+	+
Interrompu	+	+	+	+	+	+	−	−	−	−	−	−	+	+	+	+/−	+	−
Continu	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+/−	+	+
Compact	−	−	+	−	−	+	−	−	+	−	−	+	−	−	+	+/−	−	−
Aigu	−	+	+/−	−	+	+/−	−	+	+	−	+	+	−	+	+	+/−	+	−
Voisé	−	−	−	+	+	+	−	−	−	+	+	+	+	+	+	+	+	+

Traits Acoustiques du Système Consonantique du Francais (d'après M. ROSSI)

Figure 13

La figure 13 présente les descriptions des voyelles et des consonnes du français, au moyen d'un ensemble de traits, empruntées respectivement à [MALM-72] et [ROSS-77].

V.2. CLASSIFICATION ARTICULATOIRE

On utilise, pour cette classification, trois types de paramètres qui définissent une utilisation des organes de la phonation. Ces paramètres sont :

- mode d'articulation
- organe articulant
- lieu de l'articulation

V.2.1. Mode articulatoire

Le mode articulatoire est lié aux diverses sources d'excitation du conduit vocal, elles peuvent être utilisées simultanément à des degrés divers. Les caractéristiques du mode articulatoire sont :

- Sonore : vibration des cordes vocales
- Sourd : absence de vibration des cordes vocales
- Nasal : participation du conduit nasal à la production du phonème.
- Latéral : passage de l'air sur les parois latérales du conduit vocal obstrué en son centre par la langue.
- Occlusif : fermeture partielle ou totale du conduit vocal suivi d'un relachement brusque.

- Constrictif : rétrécissement du passage de l'air en un poi
  du conduit vocal.
- Vibrant : la constriction du conduit vocal n'est pas perma
  elle varie sous l'action de la pression de l'air ex
  et de la tension musculaire qui ramène ensuite l'art
  lateur dans sa position initiale.

V.2.2. L'organe articulant

L'organe qui permet l'articulation caractérise les phonèmes
la façon suivante :

- Labial : participation des lèvres.
- Apical : implication de l'extrémité de la langue.
- Dorsal : utilisation du dos de la langue.
- Uvulaire : action de la luette.

V.2.3. Le lieu de l'articulation

Le lieu de l'articulation désigne la zone du conduit vocal qu
participe à la formation du son. Il est désigné par les noms suivants

- Labial : niveau des lèvres.
- Glottal : niveau de la glotte.
- Pharyngal : niveau du pharynx.
- Dental : niveau des dents.
- Alvéolaire : niveau des alvéoles.
- Laryngal : niveau du larynx.
- Velaire : niveau du voile du palais.

- Palatal : niveau du palais.

Tous ces éléments ne sont pas utilisés dans la réalisation des phonèmes du français. La figure 14 présente une classification articulatoire de ces unités empruntée à MALMBERG [MALM-72].

VI. CONCLUSION

Quels enseignements tirer de ces observations pour la reconnaissance de la parole ?

Au premier abord, le problème paraît ardu, en raison :

1. de l'instabilité du signal.
2. du grand nombre de paramètres.
3. de la variété des évènements acoustiques.
4. de l'influence du contexte.
5. des variantes d'un même phonème.
6. des conditions de prononciation (locuteur, accentuations, longueur de l'énoncé ...).

Le système de reconnaissance doit pouvoir :

- traiter les divers évènements par des procédures spécialisées, modulaires, comportant de nombreuses règles.
- être très souple.

	a	ɑ	ɔ	ɛ	œ	o	e	ø	i	y	u	ɑ̃	ɔ̃	ɛ̃
Orales	+	+	+	+	+	+	+	+	+	+	+			
Nasales					-							+	+	+
Labiales					+			+		+				
Palatales	+			+	+		+	+	+	+				+
Vélaires		+	+			+					+	+	+	
Ouvertes	+	+										+		
Mi-ouvertes			+	+	+								+	+
Mi-fermées						+	+	+						
Fermées									+	+	+			

Classification articulatoire des voyelles du français d'après MALMBE

	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	ɲ	r	l	w	j
Occlusives	+	+	+	+	+	+													
Spirantes							+	+	+	+	+	+						+	+
Nasales													+	+	+				
Latérales																	+		
Vibrantes																+			
Sonores				+	+	+				+	+	+	+	+	+	+	+	+	+
Sourdes	+	+	+				+	+	+										
Labiales	+			+			+			+			+					+	
Apicales		+			+									+		+	+		
Prédorsales								+			+								
Dorsales			+			+									+			+	+
Uvulaires																+			
Dentales		+			+		+			+				+			+		
Alvéolaires								+			+						+		
Palatales															+				+
Vélaires			+			+									+				

Classification articulatoire des consonnes du français d'après MALMBE

Figure 14

CHAPITRE III

RECONNAISSANCE DE LA PAROLE

CHAPITRE III

RECONNAISSANCE DE LA PAROLE

I. LES APPROCHES EN RECONNAISSANCE DE LA PAROLE

On distingue deux approches dans le domaine de la reconnaissance de la parole :

- l'approche globale.
- l'approche analytique.

I.1. L'APPROCHE GLOBALE

Dans l'approche globale appelée aussi reconnaissance par mots, on désigne par mots soit des mots au sens lexical, soit des expressions, voire des phrases. Cette approche calque à peu près le fonctionnement du cerveau humain lorsqu'il comprend un énoncé oral dans son ensemble sans s'arrêter aux divers composants du message.

L'approche globale consiste à effectuer une comparaison globale, à l'aide de paramètres acoustiques entre des formes de références stockées dans la mémoire d'un calculateur et ce qui a été prononcé par un locuteur [HATO-74-b], [DIMA-81]. Donc, il est nécessaire de conserver en mémoire les différentes représentations possibles de tous les mots du vocabulaire. Le point délicat de cette approche est de trouver la meilleure représentation possible de mots. Il n'est pas possible de conserver toute l'information car il faudrait alors plusieurs dizaines

de milliers de bits par seconde de parole. Diverses méthodes exis pour compresser cette information : méthodes spectrales par banc filtres ou par analyse de Fourier par exemple, méthodes temporell par passage par zéro ou L.P.C. [HATO-76-a], [MART-76].

Par contre, le traitement acoustique préliminaire est ass simplifié, il n'y a pas de problème de segmentation, tous les mot doivent être séparés par environ 300 ms, ce qui permet une segmen triviale en mots.

Il existe actuellement des systèmes de reconnaissance de m isolés dont le taux de reconnaissance est de 99% pour des vocabul limités. L'utilisation de ces systèmes suppose une phase préalabl d'apprentissage, au cours de laquelle la machine met en mémoire l paramètres acoustiques correspondant aux mots du répertoire pron par l'utilisateur.

Dans la phase de reconnaissance, la machine compare le sig reçu avec les mots de référence jusqu'à trouver le mot le plus vo

La plupart des systèmes de reconnaissance des mots isolés mono-locuteur, c'est-à-dire les références d'un locuteur ne sont utilisables par un autre. Chaque locuteur, avant l'utilisation d système doit construire ses propres références. Les tendances act visent des systèmes multilocuteurs.

L'un des inconvénients des systèmes de reconnaissance de m isolés est le fait qu'ils n'acceptent que des vocabulaires restr à cause de l'encombrement de la mémoire et du temps de calcul néc

La programmation dynamique est largement utilisée dans ces systèmes pour compenser les variations de durée et de rythme dans les mots à comparer.

Un autre pôle de recherche s'est développé ces dernières années : la reconnaissance de mots enchaînés. Ces systèmes sont utilisés dans des applications ne nécessitant qu'un petit nombre de mots dans le vocabulaire. Après la prononciation d'une suite de mots, le système reconnaît les différents mots. Dans notre laboratoire, un système de mots accolés est en cours de réalisation par J. DIMARTINO, dans le cadre de sa thèse de docteur ingénieur.

I.2. L'APPROCHE ANALYTIQUE

Les limites des méthodes globales ont conduit les chercheurs vers un second type de méthodes, dites analytiques ou phonétiques dont l'objectif est de déterminer dans un premier temps des éléments minimaux ou sons élémentaires (phonèmes, diphonèmes, syllabes...) pour ensuite reconstruire la phrase prononcée comme séquence de ces sons élémentaires.

Le processus de traitement nécessitera donc dans ce cas différentes étapes : la segmentation, l'identification des segments et la reconnaissance de la phrase. Toute notre étude sera centrée sur la première et la deuxième étape. La troisième étape, constituée de plusieurs niveaux, est prise en charge au sein de notre équipe par d'autres chercheurs, elle n'est que sommairement présentée dans ce chapitre, bien que étroitement liée à notre propre travail.

L'approche analytique se prête mieux que la précédente à la reconnaissance multilocuteurs et au traitement de la parole continue. Les chercheurs qui travaillent à la reconnaissance de mots au sein de gros vocabulaires utilisent aussi cette méthode [PERE-79-a], un tel travail est mené au CRIN par J.F. MARI dans le cadre de son doctorat d'état.

II. LES DIFFERENTS NIVEAUX D'ANALYSE DANS LES SYSTEMES DE RECONNAISSANCE

II.1. INTRODUCTION

Le message acoustique à analyser peut être considéré de plusieurs manières :

- c'est un signal continu très complexe.
- c'est une suite de phonèmes :

  /j/, /ø/, /v/, /u/, /d/, /r/, /e/ , ...

- c'est une suite de mots après assemblage des phonèmes entre eux :

  "je" , "voudrais" , "le" , "poste" , ...

- c'est une suite de phrases après assemblage des mots entre eux en reconnaissant leurs valeurs grammaticales :

  "je voudrais le poste numéro deux ; ... "

- c'est une action à réaliser quand la signification de la phrase a été perçue :

  "action : COMMUTER LE POSTE NUMERO DEUX SUR LA LIGNE D'APPEL".

- c'est une question à poser à l'utilisateur quand le message est mal compris ou dans le cas d'ambiguïté.

Dans un système de reconnaissance automatique de la parole, on distingue un certain nombre de niveaux. Ce découpage n'est pas universel. Certains auteurs regroupent dans un même niveau, deux niveaux adjacents, d'autres éliminent certains niveaux facultatifs. Cependant, tout le monde est d'accord sur la nécessité de ces niveaux hiérarchiques pour la compréhension automatique de la parole :

- niveau acoustique.
- niveau phonétique.
- niveau phonologique.
- niveau prosodique.
- niveau lexical.
- niveau syntaxique.
- niveau sémantique.
- niveau pragmatique.

Un système de reconnaissance de la parole correspond à une mise en oeuvre particulière de ces différents niveaux [HATO-81-b]. Les informations obtenues par l'analyse d'un niveau donné représentent l'entrée du niveau suivant.

Dans les paragraphes suivants on va voir les fonctions de ces différents niveaux.

II.2. LE NIVEAU ACOUSTIQUE

Parmi les tâches effectuées par le niveau acoustique on peut citer :

- la détection des débuts et fins des messages parlés (séparation parole - non parole).
- la normalisation du signal vocal.
- le traitement des bruits qui entachent le signal : séparation parole - bruit.
- l'extraction des paramètres les plus significatifs.

II.3. LE NIVEAU PHONETIQUE

Il s'agit ici de segmenter le message parlé en phonèmes puis de reconnaître ces phonèmes.

Deux techniques de base sont utilisées pour réaliser la transformation du signal continu en une chaîne discrète d'éléments phonétiques :

- segmentation du signal vocal [HATO-79], [LAZR-81], [MERC-82], [SANC-78], puis identification des segments par des méthodes qui relèvent à la fois de la phonétique et de la reconnaissance des formes.

- identification systématique des tranches de signal de durée constante (variable selon les systèmes) puis regroupement des segments identiques [MICL-81], [JELI-81].

Pour augmenter la fiabilité des chaînes de segments, on donne le plus souvent plusieurs étiquettes possibles à un segment donné par ordre de plausibilité [HATO-81-a].

Notre système repose sur un modèle mixte mettant en jeu les deux processus fonctionnant en parallèle. Pour chaque segment dans le signal, le système fournit au plus trois phonèmes candidats.

Les deux groupes de phonèmes (/p/ /t/ /k/) et (/b/ /d/ /g/), présentant des traits acoustiques peu différents, ne peuvent être différenciés dans le niveau phonétique. L'ambiguïté est ensuite levée par les niveaux supérieurs de traitement.

Dans certains systèmes, le niveau phonétique fournit des traits et des classes de phonèmes, et c'est au niveau supérieur d'affiner ces réponses [DEMO-76-a], [DEMO-76-a], [CAEL-79], [CAEL-81].

II.4. LE NIVEAU PHONOLOGIQUE

La phonologie peut inclure des statistiques sur l'occurrence d'un phonème particulier ou d'une séquence de phonèmes ou des restrictions sur la combinaison des phonèmes dans les syllabes. Suivant les systèmes, la phonologie peut prendre en considération certaines règles. Par exemple, pour le système MYRTILLE II [PIER-81] le niveau phonologique se limite à l'étude des altérations possibles d'un phonème, ou d'un mot suivant son contexte à l'exclusion des aspects combinatoires de traits ou de phonèmes. La raison de cette définition restrictive de la phonologie réside dans le fait que c'est cet aspect d'altérations qui est le plus important en traitement automatique de la parole.

II.5. LE NIVEAU PROSODIQUE

La prosodie étudie les variations de hauteur du fondamental, de durée et d'intensité des éléments correspondant aux phonèmes qui se trouvent dans une séquence donnée [HATO-78-a], [CARB-83].

On peut distinguer :

1. Les marqueurs prosodiques délimitant des unités successives : par exemple dans la phrase :

(il va pleuvoir) (demain) (dans les vosges)

mais certaines pauses peuvent entraîner des erreurs par exemple les pauses d'hésitation ou de respiration. Ces marqueurs sont le plus souvent de type de pose ou schéma prosodique [MART-75].

2. La mélodie générale de la phrase.

La mélodie générale de la phrase permet en français de distinguer des phrases identiques, dont l'une serait énonciative, l'autre interrogative, la troisième impérative ou exclamative.

Une des raisons qui fait que l'utilisation de la prosodie est réduite dans les systèmes de reconnaissance automatique de la parole est le fait qu'une même phrase peut être prononcée de façon très différente d'un locuteur à l'autre et parfois pour le même locuteur.

II.6. LE NIVEAU LEXICAL

Ce niveau se décompose en deux parties complémentaires :

i) l'établissement du lexique, c'est-à-dire la représentation des mots et de leurs variantes phonologiques.

ii) la recherche lexicale, c'est-à-dire la recherche et la vérification d'un mot dans le treillis représentatif de la phrase.

II.7. LE NIVEAU SYNTAXIQUE

La syntaxe regroupe en général l'ensemble des informations contenues dans la grammaire [PIER-81].

Elle est utilisée soit pour sélectionner des mots proposés par l'analyse lexicale soit pour agir comme prédicteur et proposer une liste de mots qui peuvent se trouver dans la phrase prononcée.

Le choix d'un bon niveau de description syntaxique est prépondérant lors de la mise en oeuvre d'un système.

II.8. LE NIVEAU SEMANTIQUE

La sémantique est l'étude du langage considéré du point de vue du sens.

Elle a deux fonctions :

- supprimer ou confirmer certaines hypothèses, ce qui permet de restreindre le nombre de solutions à examiner.

- déterminer le sens du message dans le cadre de l'application considérée. Ceci a pour but soit de guider le dialogue, soit de provoquer l'action demandée par le locuteur, soit de guider l'analyse des réponses.

II.9. LE NIVEAU PRAGMATIQUE

Le niveau pragmatique étudie l'ensemble des informations sur le contexte de la conversation dans une tache donnée.

Remargue :

Le niveau sémantique et le niveau pragmatique insérés dans les systèmes de reconnaissance automatique de la parole sont strictement dépendants de l'application considérée, contrairement au niveau syntaxique qui peut être commun à plusieurs applications.

III. TRAITEMENT ACOUSTIQUE ET PARAMETRISATION DU SIGNAL VOCAL

III.1. CODAGE DE L'ONDE SONORE

L'analyse acoustique de la parole vise à extraire du signal un ensemble de paramètres qui doivent, d'une part, être aussi peu nombreux que possible, pour alléger les traitements futurs, et, d'autre part, condenser toute l'information linguistique disponible. La satisfaction de ces objectifs contradictoires exige que les variables retenues soient particulièrement pertinentes pour caractériser toutes les connaissances ayant une influence sur le signal de la parole.

Certaines analyses sont de types temporel et d'autres de types fréquentiel. Mais certaines méthodes ont des caractéristiques des deux types.

Les analyses de type temporel, ainsi que celles de type fréquentiel font appel aux techniques digitales. Elles requièrent, dans une première étape, une représentation digitale de l'onde sonore. Lorsque le signal est échantillonné à une fréquence F_N (théorème de Shannon) l'onde sonore doit être filtrée au dessus de $\frac{F_N}{2}$.

Les diverses méthodes utilisées pour une représentation paramétrique du signal seront très schématiquement envisagées dans ce paragraphe. On trouvera une description précise de certaines d'entre elles dans l'ouvrage de FLANAGAN [FLAN-72] et l'article de SCHAFER [SCHA-75]. Certaines techniques s'inspirent d'un modèle de la phonation ou de l'audition tandis que d'autres, plus générales, traitent le signal de parole en dehors des connaissances spécifiques à sa nature. Les paramètres sont parfois évalués directement à partir du signal mais plus fréquemment à l'aide d'une transformation spectrale à court terme.

III.2. ANALYSE TEMPORELLE

Plusieurs méthodes, s'appliquant à la représentation temporelle du signal, ont été proposées. Parfois, leurs résultats sont exprimés dans le plan fréquentiel. Cependant, cette manière d'aborder le problème permet de mettre en évidence les évènements courts et non stationnaires qui échappent à un examen spectral nécessitant une fenêtre temporelle importante.

La mesure de la densité des passages par zéro du signal et de sa dérivée a été l'une des premières méthodes d'analyse de la parole [PETE-51]. Par des moyens analogiques, amplification et écrêtage, le signal vocal est transformé en un signal carré, d'amplitude $\mp A$ et respectant les instants de passage par zéro du signal (figure 15).

A l'écoute du signal écrêté, il est possible de reconnaître la parole prononcée initialement. Ceci montre qu'une importante partie de l'information est conservée dans les passages par zéro.

Une contre-expérience a été effectuée par DUPEYRAT [DUPE-75]. Elle consiste à remplacer les points de passage par zéro par les milieux des passages par zéro d'origine. A l'écoute du signal en créneaux, ainsi constitué, est incompréhensible.

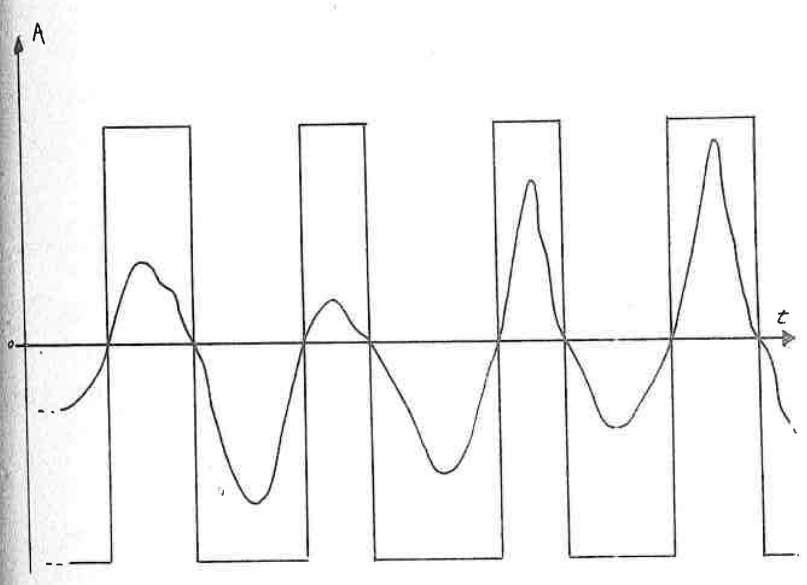
Dans le codage des passages par zéro, il n'est pas tenu compte des nombreuses fluctuations du signal qui n'entraînent pas de changement de signe. Il est donc préférable de faire porter la méthode sur la dérivée première du signal, ce qui revient à coder les temps des extrema (figure 16). Il en résulte une amélioration sensible de l'intelligibilité de la parole restituée. Pour les dérivées d'ordre supérieur à 1, la qualité se dégrade à nouveau.

La densité des passages par zéro est utilisée pour définir les deux premiers formants des sons vocaliques. Le premier formant est calculé par la quantité des passages par zéro du signal, tandis que le deuxième formant est calculé à partir de la quantité des passages par zéro de la dérivée du signal. Si la méthode est très rapide, elle ne permet cependant qu'une évaluation approximative des formants. Ce paramètre a été utilisé pour détecter certains phonèmes par exemple les consonnes /s/, /j/, /ʃ/, /z/. [HATO-81-a].

Une autre méthode utilisant la représentation temporelle du signal est le codage des extréma du signal proposée par Dupeyrat et Baudry [BAUD-72], [BAUD-74], [DUPE-75]. Elle est utilisée dans un système de segmentation et de reconnaissance de la parole, en temps réel.

La méthode proposée par BELLISSANT [BELL-78] travaillant avec le signal temporel est utilisée dans notre système, elle est détaillée dans le paragraphe V.

D'autres méthodes utilisant les paramètres temporels ont été employées par différents chercheurs [REDD-68], [DOUR-74], [PINS-63].



(AMPLITUDE EN FONCTION DU TEMPS)

Figure 15 - Expérience de Peterson et Marcou.

(AMPLITUDE EN FONCTION DU TEMPS)

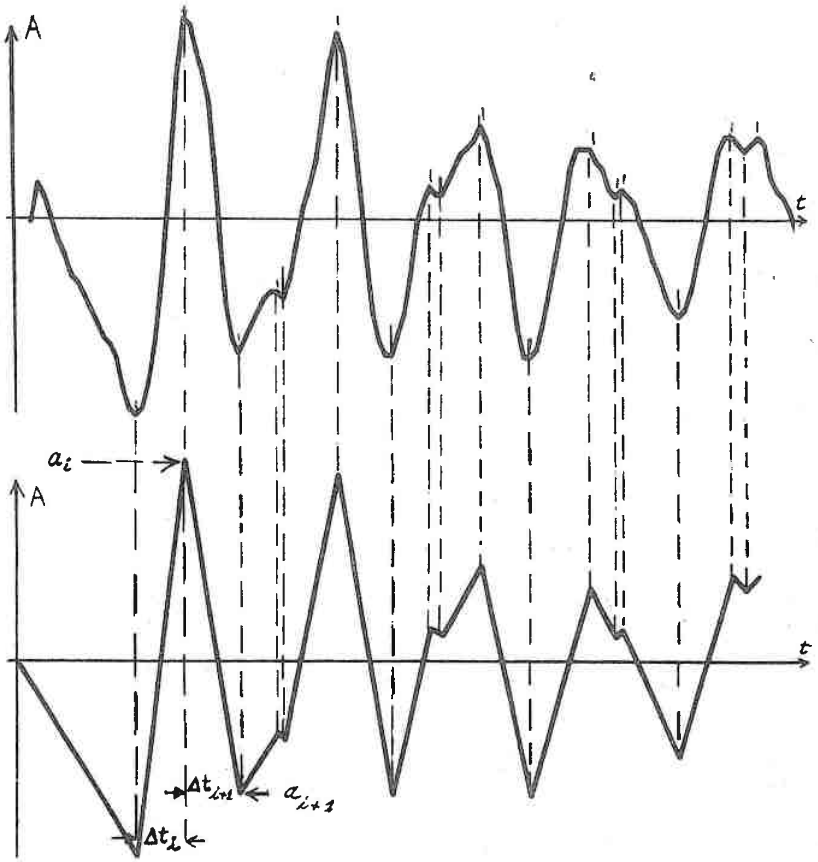


Figure 16 - Codage des extréma (bas) d'un signal origine (haut).

III.3. ANALYSE SPECTRALE

La représentation en fréquences du signal de la parole est la plus courante. Elle est avantageuse à deux points de vue : d'une part l'étude de l'audition montre que l'oreille effectue dans un premier temps une sorte d'analyse fréquentielle, d'autre part cette forme d'analyse permet une représentation du signal vocal assez fidèle.

Parmi les nombreuses méthodes d'analyse spectrale de la parole, nous citerons celles qui sont communément utilisées et qui présentent des caractéres spécifiques. Elles opèrent soit directement sur le signal (analogiques), soit à partir d'une transcription digitalisée.

L'analyse par transformée rapide de Fourrier [COOL-65] permet d'obtenir une représentation fréquentielle du signal aussi fine qu'on le souhaite pour un temps de calcul relativement peu élevé (il existe des opérateurs cablés réalisant cette transformation). Cependant, le spectre obtenu est modulé par la fréquence fondamentale ce qui rend difficile le problème de certains traitements en particulier la localisation des maxima (figure 17).

La séparation entre la source d'excitation et le canal vocal a conduit à l'élaboration d'une technique appelée CEPSTRE [SCHA-70], qui utilise deux transformées de Fourrier successives. La transformée inverse de Fourier (DFT) sur le spectre logarithmique obtenu directement à partir du signal par FFT permet d'isoler un pic dont la position correspond à la fréquence fondamentale. L'enveloppe spectrale du conduit vocal est obtenue par une nouvelle transformée de Fourier sur le spectre dont

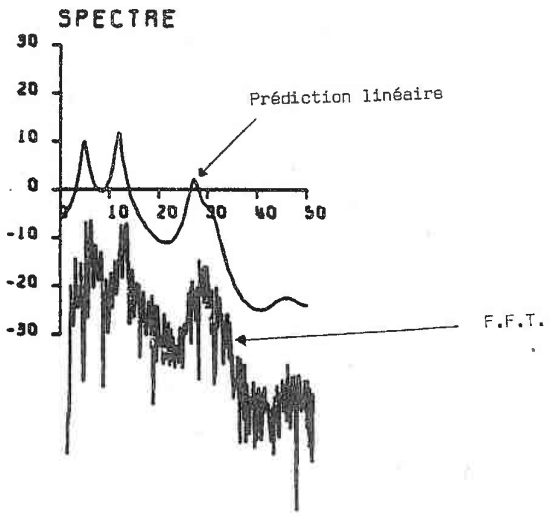


Figure 17

la partie haute a été tronquée, c'est-à-dire une mise à zéro du cepstre au delà des fréquences supérieures à une certaine valeur par exemple 200 Hz. Le spectre ainsi obtenu est lissé et fait apparaître nettement les formants (figure 17).

Les différentes étapes d'une analyse cepstrale sont schématisées sur la figure 18-a.

Une autre forme de spectre à court terme peut être fournie par une série de filtres passe-bande distribués sur les fréquences usuelles de la parole. L'ensemble des réponses des filtres lorsque le signal est appliqué à l'entrée constitue un spectre de puissance à court terme du

signal (principe de l'analyseur d'un vocoder).

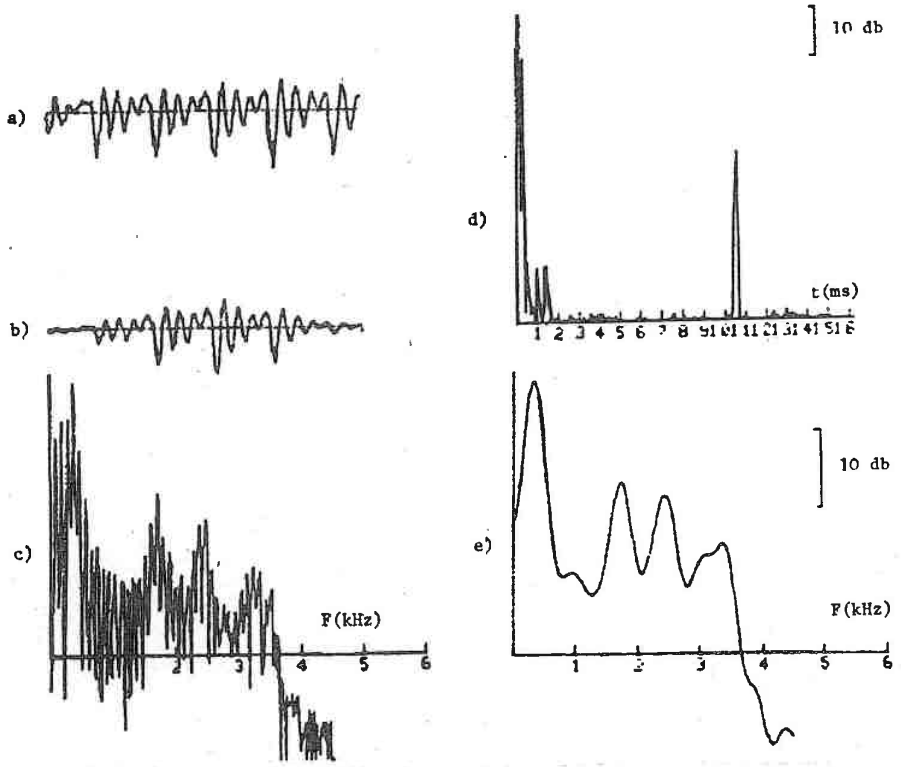
Une visualisation pratique des spectres à court terme est connue sous le nom de spectrogramme de parole ou "sonagramme".

En abscisse est porté le temps, en ordonnée la fréquence. L'amplitude relative à chaque fréquence est représentée par un noircissement plus ou moins prononcé (figure 19).

L'analyse par banc de filtres effectue une compression importante de la quantité d'information nécessaire au codage du signal.

C'est cette représentation qui est utilisée dans notre système, à partir d'un vocoder numérique à canaux, qui fournit l'énergie du signal dans les différentes bandes. La zone des fréquences traitées est située entre 200 et 6500 Hz (figure 20). Dans le système nous utilisons 16 canaux dont les sorties sont codées sur 7 bits. La fréquence d'échantillonnage est de 100 Hz.

Comme pour les sonagrammes, en abscisse est porté le temps, en ordonnée la fréquence. Mais l'amplitude relative à chaque fréquence est représentée par un code donnant l'intervalle où se trouve la valeur de l'amplitude (figure 21). Sur la figure 22 on a représenté le sonagramme du mot "CHAPEAU".



a) Signal original

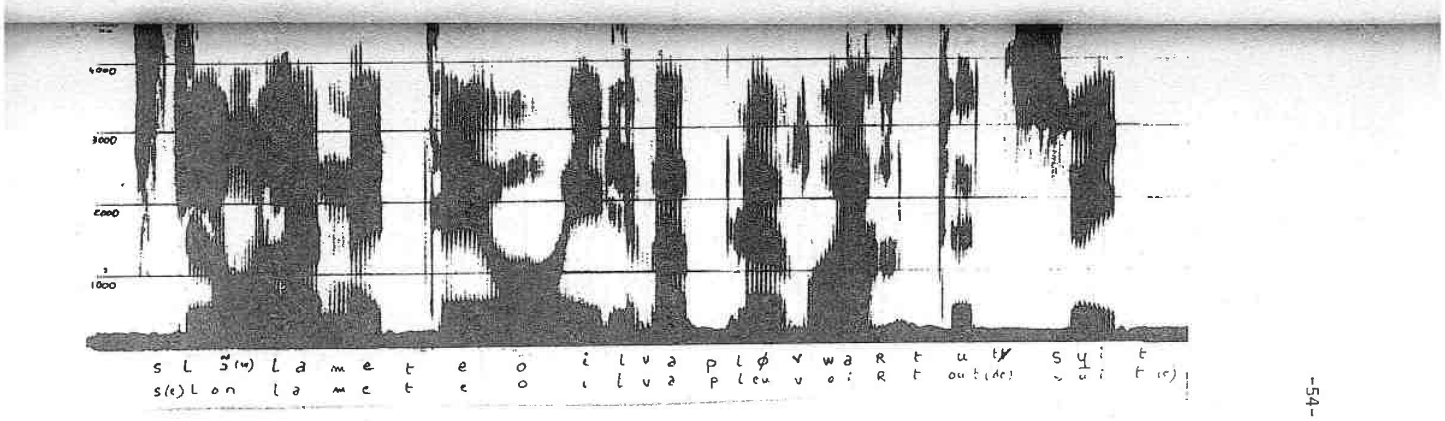
b) Après fenètre de Hamming

c) Spectre brut par FFT

d) Cepstre

e) Spectre lissé

Figure 18-a : Les étapes du Cepstre



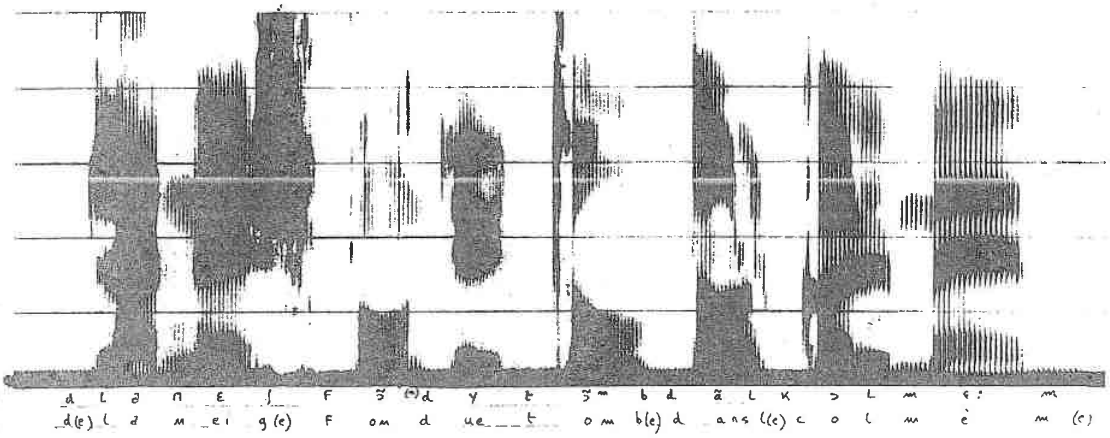


Figure 19

N° canal	bande passante	largeur de bande
1	350	200
2	550	200
3	750	200
4	950	200
5	1175	250
6	1450	300
7	1750	300
8	2050	300
9	2350	300
10	2650	300
11	2950	300
12	3300	400
13	3700	400
14	4100	400
15	4700	800
16	5900	1600

Répartition des 16 canaux de l'analyseur spectral

Figure 20

Code	valeurs d'énergie
—	0
.	16
:	32
+	48
=	64
c	80
3	96
%	112

Codes utilisés dans les sonagrammes

Figure 21

Une autre méthode d'analyse spectrale est l'analyse par prédiction linéaire. La méthode consiste à supposer qu'en l'absence de toute excitation, il existe une relation récurrente sur le signal vocal s(t) de la forme :

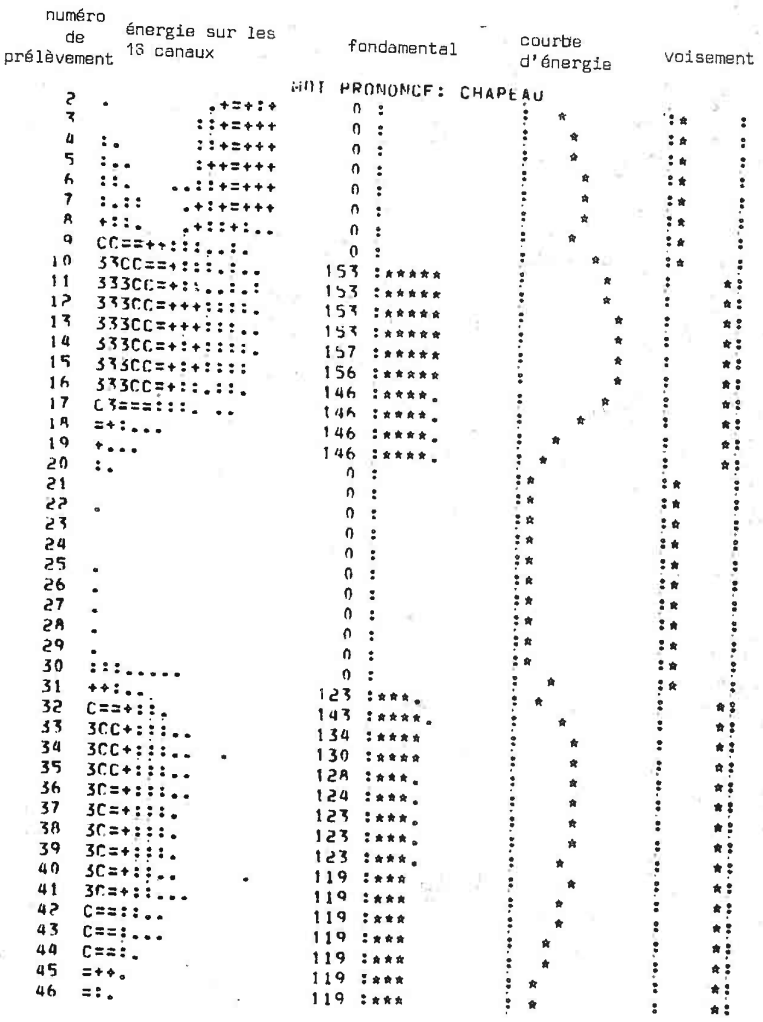
$$\hat{s}_n = - \sum_{i=1}^{M} a_i \, s_{n-i}$$

Cette relation permet de prédire la valeur de l'échantillon s_n à partir de la mesure des M valeurs précédentes si les a_i sont donnés.

L'erreur de prédiction est alors :

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

Les coefficients a_i, i = 1, ... M sont appelés coefficients de prédiction.



Sonagramme du mot "*chapeau*"
Figure 22

Il s'agit donc d'identifier la fonction de transfert du canal vocal sous forme d'un tuyau sonore à sections variables. Par une transformation de cette dernière nous obtenons un spectre lissé du signal et par suite les fréquences et les longueurs de bande des formants.

La figure 18-b schématise, en les explicitant, les résultats fournis par l'analyse de codage prédictif développé dans notre laboratoire.

Une autre forme de spectre à court terme peut être fournie par l'analyse de type impulsionnel.

III.4. COMPARAISON DES METHODES D'ANALYSE DU SIGNAL DE LA PAROLE

La comparaison des paramètres obtenus par diverses méthodes d'analyse acoustique du signal dépend de beaucoup de facteurs qui doivent être étudiés conjointement :

1. rapidité de la méthode d'analyse,

2. dépendance au locuteur,

3. sensibilité aux variations de l'environnement (dépendance du matériel d'acquisition utilisé, influence des bruits ambiants ...),

4. facilité d'extraction des paramètres,

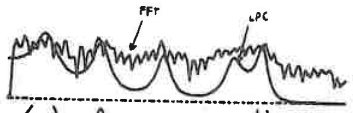
5. résistance aux variations temporelles de l'élocution (accélérations ou ralentissements).

6. place mémoire requise,

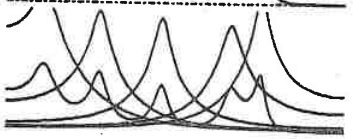
etc ...

signal temporel (voyelle /a/)

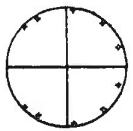
signal préaccentué (6 dB/octave)
+ fenêtre de Hamming



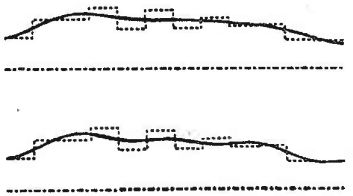
spectres de FFT et de prédiction linéaire



effet des différents pôles séparément



position des pôles dans le plan complexe



fonction d'aire du conduit vocal correspondante (en pointillé) et différents lissages de cette fonction

Figure 18-b

Traitement du signal vocal par prédiction linéaire
(méthode d'autocorrelation)
(d'après J.P. HATON [HATO-78-a])

Il n'existe pas d'étude comparative complète du comportement des principales méthodes d'analyse du signal de parole vis à vis de tous les facteurs précédemment énoncés.

Les recherches sur l'identification du locuteur ont apporté des connaissances sur les paramètres qui caractérisent le plus la voix d'un individu [WOL-72], [SAMB-75], [GREN-78]. Ces variables peuvent apporter des informations pour la reconnaissance dans la mesure où elles peuvent faire l'objet d'un apprentissage simple et qu'elles indiquent en complément celles qui sont moins marquées par des variations interlocuteurs importantes.

Quelques travaux de comparaison de plusieurs systèmes paramétriques ont été conduits dans la perspective de définir leurs performances pour segmenter, classer et identifier des unités phonétiques. Parmi ces travaux on peut citer celui de GOLDBERG [GOLD-75], de CASTELAZ [CAST-78] et de SAKAI [SAKA-79]. Sur la figure 23 les caractéristiques principales de quelques paramètres sont décrites d'après SAKAI [SAKA-79].

Features	Accuracy of extraction	Dimension	Contribution to recognition	Contribution to decoding	Cost-time
Zero-crossing	excellent	5 ∿ 10	fair	poor	excellent
Walsh-Hadamard transform	excellent	> 20	fair	fair	excellent
Spectrum (FFT)	excellent	> 20	fair	excellent	good
Autocorrelation function	excellent	5 ∿ 20	good	excellent	good
LPC	excellent	8 ∿ 12	poor	excellent	good
Parcor	excellent	8 ∿ 12	excellent	excellent	good
Cepstrum	excellent	8 ∿ 12	excellent	excellent	fair
Smoothed power spectrum	excellent	> 20	excellent	excellent	poor
Formant	good	3	excellent	good	poor
Vocal tract area function	fair	8 ∿ 12	excellent	excellent	poor
Position of lip, tongue, jaw	poor	3 ∿ 6	good	fair	poor

Caractéristiques principales de quelques paramètres d'après T. SAKAI

Figure 23

IV. LE DECODAGE ACOUSTICO-PHONETIQUE

IV.1. DESCRIPTION ET DIFFICULTES DU PROBLEME

Parmi les différents niveaux de traitement qui interviennent dans un système de compréhension automatique de la parole, le décodage acoustico-phonétique joue un rôle fondamental. Il se place en effet en frontal entre le signal vocal et les niveaux linguistiques de traitement ; par suite, de la qualité de la transcription phonétique obtenue, dépend en bonne partie la performance globale du système de compréhension.

La réalisation d'un système acoustico-phonétique se heurte à de nombreuses difficultés de segmentation du message vocal en unités phonétiques élémentaires et d'identification de ces unités du fait de la continuité du signal de la parole en fonction du temps, de la grande variabilité de la parole, de l'ambiguïté des limites entre les unités et le l'interaction entre unités voisines. On peut dire qu'une unité est composée de sa matière sonore intrinsèque à laquelle s'ajoutent la composante restante de l'unité précédente et celle anticipée de l'unité suivante. Ce problème est rendu encore plus difficile par le manque de critères stables d'un locuteur à un autre et même pour le même locuteur.

Le signal acoustique est conditionné par chacune des étapes de la communication vocale :

- réception
- transmission
- production.

Ces processus peuvent varier considérablement et leurs caractéristiques propres constituent les causes essentielles de la fluctuation des paramètres que l'on peut extraire du message.

IV.1.1. Variations dues au système de réception.

La traduction électrique du signal acoustique sera soumise aux limitations du récepteur. Une première modification intervient généralement dans l'élimination des hautes fréquences qui entraîne une altération en fonction des caractéristiques du filtre. Lors de la digitalisation du signal, la fréquence d'échantillonnage ainsi que le nombre de bits permettant de coder les valeurs vont conditionner la qualité de l'information restante. Les conséquences résultant du processus d'acquisition sont plus ou moins grandes suivant le type de voix à reconnaître, l'information perdue est plus importante dans le cas des voix aigues.

Les altérations introduites à ce niveau sont négligeables si le système a été bien pensé.

IV.1.2. Variations dues à la transmission

La qualité du support de la transmission et de ses organes (microphone, amplification, codage, ...) entraîne des distorsions plus ou moins sensibles du signal. Dans notre cas, ces problèmes sont tout à fait négligeables par rapport aux altérations résultants des bruits

ambiants. Ces derniers constituent dans la plupart des cas la source principale des modifications acoustiques du message. Ils sont parfois permanents et réguliers (ventilation, rotation des disques d'un calculateur ...) intermittents (claviers, imprimantes ...), aléatoires (portes, conversation ...) ou bien systématiques (réverbération). Nous tenons cependant à nous placer dans des conditons qui se rapprochent des conditions réelles d'utilisation d'un système d'entrées vocales.

IV.1.3. Variations dues à l'appareil phonatoire

Compte tenu de l'importance des différences physiologiques interlocuteurs des organes sollicités, on constate une personnalisation très nette de messages identiques prononcés par des individus cifférents. Au départ, la puissance de la voix est tributaire de la capacité pulmonaire et musculaire du locuteur. Ensuite, les possibilités de la source sonore sont liées à la longueur et l'épaisseur des cordes vocales, à la force de tension des muscles du larynx et à leur maîtrise (plus ou moins grande suivant les personnes). La dimension constante du conduit nasal contribue également à caractériser la voix. On constate une modification de la résonance des cavités nasales lors d'affectations particulières telles que les rhumes.

La contribution de la cavité buccale à l'élaboration des sons dépend non seulement de sa forme (17 cm de long et de section variant entre 0 et 20 cm^2 pour un homme adulte) mais surtout des possibilités

et du contrôle des articulateurs (lèvres, langue, joues, voile de palais).

IV.2. LES APPROCHES DE LA RECONNAISSANCE PHONETIQUE

Le module de reconnaissance et de segmentation de ces unités peut fonctionner selon l'un des deux principes suivants :

IV.2.1. Méthode asynchrone

Cette méthode fonctionne en deux étapes :

i) Segmentation de l'onde acoustique et localisation de la suite des unités à l'aide de paramètres temporels (nombre de passages par zéro du signal brut, nombre des passages par zéro de la dérivée du signal, l'énergie absolue, énergie quadratique ...) et spectraux (fréquence fondamentale, formants ...).

ii) Identification des unités par des méthodes relevant à la fois de la phonétique et de la reconnaissance des formes.

Parmi les travaux qui ont été réalisés dans ce cadre on peut citer ceux de J.P. HATON [HATO-81-a], [LAZR-81], [HATO-79], [SANC-78], C. BELLISSANT [BEll-78] [BELL-78-b], M. BAUDRY [BAUD-72], [BAUD-78], B. DUPEYRAT [DUPE-75], G. MERCIER [MERC-82] etc ...

IV.2.2. Méthode synchrone (appelée aussi centiseconde)

Cette méthode consiste à identifier des tranches du signal vocal de durée constante, variable selon les systèmes, puis à regrouper des segments identiques. Pour augmenter la fiabilité des chaînes de segments, on donne le plus souvent plusieurs étiquettes possibles à un segment, donné par ordre de plausibilité décroissante.

Parmi les travaux qui ont été réalisés dans cette approche on peut citer ceux de J.P. HATON [HATO-81-a], L. MICLET [MICL-81], F. JELINEK [JELI-81] etc ...

L'un des inconvénients de la méthode asynchrone est les erreurs de reconnaissance dues à une mauvaise segmentation. Ce problème provenant de la segmentation nous a conduit à utiliser simultanément la méthode synchrone et asynchrone.

Notre système repose donc sur un modèle mixte mettant en jeu des processus complémentaires fonctionnant en parallèle :

- une identification synchrone des spectres à court terme successifs fournis par le vocoder à canaux, puis identification phonémique par décision majoritaire.
- une localisation des noyaux vocaliques et identification des segments.

Notre système représente l'aboutissement actuel des différentes études menées par notre équipe dans le domaine :

- une première étude avait conduit à la conception d'un système fondé sur une segmentation phonémique du signal vocal. Ce système très souple permettait d'intégrer facilement de nouveaux algorithmes de reconnaissance [HATO-79],

- un système fondé sur une approche purement centiseconde [JELI-81] a ensuite été réalisé dans le cadre d'un système de reconnaissance analytique de grands vocabulaires [HATO-81-a].

Ces études avaient montré l'intérêt de conforter les résultats d'un tel modèle synchrone par une étude des variations de l'énergie.

- depuis 1982, un groupe de travail, dans notre laboratoire, est mis en place afin de réaliser un système expert de décodage phonétique [CARB-82]. Ce système est destiné à intégrer l'expertise mise en oeuvre par le phonéticien lors de la lecture de sonagrammes. Ce travail est destiné à fournir à long terme le cadre de système de décodage performant. A plus court terme, les retombées sur les systèmes existants sont loin d'être négligeables. Ainsi, nous avons intégré dans le système décrit ici des résultats concernant la détermination et la segmentation des noyaux syllabiques directement issus des règles de l'expert.

IV.3. UNITES DE RECONNAISSANCE

La segmentation est un des problèmes majeurs de la reconnaissance de la parole, en particulier de la parole continue. La première difficulté consiste à définir l'unité de base : mot, syllabe, diphone, phone, phonème, chacune d'elles présentant des avantages et des inconvénients [MASS-75], [KLAT-80], [VAIS-80], [SHOU-80].

Un des avantages de prendre les phonèmes comme unité de base réside dans le fait que leur nombre est assez restreint et que les mots et les phrases peuvent être considérés comme une suite de phonèmes. Par contre, la correspondance entre les phonèmes et leurs caractéristiques acoustiques et articulatoires n'est pas bien définie.

Notre système se fonde sur le phonème en tant qu'unité de reconnaissance sans pour autant négliger la notion de noyau syllabique tant dans la phase de segmentation que d'identification.

CHAPITRE IV

DESCRIPTION GENERALE DU SYSTEME

CHAPITRE IV

DESCRIPTION GENERALE DU SYSTEME

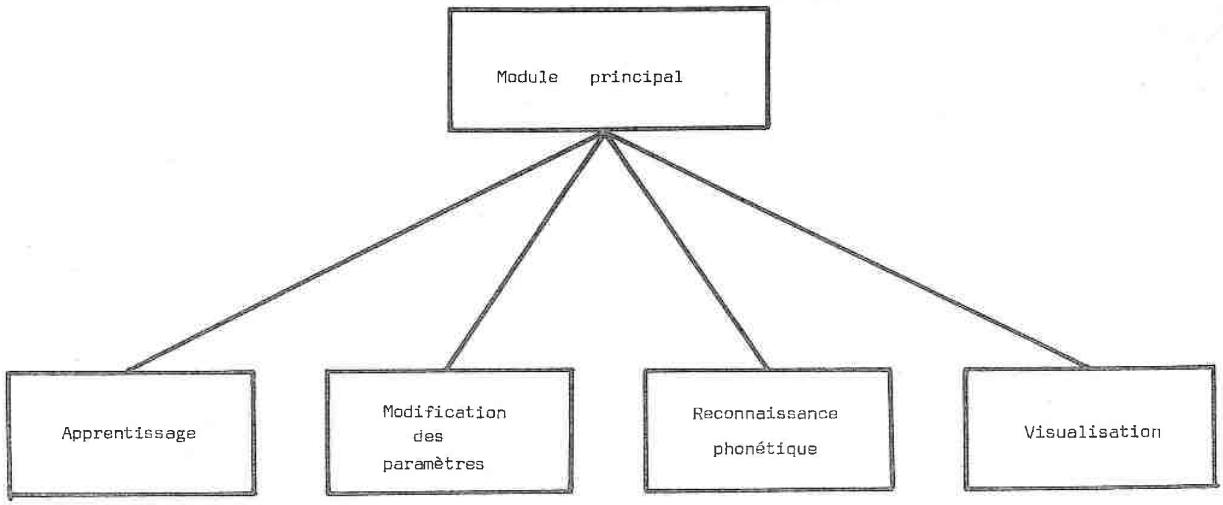
I. ARCHITECTURE GENERALE DU SYSTEME

Le système acoustico-phonétique que nous avons réalisé se compose de quatre modules (figure 24) :

1. module d'apprentissage phonétique.
2. Module de modification des paramètres du système.
3. Module de reconnaissance phonétique.
4. Module de visualisations.

Les modules d'apprentissage et de reconnaissance utilisent les mêmes programmes de prétraitement du signal (paragraphe II.2.).

Notre système, pour l'instant, est implanté en Fortran sur un mini-calculateur Mitra 125.



Organisation générale du système

Figure 24

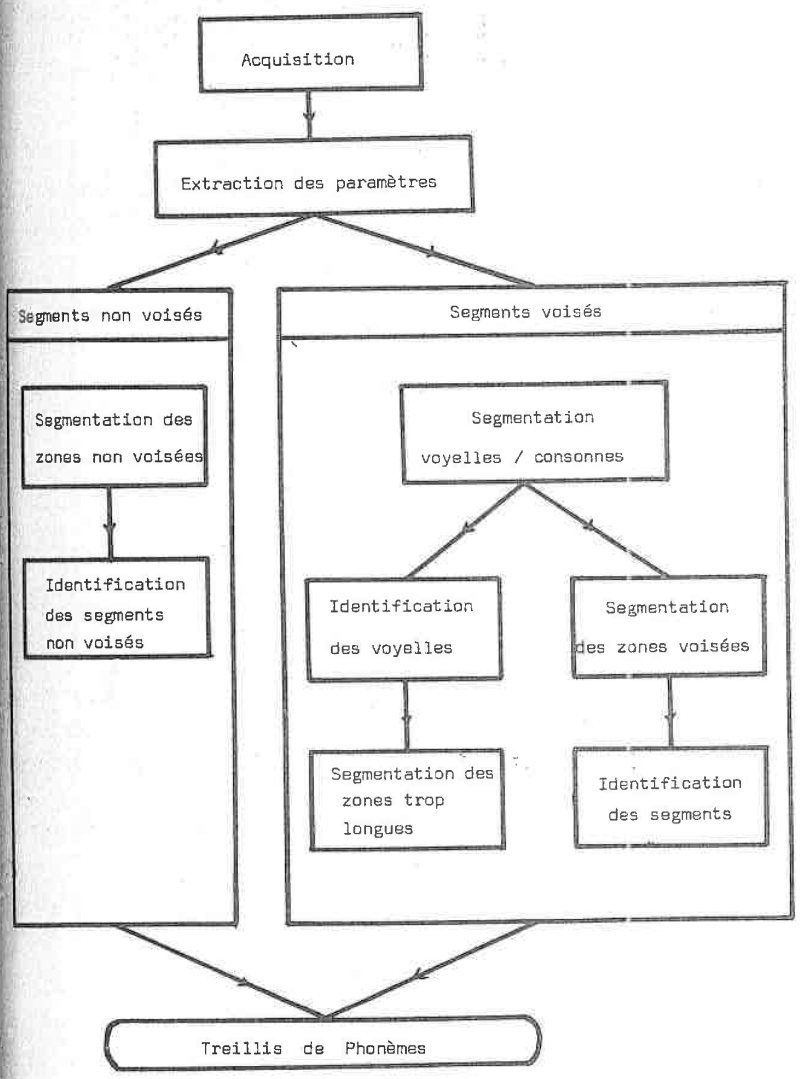
II. ORGANISATION DU MODULE DE RECONNAISSANCE

II.1. LES ETAPES DE LA RECONNAISSANCE PHONETIQUE :

Le système acoustico-phonétique est constitué d'un ensemble de modules correspondant chacun à une étape importante de l'analyse : traitement acoustique du signal, extraction des paramètres, reconnaissance des voyelles, élimination des zones de transition en début et en fin des voyelles, reconnaissance des consonnes non voisées, reconnaissance des consonnes voisées ....

Chaque module a pour fonction de faire progresser la reconnaissance phonétique à partir des hypothèses émises à l'étape précédente. A partir du module qui segmente le signal en zones voisées et non voisées nous appelons les modules correspondant à chaque type de la zone. Pour les segments voisés, si un segment est considéré, par le module de reconnaissance des voyelles, comme n'appartenant pas à la classe des voyelles alors nous appelons le module de reconnaissance des consonnes voisées. Nous verrons d'autres traitements dans les chapitres qui suivent.

Le schéma général du système est représenté sur la figure 25. Pour rendre ce schéma plus clair nous avons séparé quelques modules comme pour les modules "segmentations voyelles / consonnes" et "identification des voyelles" en effet ce dernier nécessite les étiquettes données, par le premier module à chaque prélèvement du segment, donc en réalité ces deux modules se déroulent en parallèle.



Organisation générale du module de reconnaissance

Figure 25

La procédure adoptée pour le système de reconnaissance phonétique consiste à appliquer une hiérarchie de règles logiques et un ensemble de fonctions de décision permettant de distinguer les classes des phonèmes. Les paramètres des fonctions de décision et les seuils sont, soit calculés par apprentissage, soit normalisés suivant les valeurs de certains paramètres.

Le système de reconnaissance phonétique procède par plusieurs étapes successives qui peuvent être groupées en six (figure 26) :

1. Acquisition.
2. Extraction des paramètres.
3. Reconnaissance des voyelles par la méthode spectrale.
4. Reconnaissance des voyelles par la courbe d'énergie
5. Reconnaissance des consonnes voisées.
6. Reconnaissance des consonnes non voisées.

Les modules de reconnaissance utilisent deux types d'informations :

- les références des phonèmes obtenus à la phase d'apprentissage,
- les traits acoustiques, calculés soit dans le module d'extraction des paramètres, soit au moment de leur utilisation.

Il y a donc trois types de références :

1. les références des voyelles
2. les références des consonnes voisées
3. les références des consonnes non voisées.

Dans le paragraphe III, nous allons montrer comment ces références sont calculées.

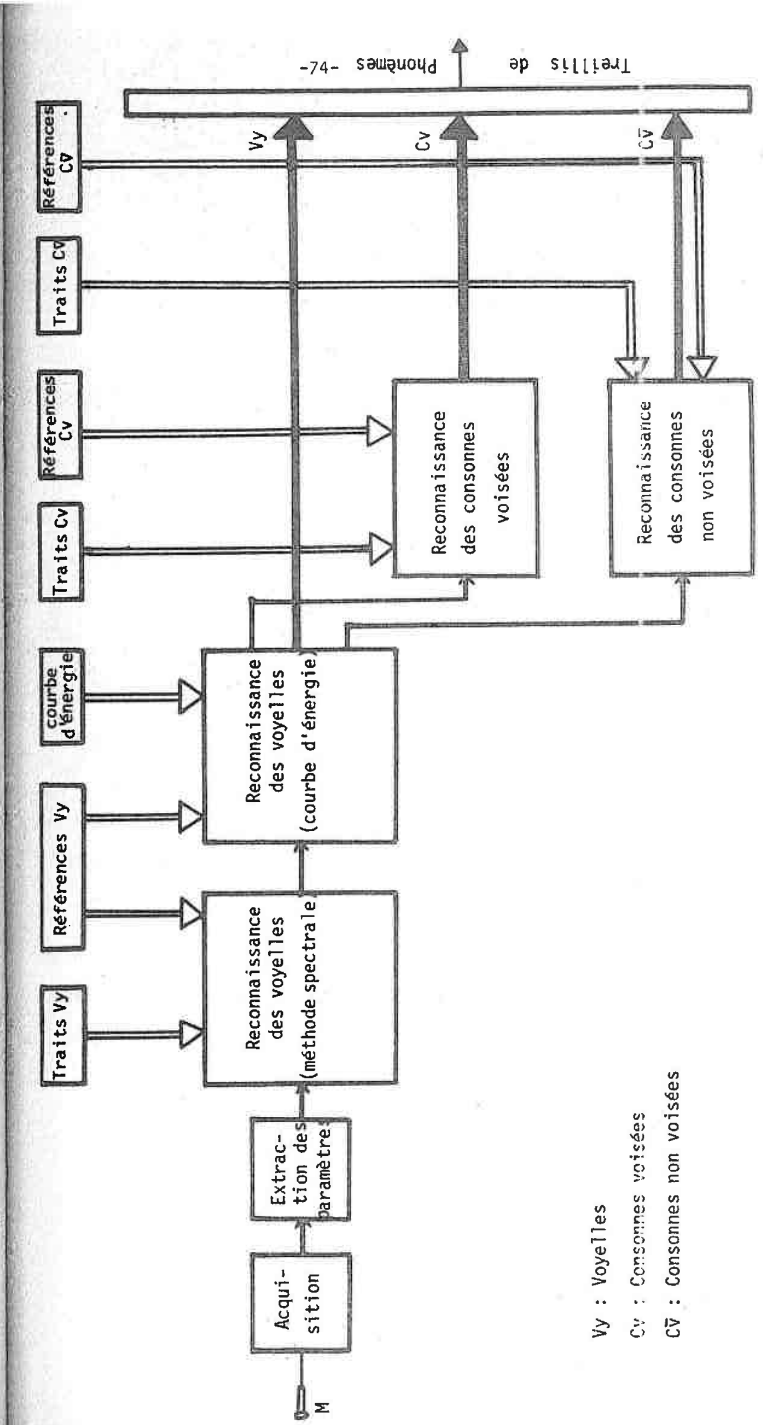


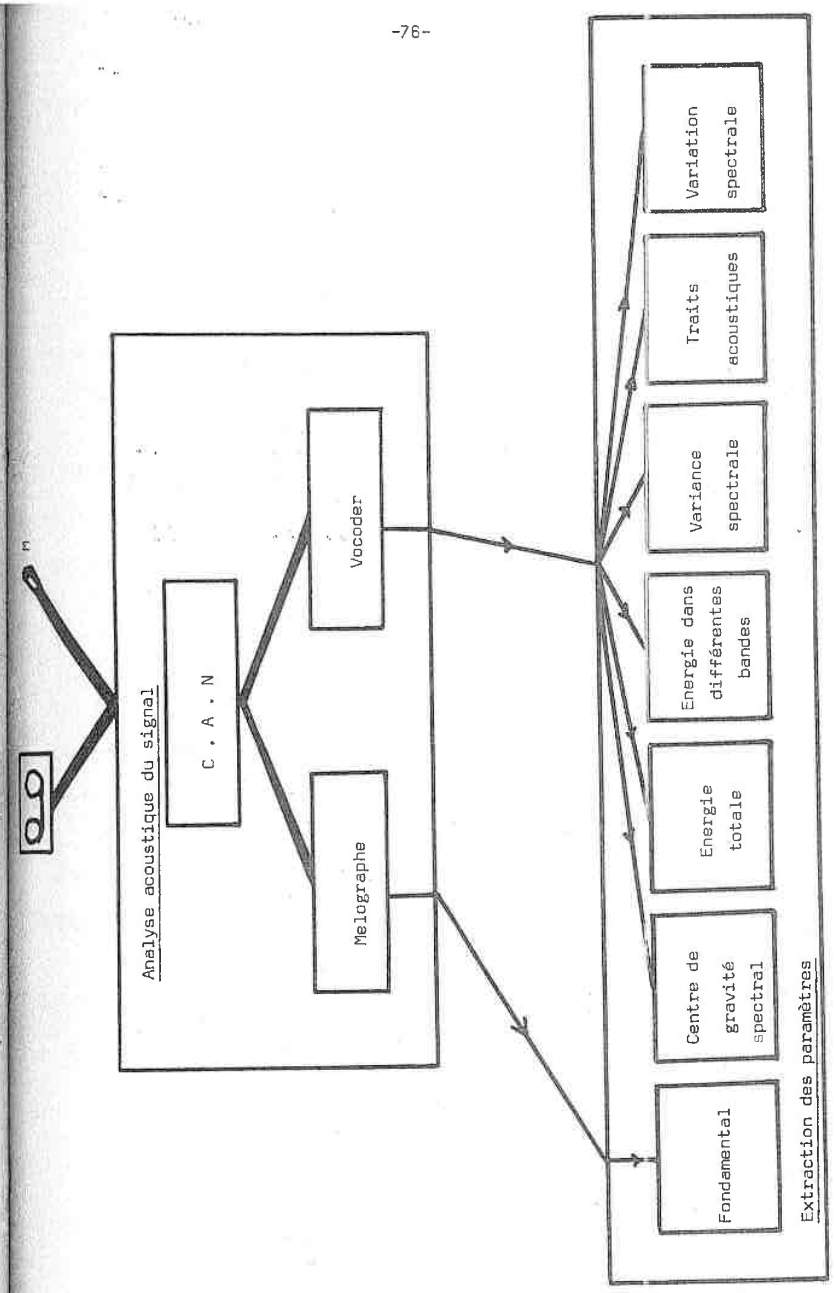
Figure 26 : Schéma général du système de reconnaissance phonétique

II.2. ANALYSE ACOUSTIQUE ET EXTRACTION DES PARAMETRES DU SIGNAL DE LA PAROLE

La représentation utilisée dans le système est une représentation spectrale obtenue à la sortie d'un vocodeur numérique à canaux dont le nombre et la répartition des filtres sont programmables. Dans le système nous prenons 16 canaux couvrant la bande 200 - 6500 Hz. Pour calculer le fondamental on utilise un mélographe, fonctionnant en temps réel, à l'aide de ce paramètre nous pouvons diviser les phonèmes en deux classes, les phonèmes voisés et les phonèmes non voisés.

En plus des valeurs de l'énergie des 16 canaux du vocodeur, on calcule d'autres paramètres de base (figure 27) l'énergie totale toutes les centisecondes, le centre de gravité fréquentiel, les énergies dans différentes bandes de fréquence, la variation d'énergie entre échantillons successifs, et un certain nombre d'indices, fondés sur la quantité d'énergie localisée dans cinq bandes particulières du spectre : grave / aigu, fermé / ouvert, compact / écarté, doux / strident, bémolisé / diésé.

Ces différents paramètres et indices sont utilisés aussi bien pour la segmentation que pour l'identification des phonèmes. Dans le paragraphe suivant nous allons voir comment nous calculons ces différents indices. Les autres paramètres seront explicités dans les chapitres qui suivent.



Analyse acoustique et extraction des paramètres du signal de la parole - Figure 27

II.3. DESCRIPTION DES TRAITS ACOUSTIQUES

Dans notre système, nous utilisons, en plus des 16 paramètres fournis par le vocoder, des indices acoustiques non formantiques, fondés sur la quantité d'énergie localisée dans cinq bandes particulières du spectre [CAEL-79], [CAEL-81].

Ces indices sont appelés :

- grave / aigu
- fermé / ouvert
- compact / écarté
- doux / strident
- bémolisé / diésé.

Ces indices ne sont pas calculés à partir des formants, puisque la détection des formants et leur suivi en phase consonantique posent d'énormes problèmes.

L'intérêt majeur de ces cinq indices réside dans la facilité des calculs et surtout le fait qu'ils ne nécessitent pas le calcul des formants.

Pour calculer ces différents traits nous définissons cinq bandes où la probabilité de trouver des formants (ou formants de bruit) est maximale (figure 28).

La première bande, B1, est définie de 150 à 400 Hz, elle correspond au fondamental Fo.

La deuxième bande, B2, est définie de 250 à 900 Hz, elle correspond au premier formant F1.

La troisième bande, B3, est définie de 800 à 3 500 Hz, elle correspond au deuxième formant F2.

La quatrième bande, B4, est définie de 2 000 à 4 200 Hz, elle correspond au troisième formant F3.

La cinquième bande, B5, est définie de 2 000 à 6 000 Hz, elle correspond au formant de bruit FS.

Les différents indices sont définis à partir de ces cinq bandes de la façon suivante (figure 28) :

grave / aigu = GA = Energie (B1) - Energie (B5 $\setminus$ B3)

fermé / ouvert = FO = Energie (B2 sup) - Energie (B2 inf)

bémolisé / diésé = BD = Energie (B3 sup) - Energie (B3 inf)

doux / strident = DS = Energie (B5 sup) - Energie (B5 inf)

compact / écarté = CE = Energie (B1) + Energie (B2) -
2 * Energie (B1 $\wedge$ B2)

Pour notre vocoder à 16 canaux les indices sont calculés de la façon suivante :

GA = Canal 51) - [Canal (13) + Canal (14) + Canal (15) + Canal (16)]

FØ = Canal (3) - Canal (1)

BD = Canal (10) + Canal (11) - Canal (5)

DS = Canal (15) - Canal (9) - Canal (10)

CE = Canal (1) + Canal (2) + Canal (6) + Canal (7) + Canal (8) +
Canal (9) - 2 * Canal (4)

Sur la figure 29, nous avons visualisé les courbes de variation de ces différents indices pour le mot *"je voudrais"*, la phrase *"Monsieur Dupont"* est représentée sur la figure 30.

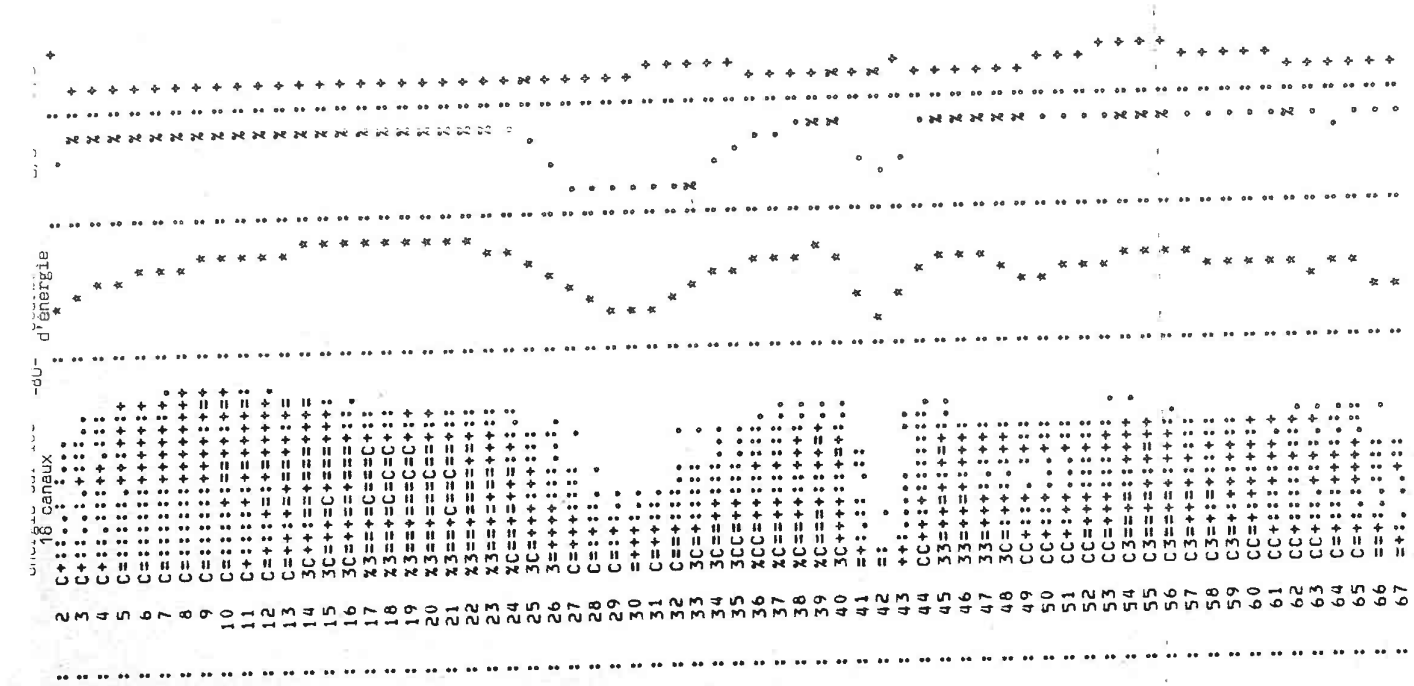


Figure 29

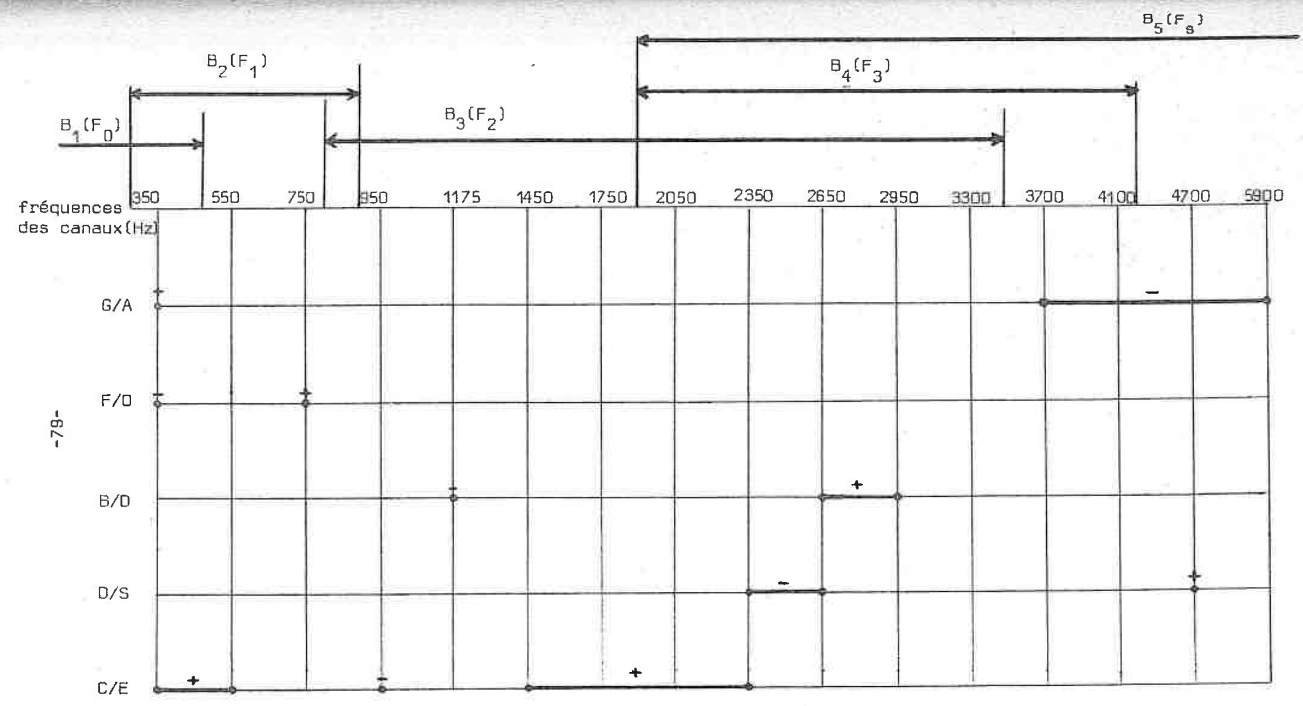


Figure 28 : Définition des traits acoustiques

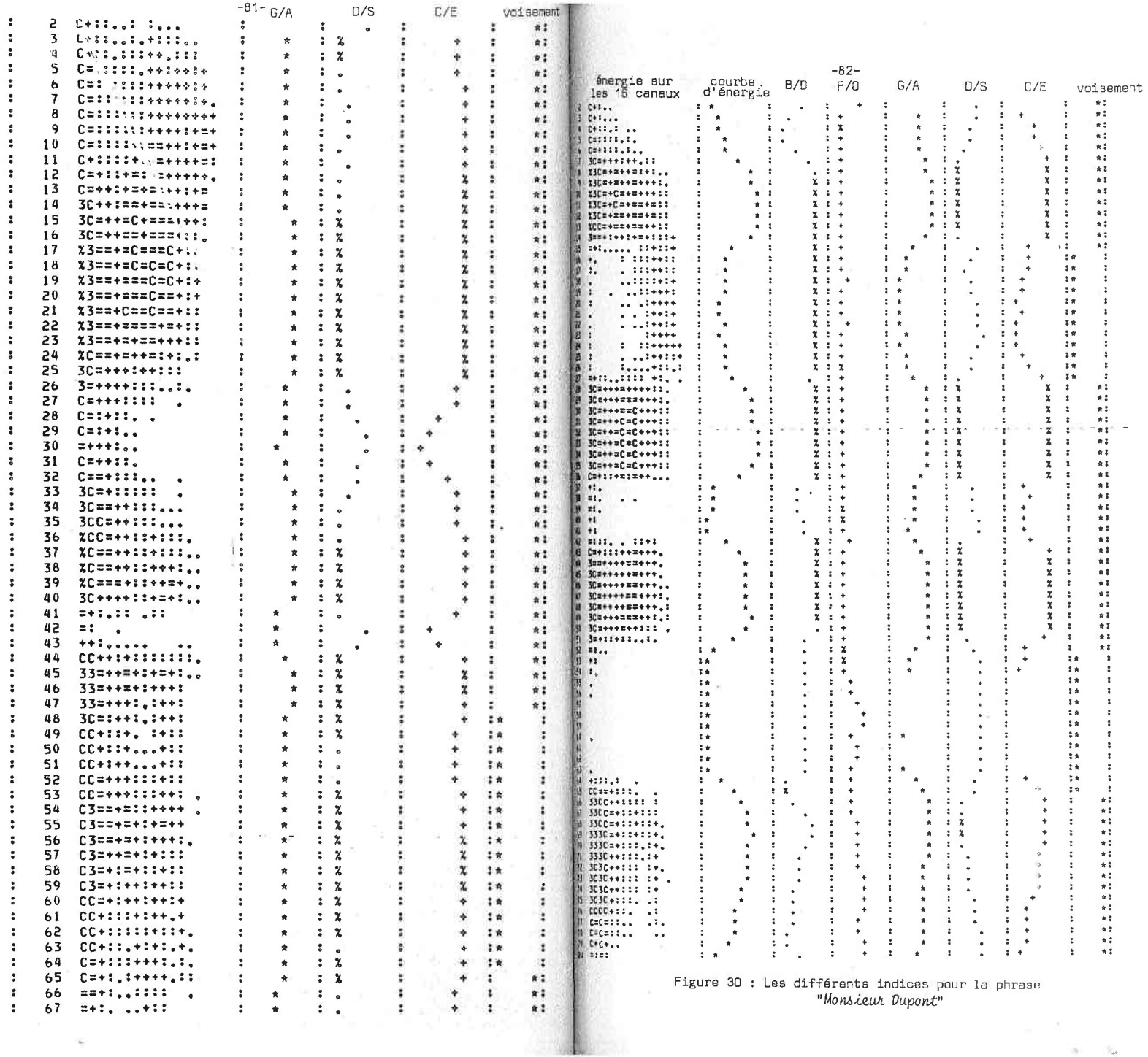


Figure 30 : Les différents indices pour la phrase "Monsieur Dupont"

Chaque prélèvement de 10 ms est représenté par un vecteur de 21 éléments dont les 16 premiers sont les valeurs de l'énergie dans les différents canaux du vocodeur et les 5 derniers correspondant aux valeurs des différents indices acoustiques (figure 31).

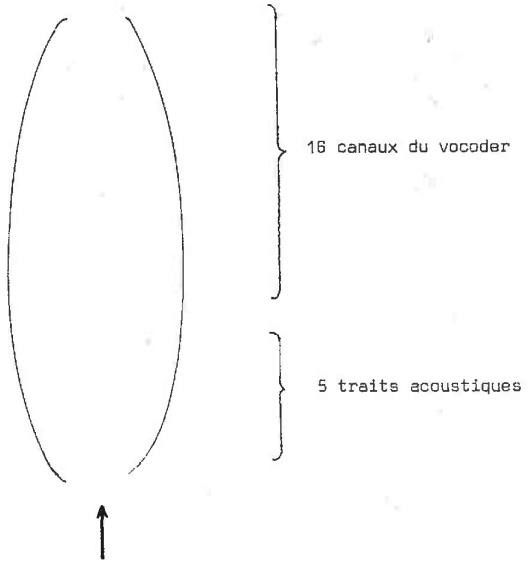


Figure 31

II.4. MODELES DE RECONNAISSANCE DANS LE SYSTEME

Comme nous l'avons signalé, le modèle de reconnaissance adopté dans notre système est de type centiseconde, c'est-à-dire, nous affectons une étiquette phonémique à chaque prélèvement de parole de 10 ms, puis nous regroupons les échantillons identiques (segmentation par identification), l'identification phonémique est faite par une décision majoritaire. Cette méthode fonctionne bien pour les échantillons des segments stationnaires, par contre elle n'est pas adaptée aux zones de transition entre deux segments consécutifs. C'est pour cela que nous avons introduit une fonction de segmentation qui nous permet de ne pas tenir compte des zones transitoires entre deux phonèmes consécutifs.

En plus du modèle centiseconde nous avons intégré dans le système des résultats concernant la détection des noyaux vocaliques directement issus du système expert de décodage phonémique en cours d'étude.

III. L'APPRENTISSAGE

La phase d'apprentissage consiste à créer et enrichir un dictionnaire phonémique dont chaque élément représente un phonème.

Les caractéristiques retenues pour chaque forme de référence sont les valeurs des 16 canaux du vocodeur ainsi que les valeurs des 5 indices, décrites dans le paragraphe précédent. Ces 21 composantes décrivent pour chaque intervalle de 10 millisecondes un point X dans $\mathbb{N}^{21}$.

L'ensemble de ces points constitue un nuage de $\mathbb{N}^{21}$ dont le centre de gravité Ω a pour coordonnées :

$$w_k = \sum_{i=1}^{N} \frac{x_{i,k}}{N}, \quad k = 1, \ldots 21$$

où $\Omega = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{21} \end{pmatrix}$

x_k est la $k^{ième}$ composante de X

et N est le nombre d'échantillons ayant servi à l'apprentissage pour le phonème considéré.

Dans les paragraphes précédents, nous avons vu que le système procède en trois phases

- la reconnaissance des voyelles
- la reconnaissance des consonnes voisées
- la reconnaissance des consonnes non voisées

pour cela nous utilisons trois types de dictionnaires

- dictionnaire des voyelles
- dictionnaire des consonnes voisées
- dictionnaire des consonnes non voisées.

Chaque phonème à une forme de référence dans le dictionnaire correspondant sous fome d'un vecteur de 21 éléments $(x_1, \ldots, x_{21})$. Nous avons vu au chapitre II que les spectres des phonèmes varient suivant le contexte, c'est pour cela qu'on met, en moyenne, deux références pour chaque phonème.

Le rôle de l'apprentissage est donc de calculer pour chaque phonème les composantes dans $\mathbb{N}^{21}$ des vecteurs moyens (ou centre de gravité) du nuage phonémique.

Nous avons environ 30 phones dans le dictionnaire des voyelles, 7 dans le dictionnaire des consonnes non voisées et 12 dans le dictionnaire des consonnes voisées, l'augmentation du nombre des références n'améliore pas beaucoup les résultats de la reconnaissance.

Actuellement, la phase d'apprentissage est faite manuellement, c'est-à-dire, nous indiquons au système le début et la fin du segment considéré. Une amélioration de ce module est en cours de réalisation, au CRIN, par K. BERRADA dans le cadre de son stage de D.E.A.

IV. MODIFICATION DES PARAMETRES

Nous utilisons, dans le système, deux types de paramètres :

- les paramètres calculés par apprentissage et
- les paramètres normalisés suivant certains critères

Ce deuxième type de paramètres est calculé au moment de son utilisation donc les valeurs des paramètres changent à chaque fois. Tandis que le premier type est initialisé au début du système. Ces paramètres sont liés aux locuteurs et doivent être adaptés pour un nouveau locuteur. Parmi ces paramètres on peut citer le seuil d'énergie au dessous duquel le prélèvement est considéré comme appartenant à la classe

des consonnes voisées. Un autre seuil important est celui de l'énergie dans les hautes fréquences au dessus duquel on étiquette le prélèvement par /j/ ou /ch/, suivant la valeur du fondamental. Pour adapter automatiquement ces paramètres à un nouveau locuteur, un projet est en cours de réalisation, au CRIN, par C. PISTER dans le cadre de sa thèse de troisième cycle, ce travail vise aussi l'adaptation des formes des références pour un nouveau locuteur.

V. VISUALISATION

Les résultats de la reconnaissance, les étapes intermédiaires, les valeurs des différents paramètres, le sonagramme de la phrase prononcée, le sonagramme avec les emplacements des noyaux vocaliques ainsi qu'un certain nombre d'informations concernant le déroulement de la reconnaissance, peuvent être visualisés à l'aide de ce module.

CHAPITRE V

RECONNAISSANCE DES VOYELLES

CHAPITRE V

RECONNAISSANCE DES VOYELLES

I. RECONNAISSANCE DES VOYELLES PAR LA METHODE SPECTRALE

I.1. DETECTION DES VOYELLES DANS LE SIGNAL

I.1.1. Segmentation du signal en deux classes

Les phonèmes voisés sont séparés des phonèmes sourds à partir des informations fournies par le mélographe. Dans ce paragraphe, nous allons nous intéresser aux phonèmes voisés.

On distingue dans les phonèmes voisés cinq classes :

- voyelles (orales et nasales)
- liquides
- nasales
- fricatives voisées
- plosives voisées

que l'on peut réduire en deux :

- voyelles
- consonnes voisées.

Etant donnée une suite d'échantillons voisés, il s'agit dans une première étape de détecter les voyelles et les consonnes voisées. A cette

fin, nous allons procéder de la façon suivante : chaque prélèvement voisé est comparé aux références des voyelles (orales et nasales), à l'aide de la distance de Hamming.

$$D_j = \sum_{i=1}^{NB} | C(k, i) - R(j, i) |$$

où NB : nombre de canaux

k : numéro du prélèvement courant

C (k, i) : valeur du $i^{ème}$ canal du prélèvement

j : numéro de la référence dans le dictionnaire

R (j, i) : valeur du $i^{ème}$ canal de la référence.

Lorsqu'un échantillon est très proche d'une de ces références, on en conclut qu'il appartient à la classe des voyelles. Par contre, lorsque cet échantillon ne correspond à aucune référence des voyelles, on en déduit son appartenance à la classe des consonnes voisées.

Cette correspondance se fait à l'aide d'un taux de ressemblance, qui est égal à la plus petite distance entre le prélèvement à reconnaître et toutes les références des voyelles :

$$D1 = \text{MIN}_{j = 1, \ldots, NBR} D_j$$

où NBR est le nombre de références des voyelles.

Pour qu'un prélèvement appartienne à la classe des voyelles, il faut que D1 soit inférieur à un seuil S appelé seuil de segmentation. Si cette condition n'est pas remplie, on en conclut que le prélèvement appartient à la classe des consonnes voisées.

Si après une suite d'échantillons appartenant à la classe des voyelles, nous trouvons un échantillon appartenant à la classe des consonnes voisées, la segmentation est alors effectuée. Cette méthode rend implicite cette dernière : c'est une segmentation par identification.

I.1.2. Etiquetage des prélèvements des segments appartenant à la classe des voyelles

Dans le cas où la plus petite distance D1, entre le prélèvement courant et les références des voyelles, est inférieure à un seuil, nous en déduisons que cet échantillon appartient à la classe des voyelles, et nous lui affectons les trois phones les plus proches.

En première position, nous lui affectons le phone correspondant à D1, puis en deuxième et troisième positions les phones correspondant à D2 et D3, avec :

$$D2 = \text{MIN } D_j \quad \text{avec} \quad D2 \neq D1$$

et

$$D3 = \text{MIN } D_j \quad \text{avec} \quad D3 \neq D2 \neq D1$$

$j = 1, \ldots$, nombre de références des voyelles.

Sur la figure 32 nous avons représenté le sonagramme du mot /b a b a/. Les résultats de la segmentation et d'étiquetage sont indiqués sur la figure 33. Le type des prélèvements est indiqué sur la dernière colonne de cette figure : le chiffre zéro signifie que le prélèvement n'appartient pas à la classe des voyelles ; par contre le chiffre deux indique qu'il appartient à la classe des voyelles. Dans cet exemple, le système a détecté deux voyelles, la première est située entre les prélèvements 13 et 25,

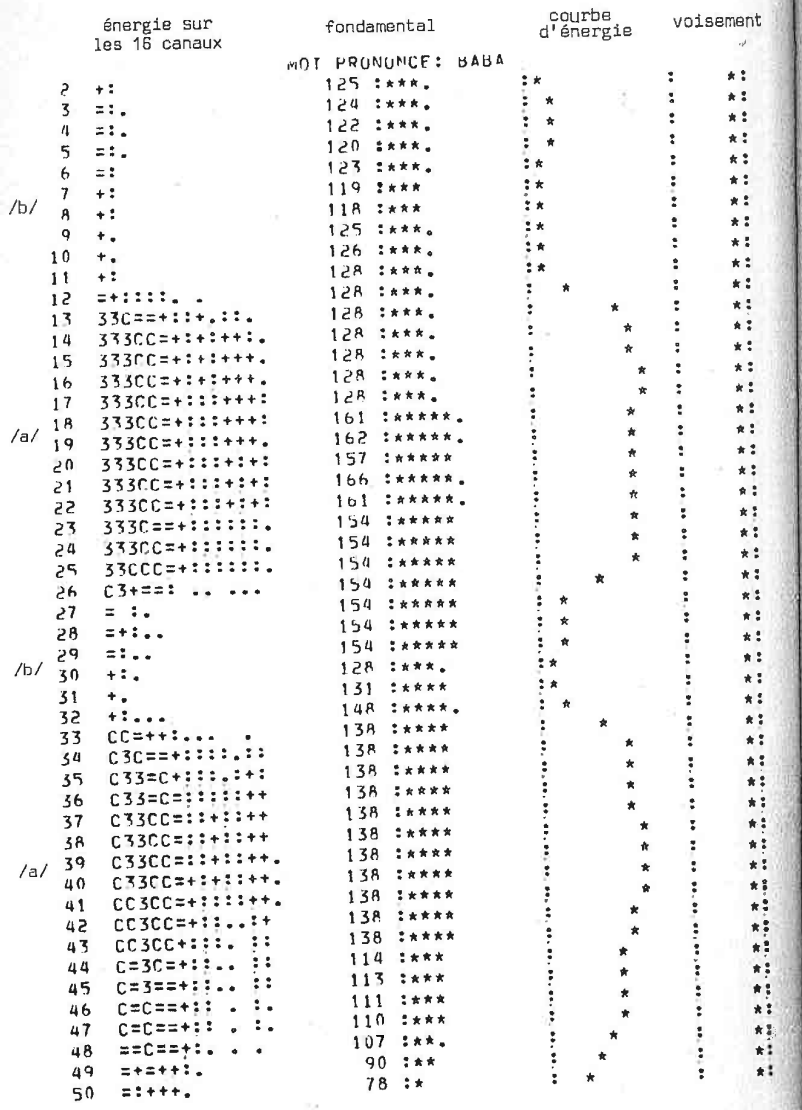


Figure 32 : Sonagramme du mot "BABA"

	ENER	TYP	PITC	VOI	VSPE	DIST1	PHO1	DIST2	PHO2	DIST3	PHO3	
2:	111:	5	125:	1	0:	794	ON	877	AN	915	OU	0
3:	129:	5	124:	1	8:	667	ON	738	OU	748	AN	0
4:	137:	5	122:	1	10:	651	ON	722	OU	736	OU	0
5:	127:	5	120:	1	14:	671	ON	742	OU	752	AN	0
6:	113:	5	123:	1	20:	699	ON	770	OU	780	AN	0
7:	93:	5	119:	1	3:	739	ON	810	OU	820	AN	0
8:	92:	5	118:	1	11:	741	ON	812	OU	822	AN	0
9:	81:	5	125:	1	5:	763	ON	834	OU	844	AN	0
10:	76:	5	126:	1	11:	773	ON	844	OU	854	AN	0
11:	87:	5	128:	1	260:	751	ON	822	OU	832	AN	0
12:	347:	5	128:	1	403:	396	ON	443	OU	455	OU	0
										LG= 1		
13:	750:	5	128:	1	139:	149	A	173	O	180	AU	2
14:	889:	5	128:	1	71:	172	A	172	A	177	A	2
15:	956:	5	128:	1	22:	148	A	164	AN	211	A	2
16:	976:	5	128:	1	21:	152	A	166	AN	221	A	2
17:	975:	5	128:	1	33:	138	A	148	AN	191	A	2
18:	952:	5	161:	1	11:	116	A	152	AN	175	A	2
19:	941:	5	162:	1	15:	131	A	141	AN	168	A	2
20:	904:	5	157:	1	17:	129	A	145	AN	172	A	2
21:	927:	5	166:	1	5:	133	A	141	AN	170	A	2
22:	922:	5	161:	1	29:	131	A	137	AN	172	A	2
23:	893:	5	154:	1	34:	128	A	136	AN	151	A	2
24:	885:	5	154:	1	62:	126	A	141	A	154	AN	2
25:	847:	5	154:	1	309:	164	A	178	A	185	A	2
										LG= 2		
26:	538:	5	154:	1	374:	256	ON	304	A	309	AN	0
27:	164:	5	154:	1	69:	674	ON	753	OU	755	AN	0
28:	197:	5	154:	1	43:	609	ON	696	OU	708	OU	0
29:	154:	5	154:	1	39:	668	ON	739	OU	749	AN	0
30:	115:	5	128:	1	36:	701	ON	772	OU	782	AN	0
31:	79:	5	131:	1	88:	767	ON	838	OU	848	AN	0
32:	167:	5	148:	1	337:	648	ON	745	OU	757	OU	0
33:	504:	5	138:	1	250:	264	ON	320	A	323	OU	0
										LG= 3		
34:	754:	5	138:	1	55:	162	A	188	A	207	AU	2
35:	803:	5	138:	1	34:	154	A	178	A	211	UN	2
36:	825:	5	138:	1	20:	163	A	167	A	207	A	2
37:	837:	5	138:	1	34:	168	A	174	A	202	A	2
38:	851:	5	138:	1	16:	159	A	165	A	190	A	2
39:	863:	5	138:	1	25:	164	A	166	A	173	A	2
40:	870:	5	138:	1	21:	154	A	163	A	186	A	2
41:	865:	5	138:	1	65:	155	A	190	AN	192	A	2
42:	806:	5	138:	1	83:	170	A	218	A	275	AN	2
										LG= 4		
43:	723:	5	138:	1	36:	202	A	241	AN	262	ON	0
44:	717:	5	114:	1	44:	187	A	210	AN	231	ON	2
45:	673:	5	113:	1	44:	199	A	230	AN	249	ON	2
46:	641:	5	111:	1	42:	205	ON	239	A	244	AN	0
47:	601:	5	110:	1	70:	206	ON	286	A	295	AN	0
48:	531:	5	107:	1	145:	256	ON	336	A	367	AN	0
49:	386:	5	90:	1	85:	412	ON	539	AN	550	A	0
										LG= 5		
50:	301:	5	78:	1	0:	527	ON	656	AN	667	A	0

Figure 33 : Détection des voyelles pour le mot "BABA"

la deuxième de 34 à 42. Sur les figures 34 et 35, nous avons représenté le sonagramme et le résultat de la détection des voyelles pour le mot *"DUPONT"*.

I.1.3. Algorithme de segmentation et d'étiquetage

Les processus de détection des voyelles sur le signal ainsi que celui d'étiquetage des prélèvements des segments des voyelles nécessitent le calcul de la plus petite distance entre le prélèvement et les références des voyelles. D'autre part, à l'aide de cette distance nous affectons le prélèvement soit à la classe des voyelles, soit à celle des consonnes voisées, d'autre part, nous calculons les étiquettes retenues pour chaque prélèvement.

En déroulant les deux processus en parallèle, on évite des calculs inutiles et par suite on réduit le temps de réponse.

Sur la figure 36, nous présentons l'algorithme de détection des voyelles et l'étiquetage des prélèvements des segments des voyelles. Sur cette figure : NBR indique le nombre de références des voyelles

NBC : nombre de canaux

I : numéro du canal

J : numéro de la référence

K : numéro du prélèvement courant

C : les valeurs des canaux du prélèvement courant

R : les valeurs des canaux des références

LG : nombre de prélèvement de la phrase prononcée.

numéro de prélèvement	énergie sur les 16 canaux	fondamental	courbe d'énergie	voisement
	MOT PRONONCE: DUPONT			
2	+.	129 :***.	:*	: *:
3	+:	128 :***.	:*	: *:
4	+:	125 :***.	:*	: *:
5	+:	125 :***.	:*	: *:
6	+:	126 :***.	:*	: *:
7	+:.	124 :***.	:*	: *:
/d/ 8	+:	128 :***.	:*	: *:
9	+:	132 :****	:*	: *:
10	+:: . ::::..	117 :***	: *	: *:
11	C++:::=+=++:::	117 :***	: *	: *:
12	3=++++++=++::.	117 :***	: *	: *:
13	3C=+++==++::.	117 :***	: *	: *:
14	3C=+++==+::::	195 :********	: *	: *:
15	3C=+++C=++:::	192 :********	: *	: *:
16	3C=+++CC=++:::	192 :********	: *	: *:
/y/ 17	3C=+++CC=+++::	192 :********	: *	: *:
18	3C=+++=C=+++::	190 :********	: *	: *:
19	3C=+++=C=+++::	188 :*******.	: *	: *:
20	3C=++==C=+++::	186 :*******.	: *	: *:
21	3C=++==+::.:.	186 :*******.	: *	: *:
22	=+:.::.:. .	186 :*******.	: *	: *:
23	+:. . .	186 :*******.	: *	: *:
24	+.	186 :*******.	:*	: *:
25	.	0 :	:*	:* :
26		0 :	:*	:* :
27	.	0 :	:*	:* :
28		0 :	:*	:* :
29		0 :	:*	:* :
/p/ 30		0 :	:*	:* :
31		0 :	:*	:* :
32		0 :	:*	:* :
33		0 :	:*	:* :
34		0 :	:*	:* :
35	:	0 :	:*	:* :
36	C=+++:..:	175 :*******	: *	: *:
37	33CC++:::.::	175 :*******	: *	: *:
38	33CC++:::.::..	175 :*******	: *	: *:
39	33CC++:::.::.:	175 :*******	: *	: *:
40	33CC++:::.::::	175 :*******	: *	: *:
41	333C++:::.:+.:	175 :*******	: *	: *:
42	333C++::..:+..	125 :***.	: *	: *:
/o/ 43	33CC++::..:: .	125 :***.	: *	: *:
44	3CCC++::. .: .	125 :***.	: *	: *:
45	CCCC++::...: .	125 :***.	: *	: *:
46	CCC=+::. ..: .	79 :*	: *	: *:
47	CCC=:::. . .	79 :*	: *	: *:
48	C=C=:::.	79 :*	: *	: *:
49	C=C=+:. . :	79 :*	: *	: *:
50	=+=+..	79 :*	: *	: *:
51	=+::	79 :*	: *	: *:
52	=:+:	79 :*	: *	: *:

Sonagramme du mot *"DUPONT"*

Figure 34

```
 2: 10000 : A  :10000 : A  :10000 : A  : 0
 3: 10000 : A  :10000 : A  :10000 : A  : 0
 4: 10000 : A  :10000 : A  :10000 : A  : 0
 5: 10000 : A  :10000 : A  :10000 : A  : 0
 6: 10000 : A  :10000 : A  :10000 : A  : 0
 7: 10000 : A  :10000 : A  :10000 : A  : 0
 8: 10000 : A  :10000 : A  :10000 : A  : 0
 9: 10000 : A  :10000 : A  :10000 : A  : 0
10:   357 : AI :  384 : I  :  473 : EI : 0
                                    LG=  1

11:   156 : I  :  188 : I  :  214 : I  : 2
12:   133 : I  :  143 : I  :  164 : AI : 2
13:    93 : AI :   94 : U  :   94 : EU : 2
14:    86 : U  :   90 : U  :  128 : EU : 2
15:    65 : U  :  125 : U  :  145 : EI : 2
16:    61 : U  :  123 : U  :  143 : EU : 2
17:    73 : U  :  139 : U  :  153 : EI : 2
18:    72 : U  :  122 : U  :  138 : EU : 2
19:    64 : U  :  108 : U  :  124 : EU : 2
20:    75 : U  :  117 : U  :  121 : EU : 2
21:   107 : U  :  144 : EU :  160 : EU : 2
                                    LG=  2

22: 10000 : U  :10000 : EU :10000 : EU : 0
23: 10000 : U  :10000 : EU :10000 : EU : 0
24: 10000 : U  :10000 : EU :10000 : EU : 0
25: 10000 : U  :10000 : EU :10000 : EU : 0
26: 10000 : U  :10000 : EU :10000 : EU : 0
27: 10000 : U  :10000 : EU :10000 : EU : 0
28: 10000 : U  :10000 : EU :10000 : EU : 0
29: 10000 : U  :10000 : EU :10000 : EU : 0
30: 10000 : U  :10000 : EU :10000 : EU : 0
31: 10000 : U  :10000 : EU :10000 : EU : 0
32: 10000 : U  :10000 : EU :10000 : EU : 0
33: 10000 : U  :10000 : EU :10000 : EU : 0
34: 10000 : U  :10000 : EU :10000 : EU : 0
35: 10000 : U  :10000 : EU :10000 : EU : 0
36:   205 : AU :  290 : OU :  300 : OU : 0
                                    LG=  3

37:   155 : O  :  170 : OU :  188 : AU : 2
38:   145 : O  :  145 : A  :  162 : AU : 2
39:   133 : A  :  153 : O  :  166 : AU : 2
40:   168 : A  :  179 : AN :  180 : O  : 2
41:   178 : A  :  185 : AN :  212 : O  : 2
42:   160 : AN :  191 : A  :  235 : A  : 2
43:    95 : AN :  259 : OU :  264 : A  : 2
44:   140 : AN :  269 : AU :  271 : A  : 2
45:   133 : AN :  230 : A  :  234 : A  : 2
46:   156 : AN :  281 : A  :  287 : A  : 2
                                    LG=  4

47:   207 : AN :  270 : A  :  272 : AU : 0
48:   286 : AU :  317 : AN :  324 : A  : 0
49:   390 : AN :  393 : AU :  397 : A  : 0
50: 10000 : AN :10000 : AU :10000 : A  : 0
51: 10000 : AN :10000 : AU :10000 : A  : 0
                                    LG=  5

52: 10000 : AN :10000 : AU :10000 : A  : 0
```

Détection des voyelles dans le mot *"DUPONT"*

Figure 35

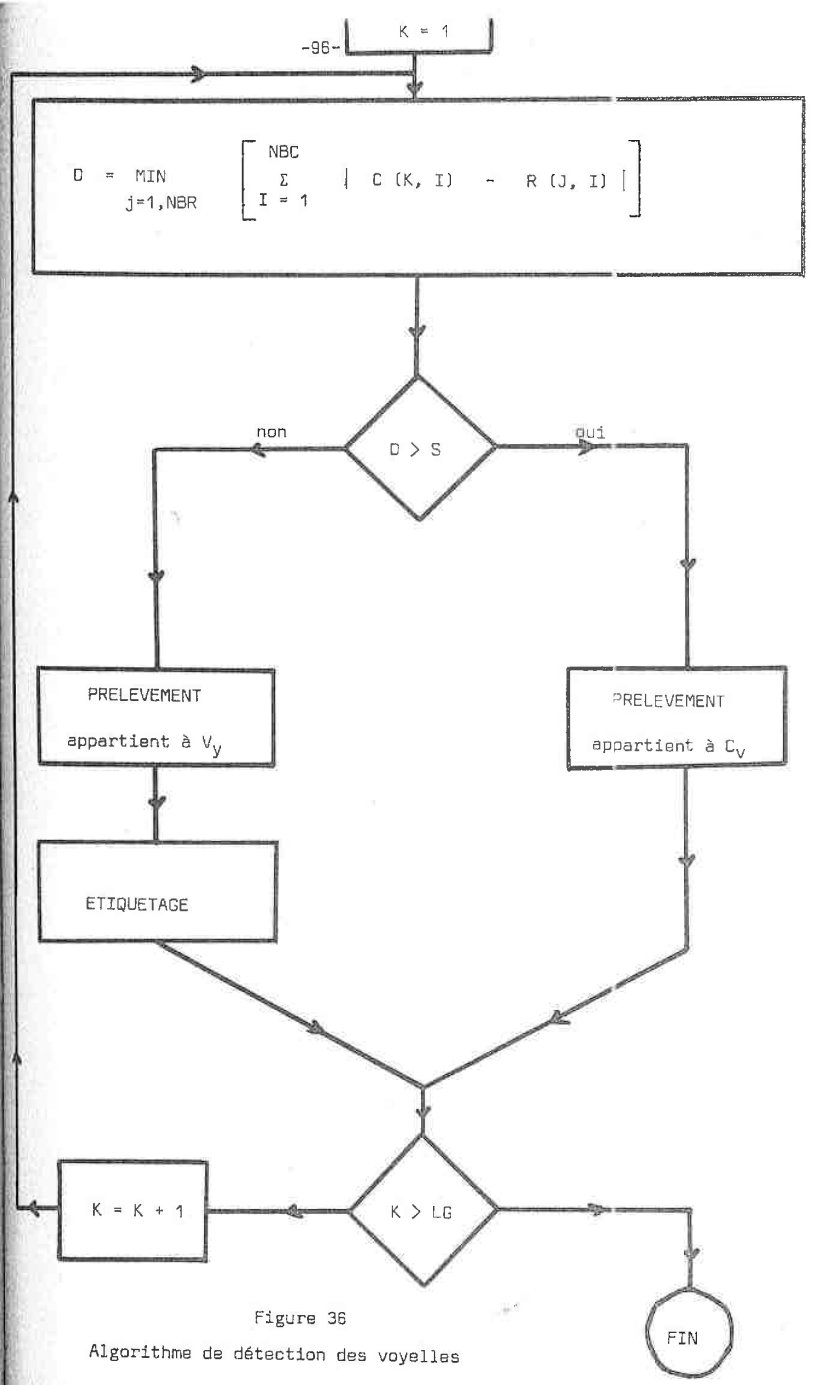


Figure 36

Algorithme de détection des voyelles

I.1.4. Normalisation du seuil de segmentation

Le seuil de segmentation pose deux types de problèmes, pouvant entraîner des erreurs de détection des voyelles :

1. Le locuteur prononce certains phonèmes, dans la phrase à reconnaître, d'une voix plus forte qu'au moment de l'apprentissage.

2. Les conditions d'acquisition au moment de la reconnaissance sont différentes de celles de l'apprentissage (bruit ambiant plus fort, réglage du gain différent ...).

Dans les deux cas, le prélèvement à reconnaître est très énergétique. Lorsqu'on calcule la plus petite distance de ce prélèvement aux références des voyelles nous obtenons une distance supérieure au seuil S et le système conclura que le prélèvement n'appartient pas à la classe des voyelles.

Pour remédier à ce problème, le seuil de segmentation est normalisé en fonction de l'énergie. La figure 37 donne l'algorithme de la normalisation.

SI (l'énergie totale du prélèvement > IE)
ET (D > S)
ALORS
le seuil S est augmenté de X
(c'est-à-dire SD = S + X)
FSI

où

IE : la valeur de l'énergie au dessus de laquelle le prélèvement est considéré appartenant à la classe des voyelles.

D : la plus petite distance entre le prélèvement et les références des voyelles.

S : le seuil de segmentation.

SD : le nouveau seuil de segmentation.

Figure 37 : Algorithme de normalisation du seuil de segmentation

Remarques :

- Si l'énergie totale du prélèvement est supérieure à IE, et si la plus petite distance entre le prélèvement et les références des voyelles est toujours supérieure au seuil de segmentation normalisé, SD, on demande au locuteur de parler moins fort.

- L'algorithme de normalisation du seuil de segmentation permet aussi de tenir compte de certaines erreurs dues à la position du microphone. Lorsque celui-ci est très proche des lèvres, l'energie des voyelles s'en trouve augmentée. A l'aide de cet algorithme le seuil de segmentation est asservi à la position du microphone.

I.2. IDENTIFICATION DES VOYELLES

Les segments des voyelles étant localisés, dans le signal, il s'agit d'identifier ces derniers, à partir des trois phones les plus plausibles affectés à chaque prélèvement.

I.2.1. Identification des segments appartenant à la classe des voyelles

L'identification des segments des voyelles se fait par un processus de décision majoritaire parmi les trois phones retenus pour chaque échantillon.

Pour chaque voyelle candidate, nous calculons son taux d'occurrence dans le segment de la façon suivante :

$$T_i = \sum_{j=1}^{N} P_j , \quad i = 1, \ldots, NBR$$

avec

NBR : nombre de références des voyelles

N : nombre d'occurrences de la voyelle i dans le segment.

P_j peut prendre les valeurs 0, 1, 2 ou 3 en fonction de la position du phonème dans le treillis.

P_j prend la valeur 3 si la $i^{ème}$ voyelle se trouve au premier rang du treillis, la valeur 2 s'il est au deuxième rang, la valeur 1 s'il est au troisième rang. Et si le phonème ne figure pas dans le treillis, P_j prend la valeur zéro.

Les voyelles correspondant aux trois plus grandes valeurs des T_i, classées par ordre décroissant, fournissent le triplet identifiant le segment. Désignons par T1, T2 et T3 les plus grandes valeurs des T_i,

T1 = Max T_i, i = 1,..., NBR

T2 = Max T_i avec T2 $\leqslant$ T1

T3 = Max T_i avec T3 $\leqslant$ T2

Ainsi le segment est identifié par les voyelles correspondant à T1, T2 et T3.

Prenons l'exemple d'un segment de voyelles composé de trois prélèvements. Par comparaison aux formes de références des voyelles, nous avons étiqueté chaque échantillon par les trois phones suivants :

1. A O I
2. O A I
3. A O AU

alors $T_A = 3 + 2 + 3 = 8$

$T_O = 2 + 3 + 2 = 7$

$T_I = 1 + 1 = 2$

$T_{AU} = 1$

les autres voyelles ne sont pas présentes dans le segment, donc $T_X = 0$ d'où le segment sera identifié par

/A/ /O/ /I/

puisque $T_A > T_O > T_I$

Les résultats de la reconnaissance des voyelles de l'exemple précédent (figure 32) sont décrits sur la figure 38. La deuxième colonne indique le type de segment. A ce niveau dans le système, on distingue trois types :

- macro-classe = 1 : signifie que le segment appartient à la classe des consonnes voisées
- macro-classe = 2 : signifie que le segment appartient à la classe des consonnes non voisées
- macro-classe = 3 : signifie que le segment appartient à la classe des voyelles.

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	1	V	V	V	2	11
2	3	A	AN		13	13
3	1	V	V	V	26	8
4	3	A			34	9
5	1	V	V	V	43	6

numéro du prélèvement du début du segment dans le sonagramme

nombre de prélèvements dans le segment

Figure 38 : résultats de la reconnaissance des voyelles

Remarque :

Le deuxième segment est identifié uniquement par deux voyelles, la troisième est négligée devant la seconde. De même, le quatrième segment est identifié par la voyelle /A/ à 100 %.

Les résultats de la reconnaissance des voyelles du mot "DUPONT" (figure 34) sont indiqués sur la figure 39.

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	Nd Pr
1	1	V	V	V	2	
2	3	U			11	
3	1	V	V	V	22	
4	2	N	N	N	25	
5	3	A	AN	U	37	
6	1	V	V	V	47	

Figure 39 : reconnaissance des voyelles pour le mot "DUPONT"

I.2.2. Réduction du treillis

L'identification des segments des voyelles est réalisée par un processus de décision majoritaire parmi les trois phones retenus pour chaque échantillon. Les voyelles ayant un taux d'occurrence faible ne doivent pas apparaître dans le triplet identifiant le segment.

En se basant sur le rapport

$$\frac{T_{i-1}}{T_i} \qquad i \in [2,3],$$

nous pouvons négliger le phonème de rang i par rapport à celui de rang i - 1, si ce rapport est supérieur à un seuil fixe.

Dans l'exemple du paragraphe précédent les valeurs résultats $T_A = 8$, $T_O = 7$ et $T_I = 2$. En prenant $S = 2,5$, le rapport T_A/T_O n'est pas supérieur à S. Par contre T_O/T_I est supérieur à S, donc le segment sera identifié seulement par les deux voyelles :

/A/ /O/

I.2.3. Fusion de deux segments de voyelles consécutifs

En comparant les résultats de la reconnaissance des segments des voyelles consécutifs, on peut corriger certaines erreurs dues à la sur-segmentation du signal.

Si deux segments de voyelles consécutifs satisfont un certain critère alors ces deux segments seront fusionnés.

Le critère de fusion se présente comme suit :

si V11 V12 V13 représente le premier triplet

et V21 V22 V23 le deuxième triplet

nous fusionnons les deux segments si :

V11 = V21 ou V22 ou V23

ou

V12 = V21 ou V22 ou V23

ou

V13 = V21 ou V22 ou V23

Les trois voyelles identifiant le segment fusionné sont celles qui ont les plus grandes valeurs V11, V12, V13, V21, V22 et V23.

I.2.4. Schéma général de reconnaissance des voyelles

Le module de reconnaissance des voyelles par la méthode spectrale peut être schématisé comme sur la figure 40. En introduisant les paramètres de voisement et d'énergie, on évite des calculs inutiles et par suite on réduit le temps de réponse.

II. RECONNAISSANCE DES VOYELLES PAR LA COURBE D'ENERGIE

II.1. INTRODUCTION

La prononciation en contexte d'un phonème, modifie ses caractéristiques. Ces variations sont aussi liées au mode d'élocution adopté (rapide, lente ...). Pour tenir compte des ces altérations au niveau de la détection des voyelles, nous avons développé une méthode qui se déroule en parallèle avec la méthode présentée dans le paragraphe précédent.

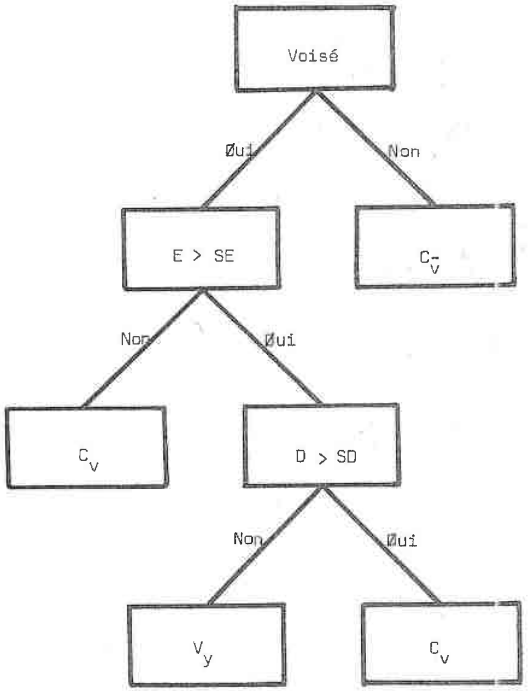


Figure 40 : processus de reconnaissance des voyelles

avec E : énergie d'un échantillon

D : distance entre l'échantillon et les références des voyelles

C_v : consonne voisée

$C_{\bar{v}}$: consonne non voisée

V_y : voyelle

Comme les voyelles sont les plus énergétiques de tous les phonèmes, en détectant les maxima de la courbe d'énergie, nous localisons les noyaux vocaliques, et par suite la présence ou l'absence de voyelles dans un segment donné.

L'algorithme de reconnaissance des voyelles par la courbe d'énergie se déroule en deux phases successives :

- détection des noyaux vocaliques,
- segmentation et identification des voyelles.

II.2. DETECTION DES NOYAUX VOCALIQUES

II.2.1. Choix de la courbe d'énergie

Les pics dans la courbe d'énergie indiquent les zones stables des voyelles d'une phrase, à l'exception de certains cas particuliers (barre d'explosion, les sonantes en contact avec certains phonèmes, etc ...). Pour éliminer les faux pics qui ne représentent pas des noyaux vocaliques, d'une part, nous ne prenons pas en compte les maxima non stables, d'autre part, nous prenons une bande de fréquence particulière pour calculer l'énergie. En effet, en prenant la bande 100-6000 Hz, les fricatives, ainsi que les liquides, perturbent beaucoup la courbe d'énergie (figures 41-44). D'où la nécessité de bien choisir la bande de fréquence pour calculer l'énergie, afin que la courbe d'énergie soit perturbée le moins possible par les consonnes voisées.

L'énergie dans les hautes fréquences induit des erreurs lors de la recherche des noyaux vocaliques, puisque d'une part les voyelles ont très peu d'énergie dans cette bande, et que d'autre part, les fricatives en présentent beaucoup.

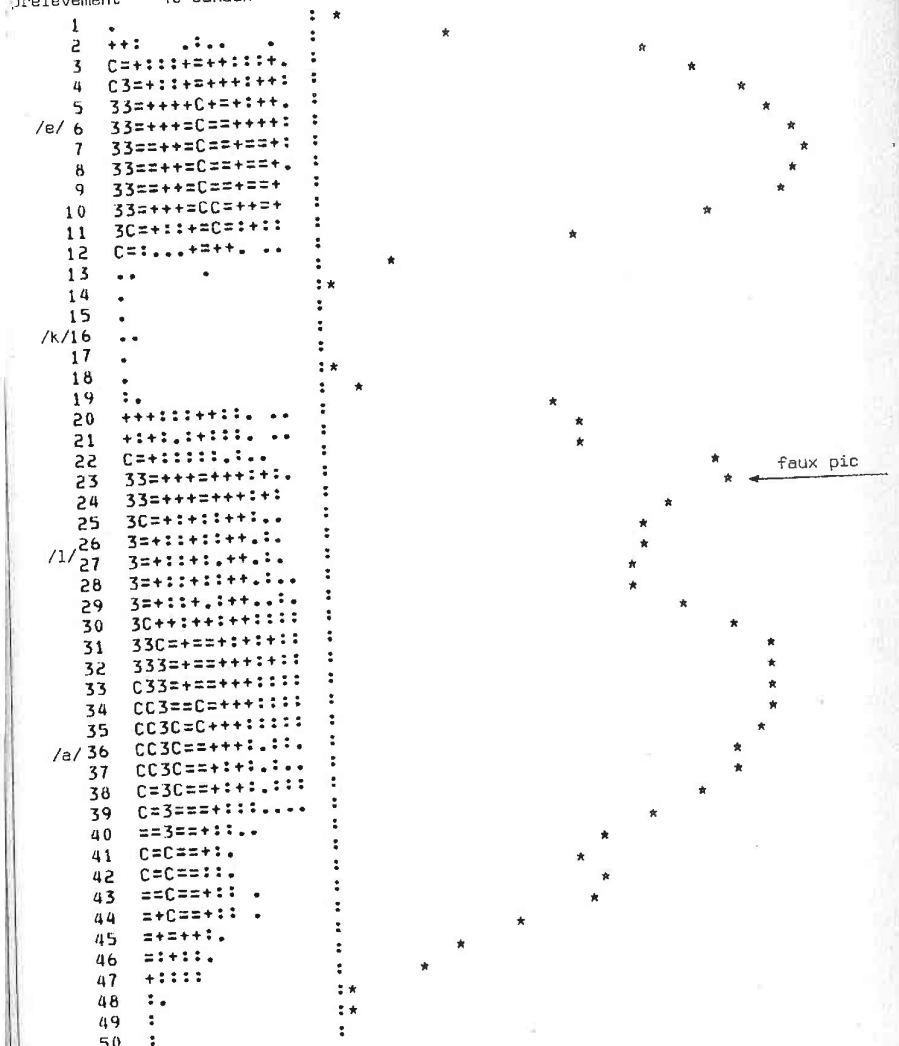
Il en est de même en ce qui concerne les basses fréquences, car les sonantes peuvent présenter beaucoup d'énergie dans cette bande. Donc, l'énergie utile pour détecter les noyaux vocaliques se situe à une fréquence moyenne.

Dans notre système, nous prenons l'énergie dans la bande 500 - 3000 Hz.

II.2.2. Recherche des pics sur la courbe d'énergie

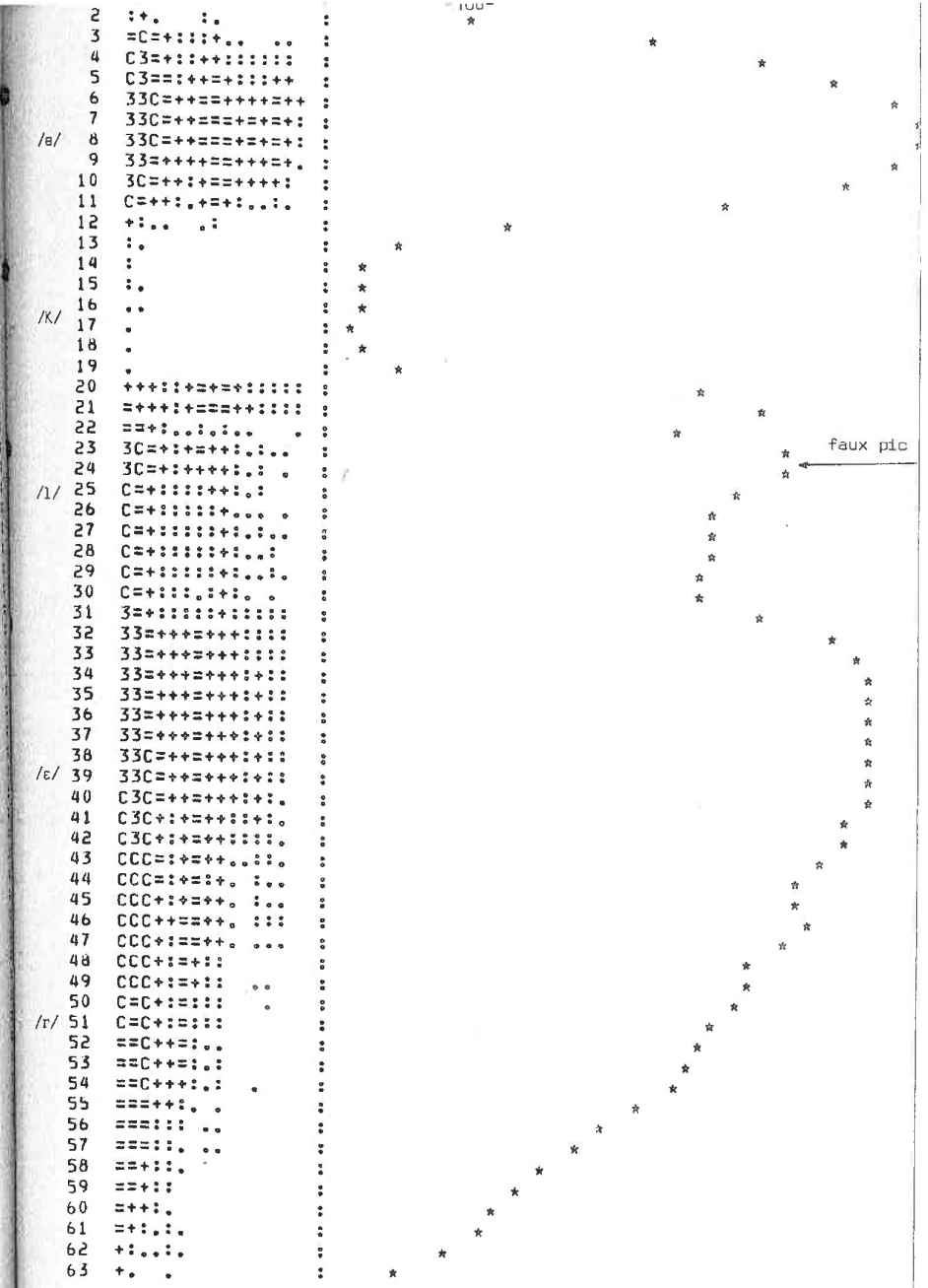
Afin de repérer les noyaux vocaliques, nous examinons la suite des pics et des vallées dans la courbe de l'énergie. Dans la plupart des cas, la voyelle est définie par un pic d'énergie. Or les consonnes sonantes voisées possèdent également une forte énergie ; d'où l'apparition de faux pics sur la courbe.

Il est donc nécessaire de définir la largeur du pic, c'est-à-dire l'étendue de la voyelle, de manière à éliminer les consonnes sonantes qui entourent la voyelle. Ceci pourrait conduire à ne conserver que l'extrême pointe du pic, et à la limite, un seul point [CARB-83]. Le processus de détection des noyaux vocaliques est schématisé par l'algorithme de la figure 45.



MOT PRONONCE : ECLAT

Figure 41 : Sonagramme du mot "*éclat*"



MOT PRONONCE : ECLAIR

Figure 42 : Sonagramme du mot "*éclair*"

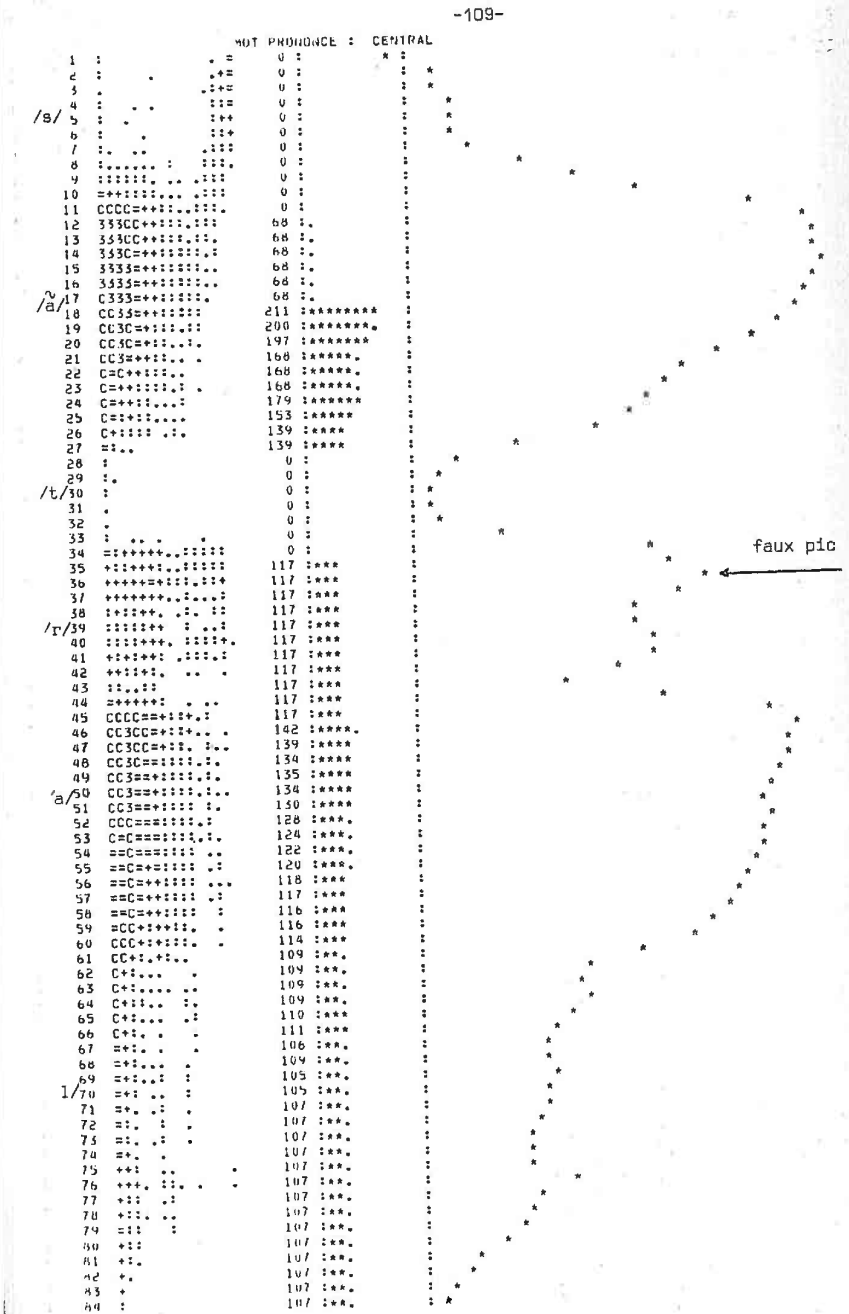


Figure 43 : Sonagramme du mot "*central*"

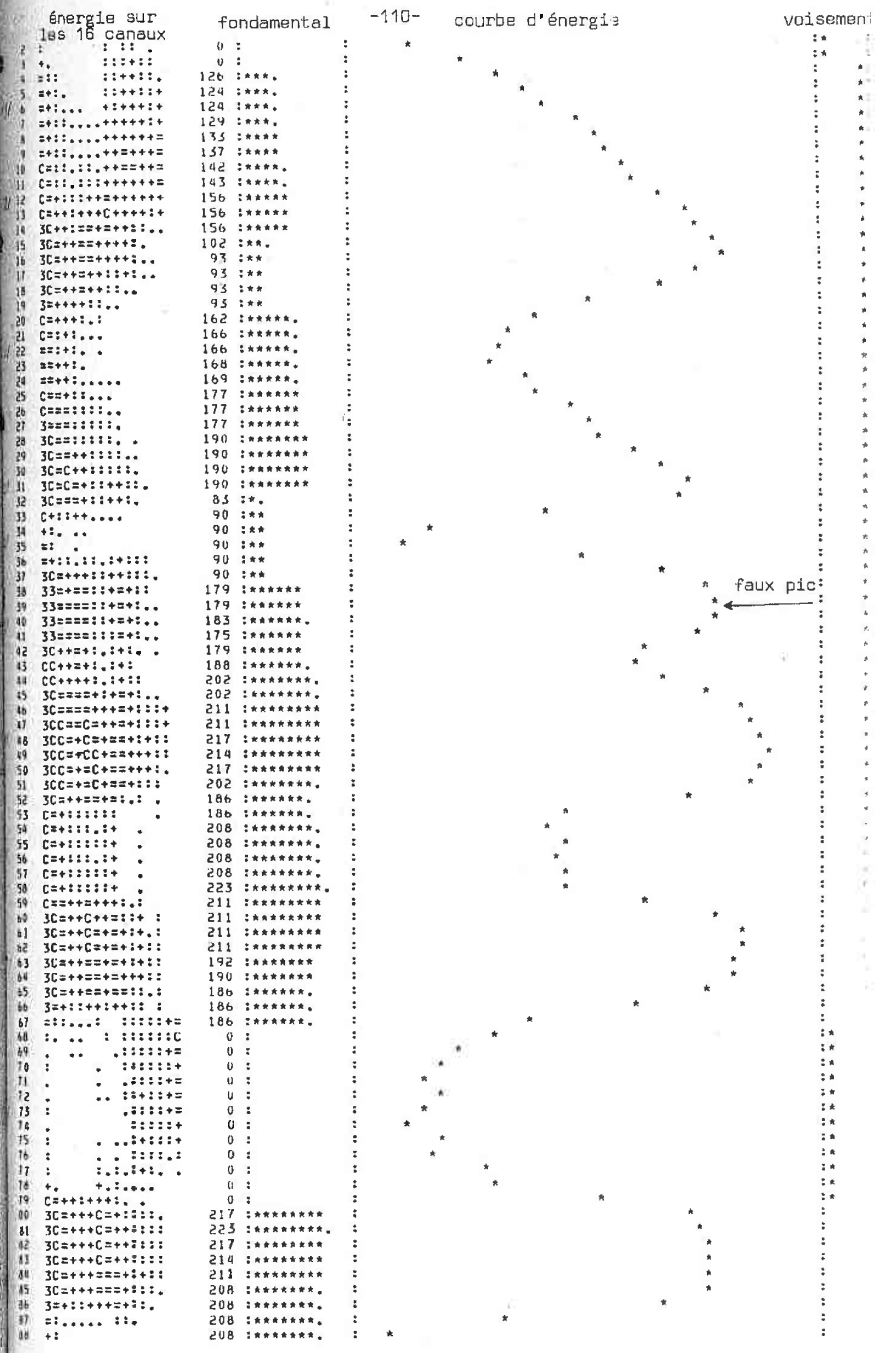


Figure 44 : Sonagramme de la phrase "*je voudrais Monsieur*"

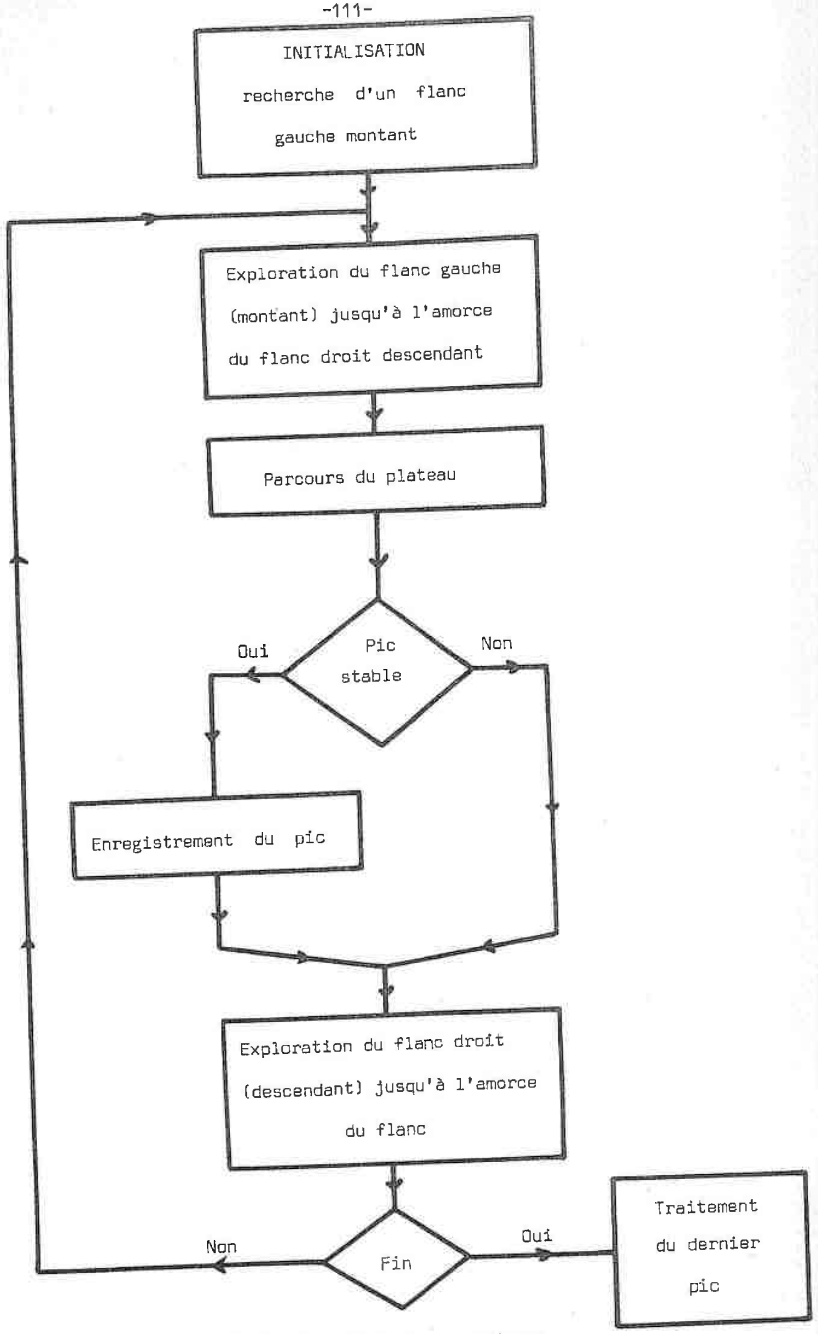


Figure 45 : Recherche des noyaux vocaliques

II.2.3. Stabilité des pics

Ainsi que le montrent les figures 41 - 44, on trouve fréquemment des pics, sur la courbe d'énergie, qui sont le plus souvent énergétiquement proches des voyelles et qui représentent soit la transition d'une voyelle vers une consonne, soit une sonante, soit un "burst" très marqué.

Ces pics ont une caractéristique particulière : leur faible stabilité par rapport aux pics des voyelles. Autrement dit, les voyelles présentent un plateau qui dure un certain temps, par contre les pics qui ne représentent pas des voyelles ont un plateau très bref.

En utilisant le nombre de prélèvements dont l'énergie est supérieure à un seuil S, nous pouvons distinguer les pics, représentant les voyelles, des autres pics. Ce seuil est variable, nous le prenons égal au 2/3 du maximum de l'énergie de la phrase prononcée.

Lorsque le nombre de prélèvements (au dessus de S), est suffisamment élevé nous enregistrons le pic, dans le cas contraire on considère que le pic ne représente pas un noyau vocalique.

II.2.4. Traitement du dernier pic

Le plus souvent la dernière voyelle est moins énergétique que les autres (dans un mot ou une phrase à reconnaître). Pour tenir compte de cet affaiblissement d'énergie, nous diminuons la valeur du seuil S. Dans le système nous prenons $\frac{3}{4}$ S pour nouvelle valeur du seuil. Le traitement du dernier pic est schématisé sur la figure 46.

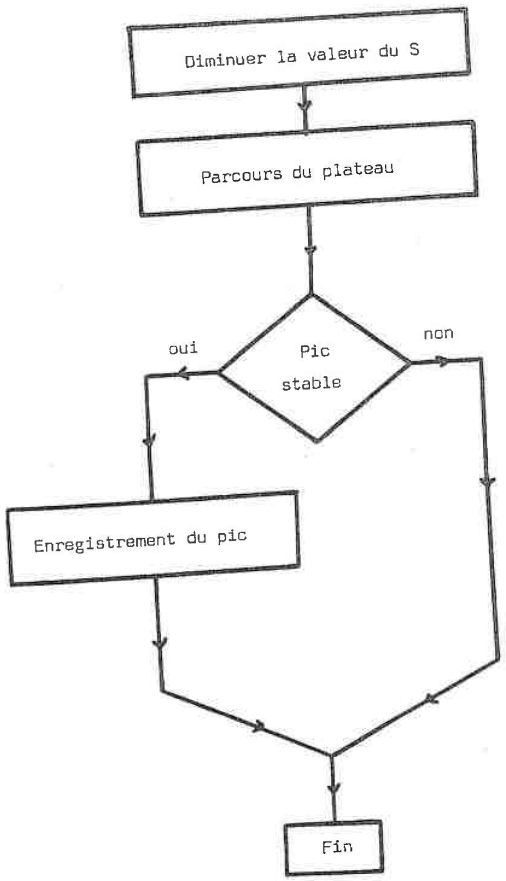


Figure 46 : Traitement du dernier pic

II.3. RECONNAISSANCE DES VOYELLES DANS LES SEGMENTS VOISES

Ce paragraphe décrit la recherche des voyelles dans les segments appartenant à la classe des consonnes voisées. Si une voyelle est présente dans de tels segments une fonction de segmentation acaptée est nécessaire pour trouver les frontières de la voyelle et pour l'identifier.

Donc, nous nous trouvons en face de trois types de problèmes :

- détection des voyelles dans les segments voisés,
- segmentation des zones voisées,
- identification des voyelles.

II.3.1. Détection des voyelles dans les segments voisés

Dans les paragraphes précédents, nous avons extrait deux types d'informations : d'une part, nous avons segmenté le signal vocal en trois classes :

- les voyelles,
- les zones voisées,
- les zones non voisées,

d'autre part, nous avons détecté sur la courbe d'énergie les noyaux vocaliques.

Pour rectifier certaines erreurs de segmentation et d'identification, nous avons mis en oeuvre un processus liant deux modules, celui de la reconnaissance des voyelles par la méthode spectrale, et celui de la détection des noyaux vocaliques par la courbe d'énergie (figure 47).

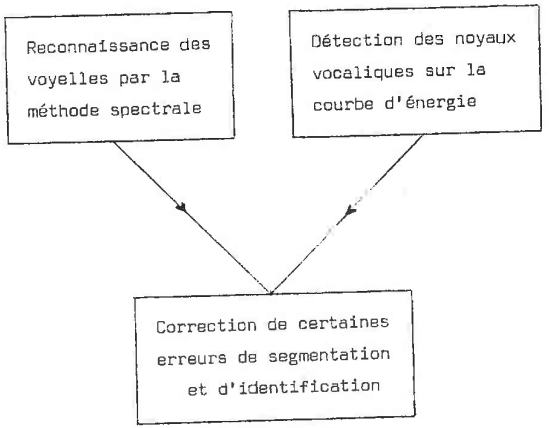


Figure 47 : Reconnaissance des voyelles par la courbe d'énergie

Si on détecte un noyau vocalique, dans un segment appartenant à la classe des consonnes voisées, cela nous indique la présence d'une voyelle qui n'a pas été détectée par le module de reconnaissance des voyelles utilisant la méthode spectrale, et donc il s'agit de localiser précisément puis d'identifier cette voyelle.

II.3.2. Calcul du seuil de segmentation

La fonction de segmentation adoptée est la même que dans le module de reconnaissance des voyelles par analyse spectrale, qui consiste en une segmentation par identification. Mais le seuil qui permet de décider si un échantillon appartient ou non à la classe des voyelles est normalisé.

Les voyelles étant caractérisées essentiellement par leurs parties stables dans le signal, nous prenons le pic dans la courbe d'énergie, comme identificateur de la voyelle. Donc, on prend pour valeur du seuil de segmentation la plus petite distance entre l'échantillon correspondant au pic de la courbe d'énergie et toutes les références des voyelles. Autrement dit, nous comparons le prélèvement correspondant au pic dans la courbe d'énergie à toutes les références des voyelles à l'aide de la fonction :

$$D_j = \sum_{i=1}^{NB} | C(k,i) - R(j,i) |$$

où NB est le nombre de canaux

k est le numéro du prélèvement correspondant au pic

C les différentes valeurs des canaux du prélèvement

R les différentes valeurs des canaux de la référence.

Puis nous cherchons la plus petite valeur de ces D_j :

$$D = MIN \{ D_j \}, \quad j = 1, \ldots, NBR.$$

NBR étant le nombre de références des voyelles.

On choisit comme seuil de segmentation la valeur de D augmentée d'une petite quantité ε pour tenir compte du contexte de la voyelle :

$$S = D + \varepsilon$$

II.2.3. Segmentation des zones voisées

La présence d'un noyau vocalique dans un segment appartenant à la classe des consonnes voisées indique l'existence d'une voyelle dans ce segment. Elle peut se trouver, soit au début, soit au milieu, soit à la fin du segment (figure 48).

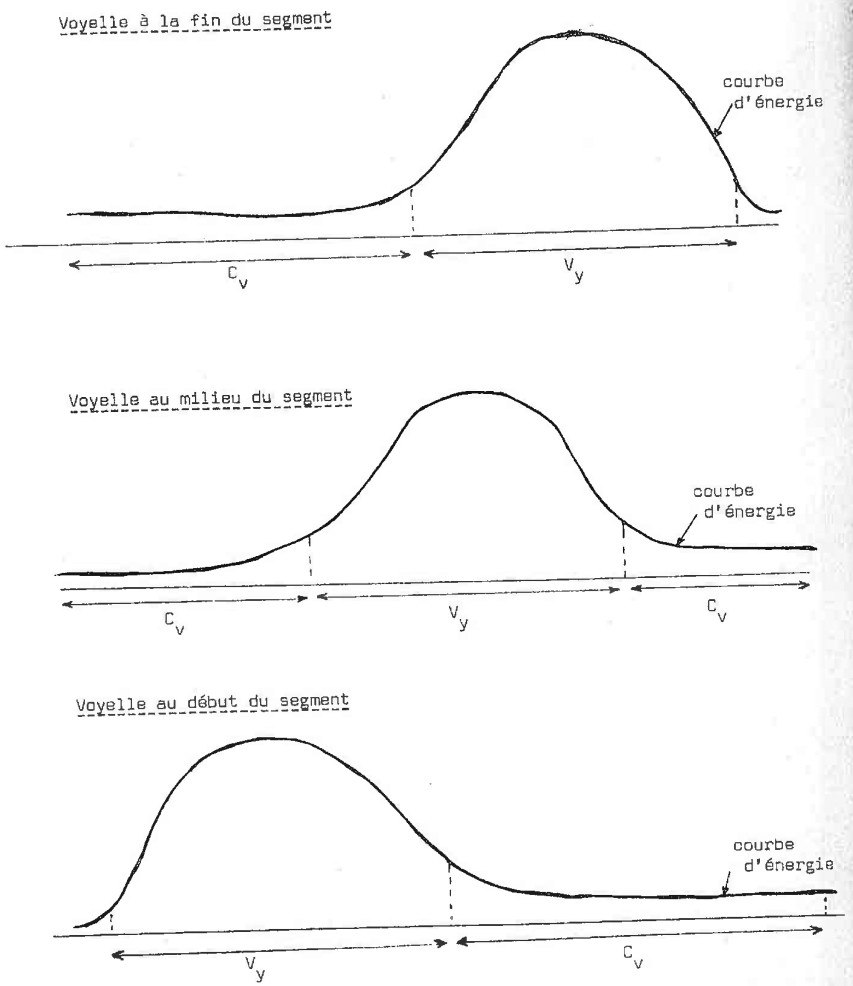


Figure 48 : Positions de la voyelle dans un segment voisé.
V_y : Voyelle
C_v : une ou plusieurs consonnes voisées

La partie stable de la voyelle se situe à proximité du sommet de la courbe d'énergie. La localisation de la voyelle dans le segment est obtenue à partir du prélèvement correspondant au pic. Pour trouver la frontière droite de la voyelle nous parcourons le signal de gauche à droite, à partir du pic et inversement pour la frontière gauche (figure 49).

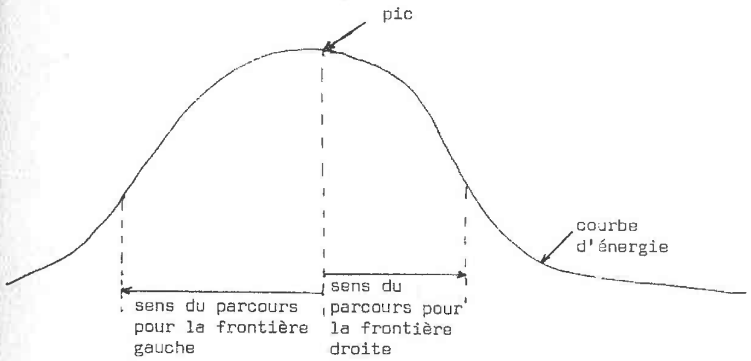


Figure 49 : Recherche de la frontière gauche et droite de la voyelle

La segmentation se fait donc en deux étapes : recherche des limites droite et gauche de la voyelle. La marque de segmentation sera placée au premier échantillon dont le taux de dissemblance avec les références des voyelles est inférieur au seuil de segmentation.

II.3.3.1. Détermination de la frontière droite de la voyelle

En parcourant le signal vocal de gauche à droite à partir du pic, nous comparons chaque prélèvement à toutes les références des voyelles à l'aide de la fonction $D_i = \Sigma |C - R|$. Ensuite, on détermine le minimum de toutes ces distances :

$$D1 = MIN \{D_i\}, \quad i = 1, \quad NBR$$

Si D1 est inférieur au seuil de segmentation S, nous considérons alors que le prélèvement appartient à la classe des voyelles, et on lui affecte les trois phones correspondant respectivement à D1, D2, D3

où $D2 = MIN \quad D_i$ avec $D2 \geqslant D1$

et $D3 = MIN \quad D_i$ avec $D3 \geqslant D2$.

Dans le cas contraire, on considère que le prélèvement n'appartient pas à la classe des voyelles.

La frontière droite de la voyelle est placée au premier prélèvement n'appartenant pas à la classe des voyelles.

II.3.3.2. Détermination de la frontière gauche de la voyelle

Pour trouver la frontière gauche de la voyelle, nous procédons de la même manière que pour déterminer la frontière droite mais en parcourant le signal de droite à gauche à partir du pic. Nous prenons la même valeur du seuil de segmentation que dans le paragraphe précédent.

De même, la frontière gauche de la voyelle est mise au premier prélèvement n'appartenant à la classe des voyelles.

II.4. IDENTIFICATION DE LA VOYELLE

L'identification de la voyelle se fait par un processus de décision majoritaire parmi les trois phones retenus pour chaque échantillon.

Pour chaque prélèvement du segment, nous affectons le poids 3 au premier élément du treillis, le poids 2 au deuxième et le poids 1 au troisième. En additionnant les différents poids du même phonème dans le segment, nous obtenons pour chaque voyelle son taux d'occurrence dans le segment. Le treillis identifiant le segment sera composé par les voyelles ayant les plus grands taux, dans l'ordre décroissant.

Si nous remarquons que le taux de la troisième voyelle dans le treillis est négligeable par rapport à celui de la ceuxième alors le segment sera identifié uniquement par les deux premières voyelles du treillis.

De même, si le taux de la deuxième voyelle du treillis est négligeable par rapport à celui de la première alors le segment sera identifié uniquement par la première voyelle du treillis.

III. SEGMENTATION ET IDENTIFICATION DES VOYELLES LONGUES

Si deux voyelles (ou plusieurs) se trouvent en contact et s'il n'y a pas entre elles une zone transitoire alors ces voyelles seront considérées comme formant un seul segment, car ni le module de reconnaissance des voyelles par la courbe d'énergie, ni le module de reconnaissance des voyelles par la méthode spectrale ne peut détecter les deux voyelles.

En effet, d'une part, le fait qu'il n'existe pas de zone de transition entre les voyelles, implique qu'il n'apparaîtra qu'un seul pic sur la courbe d'énergie. Donc le module de reconnaissance des voyelles par la courbe d'énergie en déduit qu'il n'y a qu'une seule voyelle (figures 50 et 51).

D'autre part, puisque tous les prélèvements appartiennent à la classe des voyelles, il n'y a pas de marque de segmentation entre les voyelles. Et par suite, le module de reconnaissance des voyelles par la méthode spectrale en déduit qu'il n'y a qu'une seule voyelle (figure 52). Les résultats du module de la reconnaissance des voyelles sont indiqués sur la figure 53.

Pour remédier à ce problème, on consulte le nombre de prélèvements des segments des voyelles :

si ce nombre est inférieur à un seuil on conclut qu'il n'y a qu'une seule voyelle,

sinon on segmente la zone à partir d'une fonction de variation spectrale :

$$VSP = \sum_{i=1}^{NB} | canal\ (i, j) - canal\ (i, j-1) |$$

où NB est le nombre de canaux de l'analyseur,

j est le numéro du prélèvement courant.

L'identification des segments est alors réalisée à partir des étiquettes des échantillons (figure 54). Le résultat de la reconnaissance des voyelles pour le mot "IA" est indiqué sur les figures 55 et 56. Un autre exemple est donné sur la figure 57.

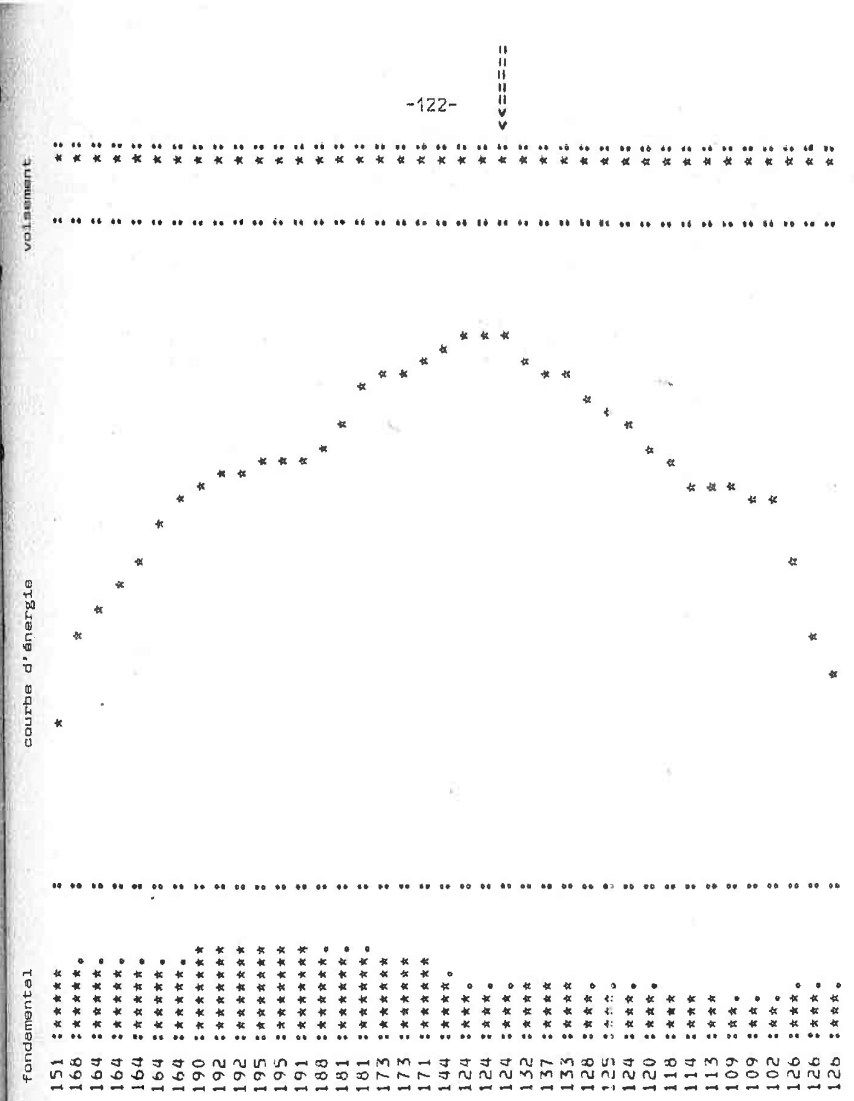


Figure 50 : Spectrogramme de /O a/

Figure 51 : Spectrogramme de /y a/

2:	10000	:	A	:	10000	:	A	:	10000	:	A	: 0
3:	140	:	I	:	145	:	AI	:	213	:	EI	: 2
4:	107	:	I	:	154	:	AI	:	180	:	EI	: 2
5:	107	:	I	:	149	:	I	:	178	:	EI	: 2
6:	153	:	I	:	154	:	AI	:	179	:	I	: 2
7:	143	:	I	:	144	:	AI	:	166	:	EI	: 2
8:	117	:	I	:	137	:	I	:	142	:	EI	: 2
9:	108	:	I	:	122	:	I	:	159	:	EI	: 2
10:	98	:	I	:	98	:	I	:	167	:	EI	: 2
11:	79	:	I	:	115	:	I	:	192	:	EI	: 2
12:	79	:	I	:	125	:	I	:	202	:	EI	: 2
13:	85	:	I	:	153	:	I	:	219	:	EI	: 2
14:	109	:	I	:	163	:	I	:	195	:	EI	: 2
15:	150	:	I	:	178	:	EI	:	200	:	I	: 2
16:	250	:	EI	:	252	:	I	:	264	:	I	: 2
17:	200	:	EI	:	252	:	I	:	271	:	AI	: 2
18:	157	:	EI	:	261	:	U	:	273	:	U	: 2
19:	188	:	EI	:	270	:	EU	:	274	:	U	: 2
20:	223	:	EI	:	289	:	U	:	291	:	EU	: 2
21:	187	:	EI	:	251	:	U	:	269	:	EU	: 2
22:	210	:	EI	:	241	:	AI	:	248	:	EU	: 2
23:	197	:	UN	:	217	:	O	:	282	:	EU	: 2
24:	216	:	UN	:	273	:	UN	:	304	:	A	: 2
25:	245	:	UN	:	287	:	A	:	304	:	ON	: 2
26:	247	:	UN	:	281	:	A	:	286	:	ON	: 2
27:	209	:	UN	:	244	:	ON	:	257	:	A	: 2
28:	194	:	UN	:	205	:	ON	:	237	:	A	: 2
29:	179	:	UN	:	194	:	A	:	200	:	ON	: 2
30:	165	:	A	:	181	:	AN	:	186	:	UN	: 2
31:	142	:	AN	:	173	:	A	:	189	:	A	: 2
32:	157	:	AN	:	178	:	A	:	197	:	A	: 2
33:	197	:	A	:	219	:	A	:	233	:	A	: 2
34:	195	:	A	:	222	:	AN	:	235	:	A	: 2
35:	183	:	AN	:	230	:	A	:	290	:	A	: 2
											LG=	1
36:	241	:	A	:	256	:	AN	:	387	:	AU	: 0
37:	297	:	A	:	358	:	AN	:	419	:	AU	: 0
38:	286	:	A	:	300	:	AU	:	323	:	AN	: 0
39:	362	:	AU	:	386	:	A	:	393	:	AN	: 0
40:	10000	:	AU	:	10000	:	A	:	10000	:	AN	: 0
41:	10000	:	AU	:	10000	:	A	:	10000	:	AN	: 0
42:	10000	:	AU	:	10000	:	A	:	10000	:	AN	: 0
											LG=	2
43:	10000	:	AU	:	10000	:	A	:	10000	:	AN	: 0

Figure 52 : Résultats de la segmentation pour "IA"

```
**************************************************************************
NU PHONEME * MACRO-CLASSE * TREILLIS DE PHONEMES * NU DEBUT * NB PREL.
**************************************************************************
    1      *        3      *    EI    I     UN     *     2     *     54
    2      *        1      *    V     V     V      *     56    *     7
**************************************************************************
```

Figure 53 : Résultats du module de reconnaissance des voyelles pour "IA"

4	107	I	154	AI	180	EI
5	107	I	149	I	178	EI
6	133	I	154	AI	179	I
7	143	I	144	AI	166	EI
8	117	I	137	I	142	EI
9	108	I	122	I	159	EI
10	98	I	98	I	167	EI
11	79	I	115	I	192	EI
12	79	I	125	I	202	EI
13	85	I	153	I	219	EI
14	109	I	163	I	195	EI
15	150	I	178	EI	200	I
16	250	EI	252	I	264	I
17	200	EI	252	I	271	AI
18	157	EI	261	U	273	U
19	188	EI	270	EU	274	U
20	223	EI	289	U	291	EU
21	187	EI	251	U	269	EU
	21	18	111			
23	197	UN	217	O	282	EU
24	216	UN	273	ON	304	A
25	245	UN	287	A	304	ON
26	247	UN	281	A	286	ON
27	209	UN	244	ON	257	A
28	194	UN	205	UN	237	A
29	179	UN	194	A	200	ON
30	165	A	181	AN	186	UN
31	142	AN	173	A	189	A
32	157	AN	178	A	197	A
33	197	A	219	A	233	A
34	195	A	222	AN	235	A
35	183	AN	230	A	290	A
36	241	A	256	AN	387	AU
	36	14	112			

Figure 54 : Résultats de la segmentation par la variation spectrale

```
**********************************************************************
* NO DU PHONEME * MACRO-CLASSE * TREILLIS DE PHONEMES * NB DEBUT * NB PREL.*
**********************************************************************
*       1       *       3      *    I     EI          *     4     *    18   *
*       2       *       3      *    UN    A     ON    *    23     *    20   *
**********************************************************************
```

Figure 55 : Résultats finals

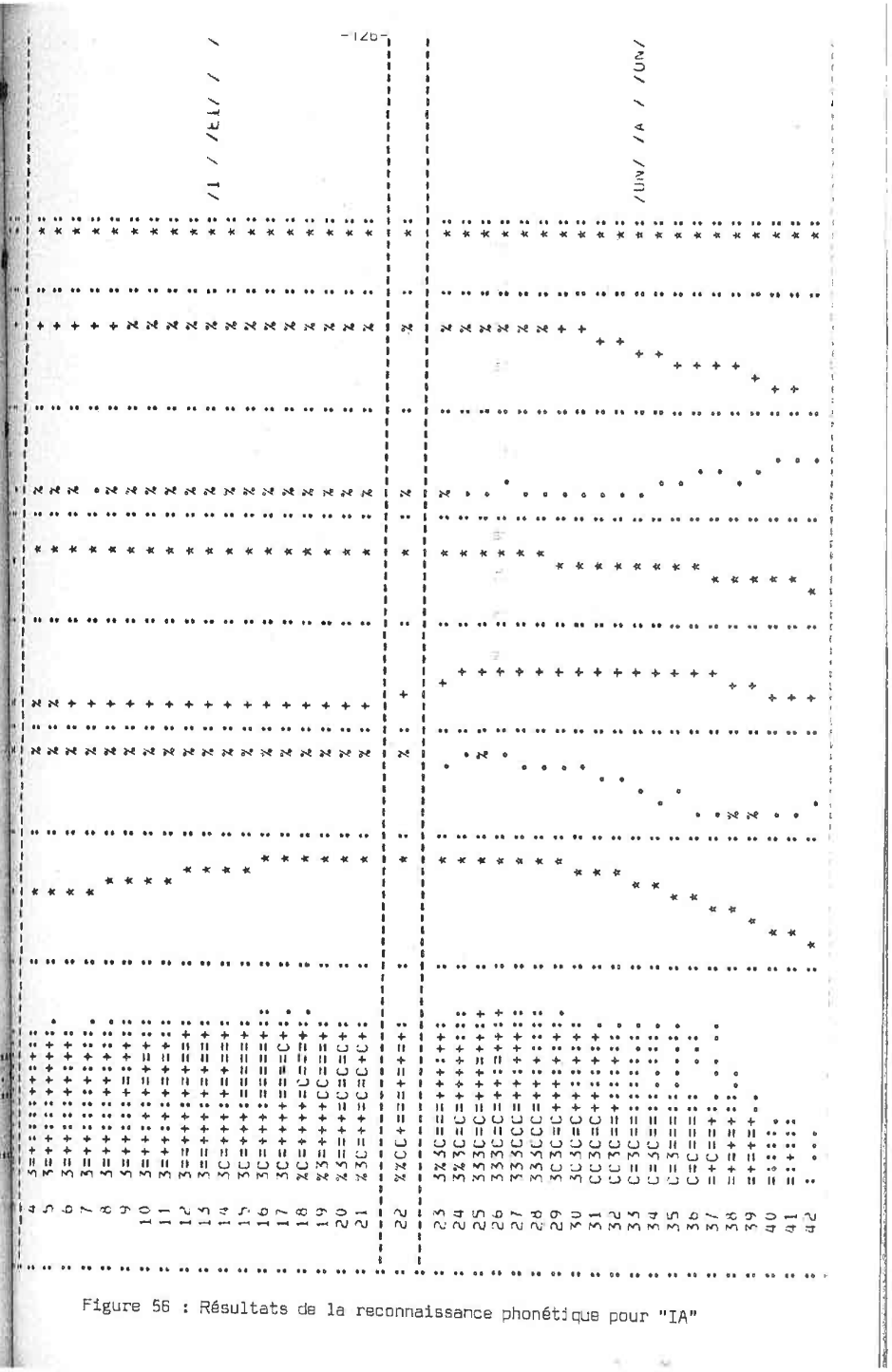


Figure 56 : Résultats de la reconnaissance phonétique pour "IA"

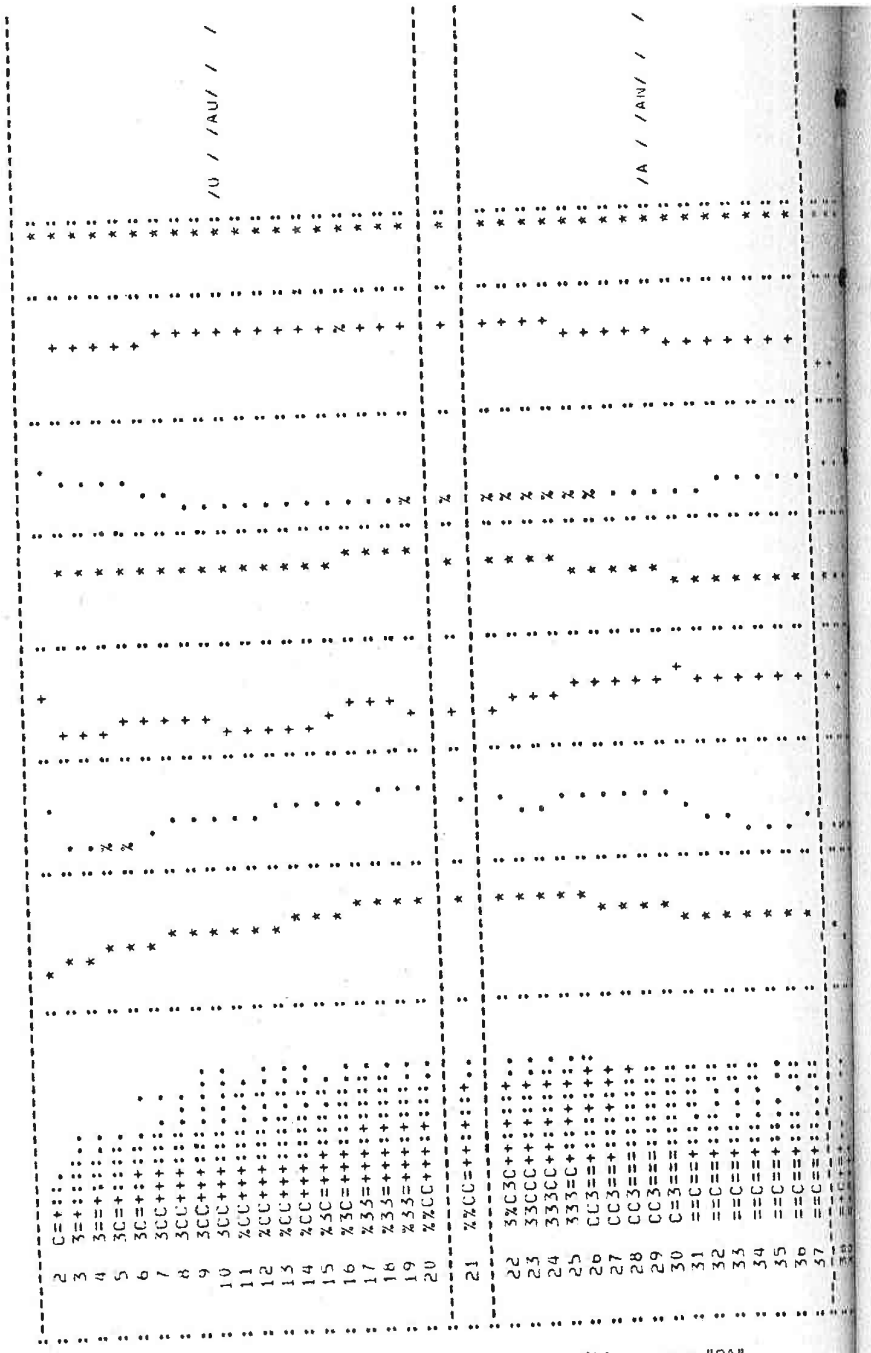


Figure 57 : Résultats de la reconnaissance phonétique pour "OA"

Les modules de reconnaissances des voyelles décrits précédemment, ne fournissent que les parties stables des voyelles. Dans le paragraphe suivant, nous décrirons une méthode pour éliminer les segments de transitions en debut et fin des voyelles.

IV. ELIMINATION DES ZONES DE TRANSITIONS DE DEBUT ET DE FIN DES VOYELLES

Sur le signal vocal, il apparaît toujours une zone de transition entre deux phonèmes consécutifs. Cette zone pose d'énormes problèmes lors de la segmentation et de l'identification des phonèmes, du fait que leurs frontières sont mal connues. Les modules de reconnaissance des voyelles par la courbe d'énergie et par la méthode spectrale, ne détectent que les zones stables des voyelles. Cela peut s'expliquer par deux raisons :

- la première, c'est que les références ne sont calculées, pendant la phase d'apprentissage, que sur les parties stables des voyelles.

- la deuxième raison est que les zones de transitions en début et fin des voyelles ne présentent pas les caractéristiques des zones stables adjacentes.

Sur l'exemple des figures 58, 59 et 60, il apparaît des zones voisées ne représentant pas des phonèmes.

Les zones de transitions, en début et fin de la voyelle sont supprimées à partir de la concavité de la courbe d'énergie fournie par la dérivée seconde d'un polynôme du second degré, obtenu par la méthode des moindres carrés sur cinq échantillons successifs [BELL-78-a] [BELL-78-B].

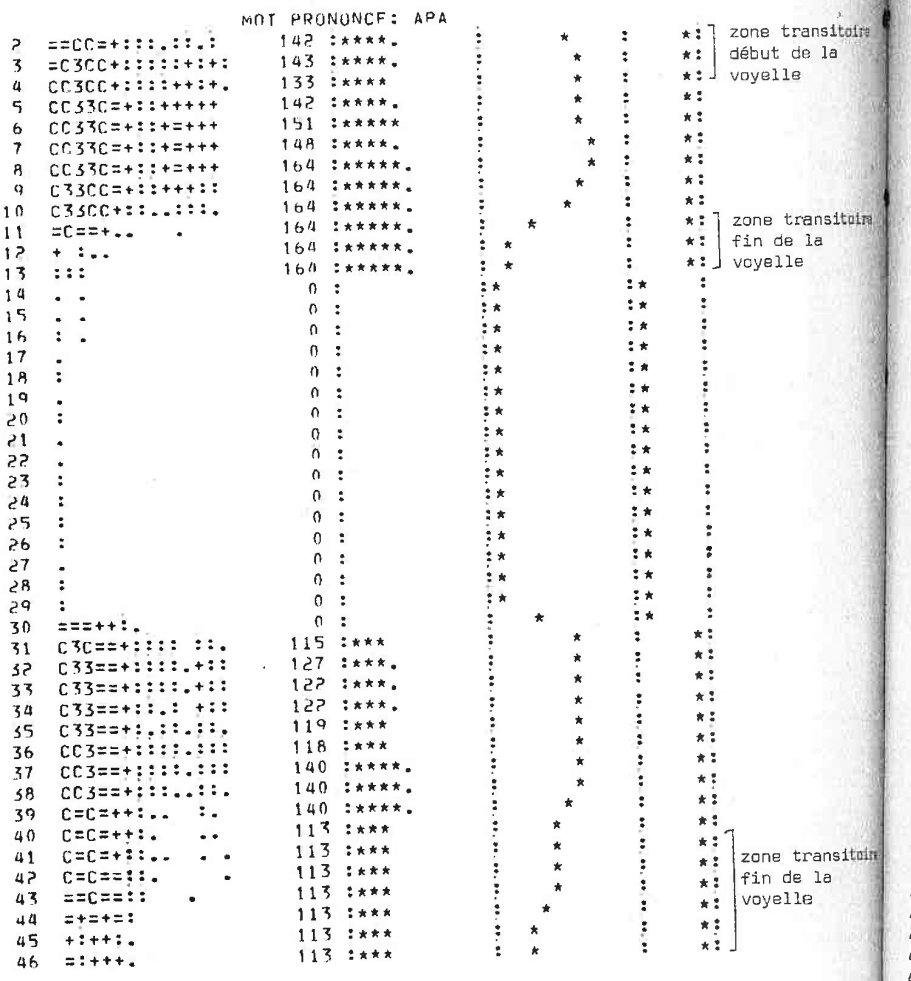


Figure 58 : Sonagramme du mot "APA"

	ENER	TYP	PTIC	VOT	VSPE	DTST1	PHO1	DTST2	PHO2	DTST3	PHO3	
2:	754:	5	142:	1	0:	429	AN	436	ON	510	A	0
3:	867:	5	143:	1	49:	263	AN	351	AN	364	ON	0
4:	886:	5	133:	1	64:	265	AN	297	AN	311	A	0
										LG= 1		
5:	920:	5	142:	1	51:	226	AN	258	A	275	A	?
6:	943:	5	151:	1	19:	217	A	217	AN	234	A	?
7:	962:	5	148:	1	13:	205	A	209	AN	224	A	?
8:	975:	5	164:	1	39:	181	A	199	AN	210	A	?
9:	940:	5	164:	1	127:	142	A	184	AN	185	A	?
10:	829:	5	164:	1	372:	178	AN	207	A	219	A	?
										LG= 2		
11:	457:	5	164:	1	307:	287	ON	410	AN	449	A	0
12:	150:	5	164:	1	79:	700	ON	825	AN	829	OU	0
13:	139:	5	164:	1	68:	678	ON	775	AN	799	OU	0
14:	71:	?	0:	0	16:	10000	ON	10000	AN	10000	OU	0
15:	55:	?	0:	0	11:	10000	ON	10000	AN	10000	OU	0
16:	66:	?	0:	0	21:	10000	ON	10000	AN	10000	OU	0
17:	45:	?	0:	0	28:	10000	ON	10000	AN	10000	OU	0
18:	53:	?	0:	0	18:	10000	ON	10000	AN	10000	OU	0
19:	35:	?	0:	0	8:	10000	ON	10000	AN	10000	OU	0
20:	33:	?	0:	0	7:	10000	ON	10000	AN	10000	OU	0
21:	26:	?	0:	0	5:	10000	ON	10000	AN	10000	OU	0
22:	31:	?	0:	0	2:	10000	ON	10000	AN	10000	OU	0
23:	33:	?	0:	0	1:	10000	ON	10000	AN	10000	OU	0
24:	32:	?	0:	0	1:	10000	ON	10000	AN	10000	OU	0
25:	33:	?	0:	0	0:	10000	ON	10000	AN	10000	OU	0
26:	33:	?	0:	0	4:	10000	ON	10000	AN	10000	OU	0
27:	29:	?	0:	0	7:	10000	ON	10000	AN	10000	OU	0
28:	36:	?	0:	0	11:	10000	ON	10000	AN	10000	OU	0
29:	47:	?	0:	0	356:	10000	ON	10000	AN	10000	OU	0
30:	403:	?	0:	0	347:	10000	ON	10000	AN	10000	OU	0
										LG= 3		
31:	750:	5	115:	1	66:	150	A	218	A	240	A	?
32:	814:	5	127:	1	29:	181	A	194	AN	197	A	?
33:	797:	5	122:	1	35:	152	A	220	A	230	A	?
34:	766:	5	122:	1	69:	153	A	269	A	273	A	?
35:	747:	5	119:	1	54:	139	A	201	A	217	A	?
36:	789:	5	118:	1	9:	175	A	215	A	235	A	?
37:	792:	5	140:	1	48:	189	A	215	A	233	A	?
38:	748:	5	140:	1	135:	132	A	238	A	249	AN	?
39:	613:	5	140:	1	72:	164	ON	220	A	235	AN	?
40:	553:	5	113:	1	77:	185	ON	279	A	302	AN	?
										LG= 4		
41:	550:	5	113:	1	41:	231	ON	335	A	350	AN	0
42:	531:	5	113:	1	65:	250	ON	344	A	381	AN	0
43:	506:	5	113:	1	134:	285	ON	373	A	410	AN	0
44:	372:	5	113:	1	98:	466	ON	570	A	595	AN	0
45:	274:	5	113:	1	17:	561	ON	692	AN	699	A	0
										LG= 5		
46:	285:	5	113:	1	0:	555	ON	684	AN	694	ON	0

zone transitoire
fin de la voyelle

Figure 59 : Résultats de la segmentation du mot "APA"

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES	NO DEBUT	NB PREL.
1	1	V V V	2	3
2	3	A AN	5	6
3	1	V V V	11	3
4	2	N N N	14	17
5	3	A	31	10
6	1	V V V	41	5

zones transitoires (début et fin des voyelles)

Figure 60 : Reconnaissance des voyelles

IV.1. FONCTION DE SEGMENTATION

Désignons par E_i l'énergie dans les basses fréquences du prélèvement numéro i des sorties du vocoder. Dans notre système, nous prenons l'énergie dans la bande 500 - 1200 Hz.

Soit P le polynôme d'ajustement au sens des moindres carrés, de degré 2, centrés sur la valeur i et construits à partir des cinq points d'abscisse i - 4, i - 2, i, i + 2, i + 4 pour les valeurs de E.

Pour chaque point i du signal vocal, nous allons calculer la dérivée seconde de ce polynôme d'ajustement au point i. On étend le domaine de définition au delà des extrémités en posant :

$$E_i = 0 \quad \text{pour}$$

$$i \in [-4, -3, -2, -1, N+1, N+2, N+3, N+4]$$

IV.2. CALCUL DU POLYNOME D'AJUSTEMENT

La valeur de la dérivée seconde de ce polynôme d'ajustement nous indique la nature de la concavité de la courbe d'énergie dans les basses fréquences.

Si la valeur de cette dérivée au point E_i est positive nous dirons que la parabole d'ajustement est concave, par contre si elle est négative nous dirons qu'elle est convexe.

Le changement de signe d'une valeur E_i à la suivante nous indique la possibilité d'une frontière entre deux phonèmes.

Nous allons considérer le polynome

$$P(x) = a_0 + a_1 x + a_2 x^2$$

ajustant au sens des moindres carrés les valeurs expérimentales E_{-2}, E_{-1}, E_0, E_1, E_2. Par définition nous cherchons les coefficients a_0, a_1 et a_2 sachant :

$$(1) \qquad \text{MIN} \left(\sum_{i=1}^{5} (E_i - a_0 - a_1 x_i - a_2 x_i^2)^2 \right)$$

pour cela nous disposons de deux méthodes :

i) la méthode géométrique

ii) la méthode analytique.

Développons par exemple, la méthode analytique :

annulons la dérivée de l'équation (1) par rapport à chaque variable.

$$\begin{cases} \dfrac{\partial}{\partial a_0} (1) = 0 \\ \dfrac{\partial}{\partial a_1} (1) = 0 \\ \dfrac{\partial}{\partial a_2} (1) = 0 \end{cases}$$

Nous obtenons :

$$\begin{cases} \sum_{i=1}^{5} \left| E_i - a_0 - a_1 x_i - a_2 x_i^2 \right| = 0 \\ \sum_{i=1}^{5} x_i \left| E_i - a_0 - a_1 x_i - a_2 x_i^2 \right| = 0 \\ \sum_{i=1}^{5} x_i^2 \left| E_i - a_0 - a_1 x_i - a_2 x_i^2 \right| = 0 \end{cases}$$

le système peut s'écrire :

$$(2) \begin{cases} \sum_{i=1}^{5} E_i = a_0 \sum_{i=1}^{5} 1 + a_1 \sum_{i=1}^{5} x_i + a_2 \sum_{i=1}^{5} x_i^2 \\ \sum_{i=1}^{5} x_i E_i = a_0 \sum_{i=1}^{5} x_i + a_1 \sum_{i=1}^{5} x_i^2 + a_2 \sum_{i=1}^{5} x_i^3 \\ \sum_{i=1}^{5} x_i^2 E_i = a_0 \sum_{i=1}^{5} x_i^2 + a_1 \sum_{i=1}^{5} x_i^3 + a_2 \sum_{i=1}^{5} x_i^4 \end{cases}$$

En posant

$$M = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \end{pmatrix}$$

le premier membre du système (2) peut s'écrire sous forme matricielle :

$$M^T . E$$

où M^T est la transposée de M

et

$$E = \begin{pmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{pmatrix}$$

Le second membre du système (2) peut s'écrire aussi sous forme matricielle :

$$M^T . M . a \qquad \text{avec} \qquad a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}$$

les coefficients du polynôme P(x).

d'où l'équation

$$M^T . M . a = M^T . E \qquad (3)$$

Nous allons ramener les points E_{i-4}, E_{i-2}, E_i, E_{i+2}, E_{i+4} en variables réduites à -2, -1, 0, 1, 2. Avec les variables réduites, la matrice M peut s'écrire :

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 2 & 4 \\ 1 & -2 & 4 \end{pmatrix}$$

l'équation (3) s'écrit

$$(4) \quad \begin{pmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & 2 & -2 \\ 0 & 1 & 1 & 4 & 4 \end{pmatrix} \begin{pmatrix} E_0 \\ E_1 \\ E_{-1} \\ E_2 \\ E_{-2} \end{pmatrix}$$

Le seul coefficient qui nous intéresse est celui de x^2 intervenant dans la dérivée seconde $P'' = 2\,a_2$.

Sa valeur fournie par la résolution du système (4) est :

$$a_2 = \frac{1}{14}\,(2\,(E_2 + E_{-2} - E_0) - E_1 - E_{-1})$$

Remarque :

Le choix des points i - 4, i - 2, i, i + 2, i + 4 plutôt que i - 2, i - 1, i, i + 1, i + 2 fait que le lissage obtenu tient moins compte des petites variations d'amplitude et prévilégie les grandes variations.

Donc pour chaque prélèvement de 10 ms; nous calculons :

$$B_i = 2\,(E_{i+4} + E_{i-4} - E_i) - E_{i+2} - E_{i-2}$$

où E_i est l'énergie dans les basses fréquences du prélèvement i.

Pour éviter que les valeurs de B oscillent rapidement autour de zéro, nous effectuons un lissage particulier sur B de la mamière suivante :

A l'intérieur d'un domaine où B conserve le même signe, on calcule la somme X des valeurs B_i :

$$X = \sum_{i=k}^{l} B_i$$

Dans ce calcul on inclut les intervalles isolés B_j tels que :

$$(B_j * B_{j-1} < 0) \quad \text{et} \quad (B_j * B_{j+1} < 0)$$

Puis, nous remplaçons chaque valeur B_i du domaine par $C_i = X$. Le graphe de C est celui d'une fonction en escalier (figure 62). Le résultat de l'élimination des zones voisées en début et fin des voyelles pour le mot "APA" est indiqué sur la figure 61, celui du mot "*chapeau*" est sur les figures 63 et 64.

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES					NO DEBUT	NB PREL.
1	3	A	AN				2	10
2	2	N	N	N			14	17
3	3	A					31	15

Figure 61 : Elimination des zones voisées au début et fin des voyelles (N indique une zone sourde)

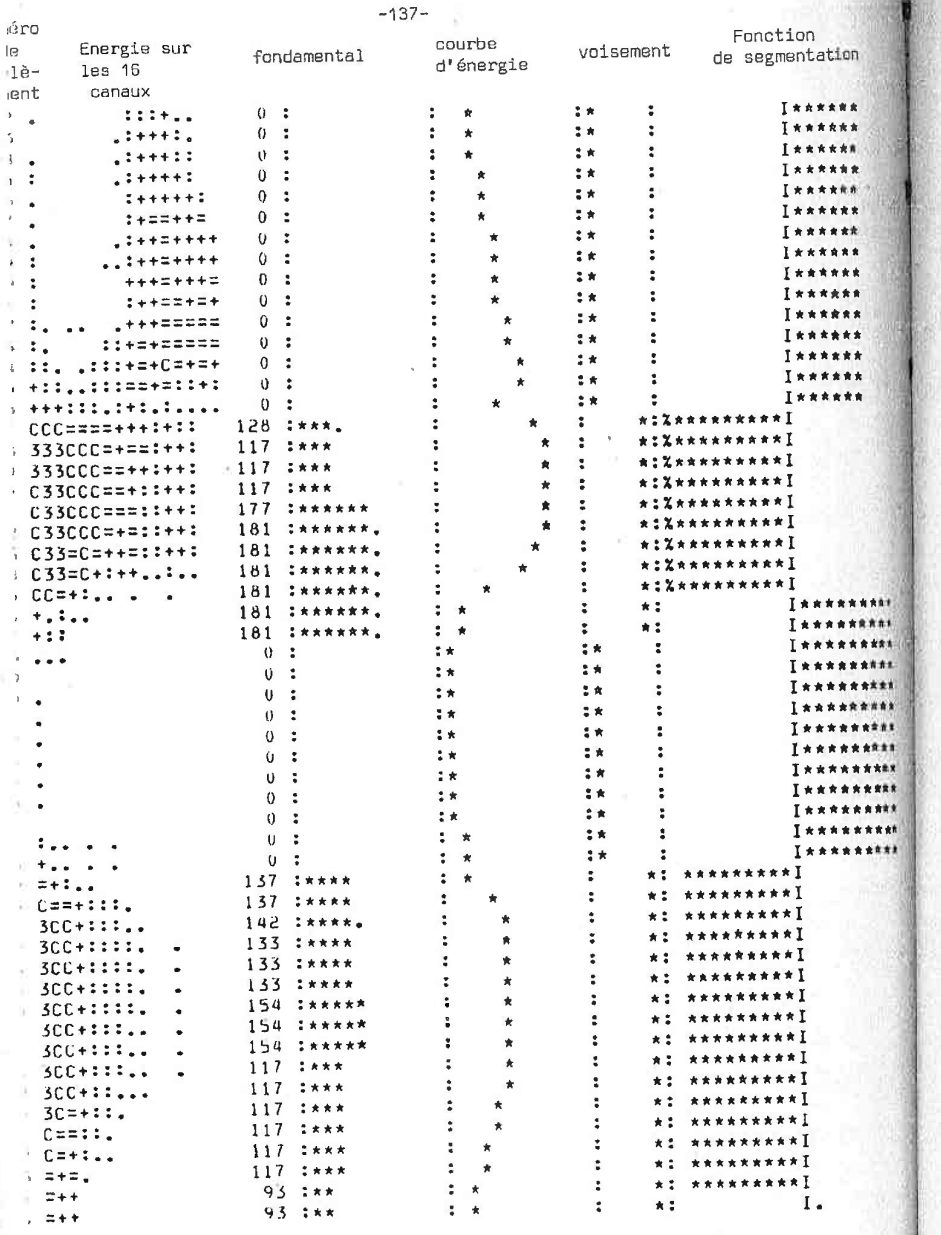
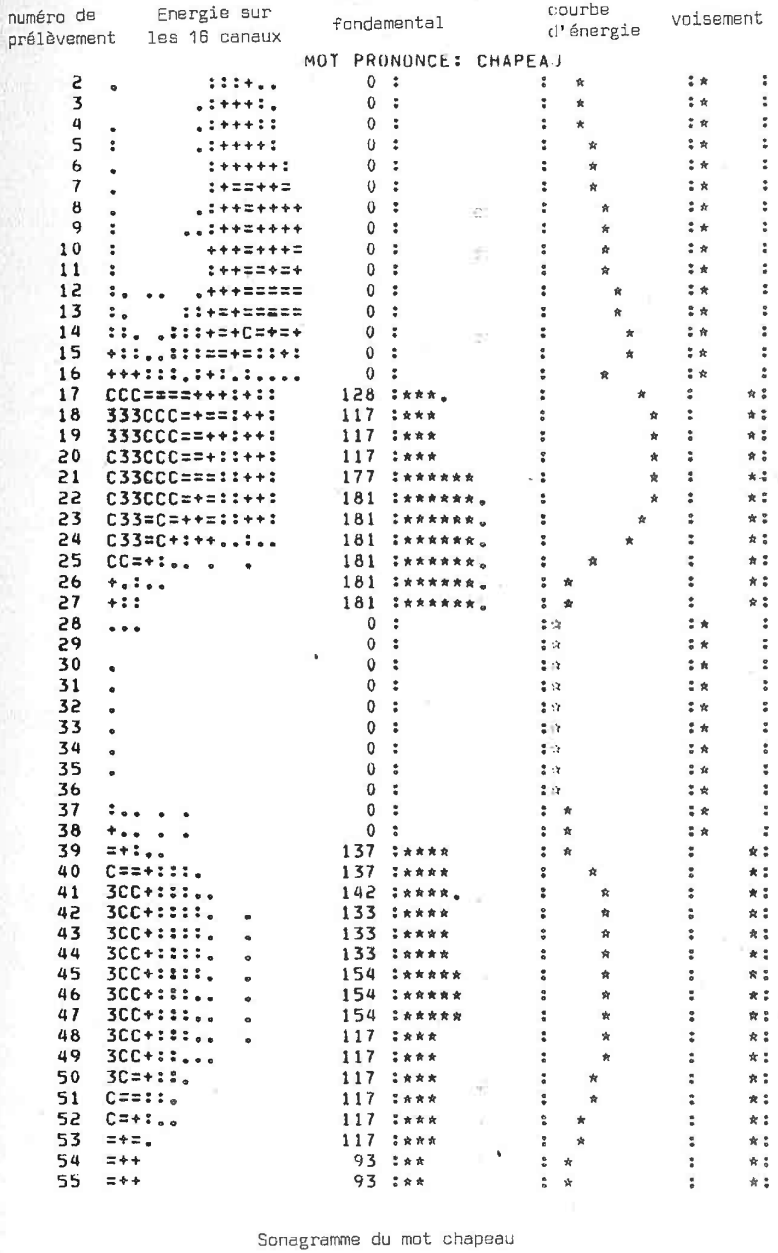


Figure 62 : Sonagramme du mot "*chapeau*" et fonction de segmentation



Sonagramme du mot chapeau

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES				NO DEBUT	NB PRE
1	2	N	N	N		2	15
	3	A	UN			17	8
	1	V	V	V		25	3
2	2	N	N	N		28	11
	3	AU	OU			41	9
	1	V	V	V		50	5

Figure 63 : Résultats de la reconnaissance des voyelles pour le mot "chapeau"

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES				NO DEBUT	NB PRE
1	2	N	N	N		2	15
	3	A	UN			17	9
2	2	N	N	N		28	11
	3	AU	OU			41	14

Figure 64 : Elimination des zones au début et fin des voyelles dans le mot "chapeau"

CHAPITRE VI

RECONNAISSANCE DES CONSONNES VOISEES

CHAPITRE VI

RECONNAISSANCE DES CONSONNES VOISEES

I. INTRODUCTION

Dans le chapitre précédent, nous avons segmenté le signal vocal en trois classes :

. les voyelles

. les consonnes voisées

. les consonnes non voisées

puis nous avons identifié les segments appartenant à la classe des voyelles. Il s'agit dans ce chapitre d'étudier les segments appartenant à la classe des consonnes voisées.

On peut regrouper les consonnes voisées en quatre classes :

- fricatives voisées /v/ /z/ /J/
- plosives voisées /b/ /d/ /g/
- liquides /l/ /r/
- consonnes nasales /m/ /n/ /ɲ/

Sachant que dans un segment, appartenant à la classe des consonnes voisées, peuvent exister plusieurs consonnes, deux types de problèmes se posent :

. la segmentation

. l'identification.

II. SEGMENTATION DES ZONES DES CONSONNES VOISEES

Dans un segment appartenant à la classe des consonnes voisées on peut trouver plusieurs consonnes voisées, comme par exemple dans la phrase "*Madame Boucher*" où figure la consonne /m/ en contact avec /b/ (figure 65).

D'où la nécessité d'une fonction de segmentation adaptée pour extraire les phonèmes dans le segment étudié.

La segmentation, en phonèmes, des zones appartenant à la classe des consonnes voisées est réalisée à partir de la fonction de variation spectrale, VSP ; entre deux prélèvements successifs :

$$VSP = \sum_{i=1}^{NB} | \text{canal}(i, j) - \text{canal}(i, j-1) |$$

La segmentation se fait suivant l'algorithme :

Pour chaque prélèvement appartenant au segment des consonnes voisées

- calcul de VSP

Si VSP > S alors

marque de segmentation

fsi

Fin pour

où S est un seuil fixe.

Sachant qu'on connaît le début et la fin des segments des consonnes voisées, il s'agit maintenant de les identifier. L'identification, des consonnes voisées, se compose de deux étapes :

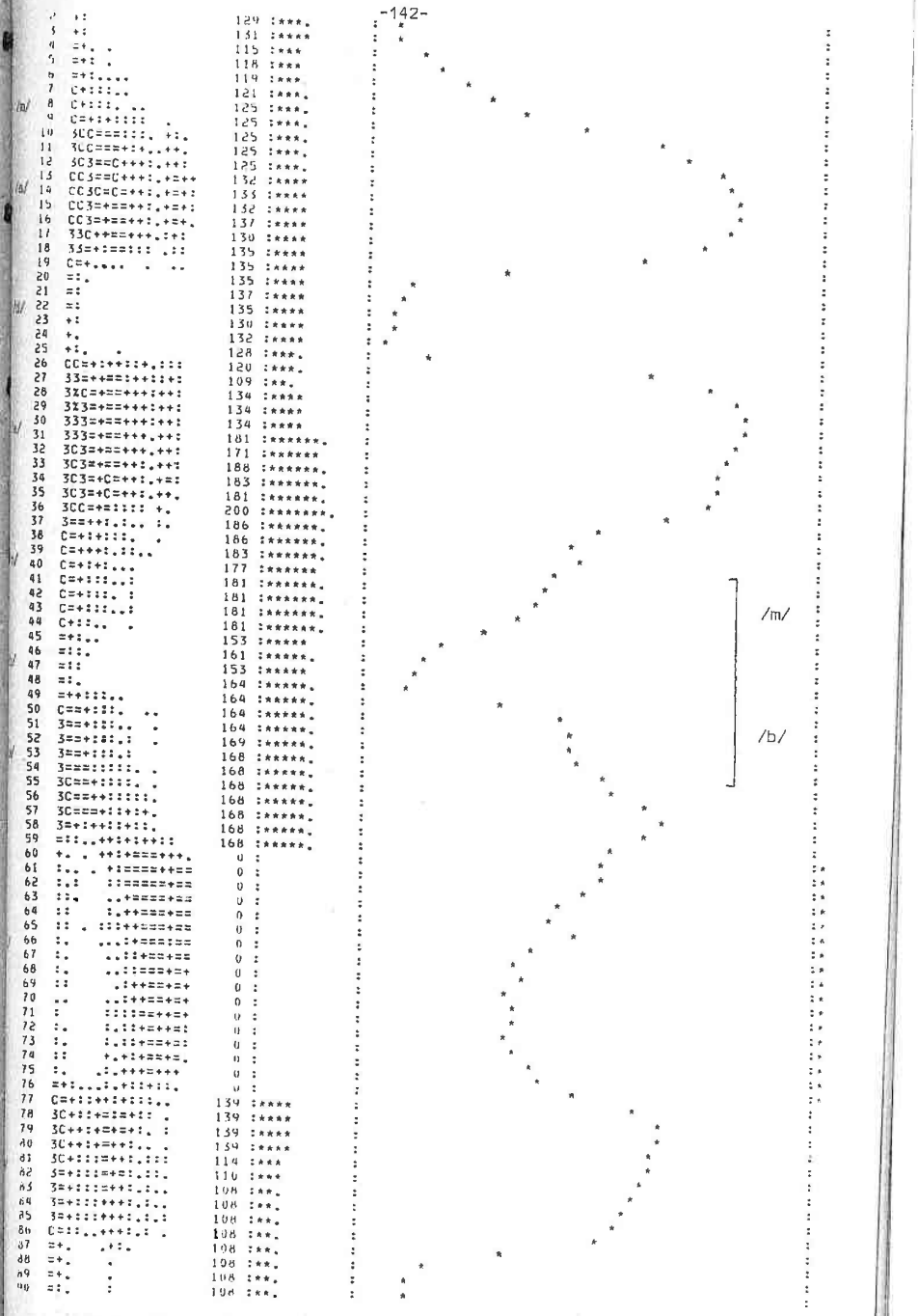


Figure 65 : Sonagramme de la phrase "*Madame boucher*"

i) étiquetage de chaque prélèvement du segment

ii) identification du segment à partir de ces étiquettes.

III. ETIQUETAGE DES ECHANTILLONS DES SEGMENTS DES CONSONNES VOISEES

De la description des consonnes voisées, dans le chapitre II, on peut déduire que certains phonèmes sont plus faciles à reconnaître que d'autres, du fait de la forme de leurs spectres ainsi que de leurs dépendances contextuelles. Les spectres de certaines consonnes, aisément reconnaissables, ont une forme bien caractéristique et sont peu dépendants du contexte, par contre les autres dépendent beaucoup de celui-ci (figure 11 et figure 12). Suivant ce critère, on peut regrouper les consonnes voisées en deux classes :

i) les consonnes peu dépendantes du contexte :

/j/,/z/ et /b/ /d/ /g/

ii) les consonnes dépendantes du contexte :

/l/ /r/ , /m/ /n/ et /v/.

En tenant compte de cette classification, les phones retenus à chaque échantillon sont obtenus à partir du graphe de la figure 66, où

HF représente l'énergie dans les hautes fréquences,

BC représente le centre de gravité de la distribution d'énergie dans le spectre, obtenu par la formule :

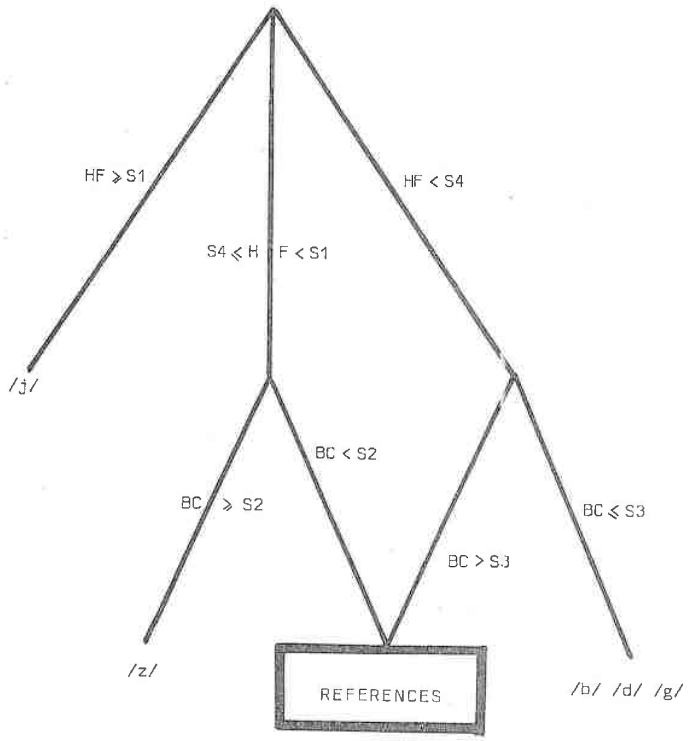


Figure 66 : Reconnaissance des consonnes voisées

$$BC = \frac{\sum_{j=1}^{NB} J * canal(j, i)}{\sum_{j=1}^{NB} canal(j, i)}$$

i étant le numéro du prélèvement courant.

Dans le cas où le prélèvement à étiqueter ne présente pas les caractéristiques des plosives voisées, ni de /j/ /z/ alors on fait appel au dictionnaire constitué par les liquides, les nasales, la consonne /v/ et des formes particulières de /j/,/z/ et /b/ /d/ /g/.

Les résultats de la reconnaissance des voyelles ainsi que l'élimination des zones de début et fin des voyelles pour la phrase *"Madame Boucher"* sont indiqués respectivement sur les figures 67 et 68.

Les résultats de la segmentation et de l'étiquetage des prélèvements des segments des consonnes voisées, sont décrits pour cette phrase sur la figure 69.

IV. IDENTIFICATION DES CONSONNES VOISEES

L'identification se fait par un processus de décision majoritaire parmi les phones affectés à chaque prélèvement du segment.

Le traitement d'identification des consonnes voisées est identique à celui des voyelles. Pour chaque consonne voisée, on calcule, à partir des étiquettes retenues pour chaque échatillon, son taux d'occurrence

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	1	V	V	V	2	7
2	3	ON	A		9	9
3	1	V	V	V	18	8
4	3	ON	AI		26	11
5	1	V	V	V	37	14
6	3	OU	AU		51	8
7	2	N	N	N	60	17
8	3	EI	U	I	77	10
9	1	V	V	V	87	3

Figure 67 : Reconnaissance des voyelles dans le groupe *"Madame Boucher"*

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL
1	1	V	V	V	2	7
2	3	ON	A		9	10
3	1	V	V	V	19	7
4	3	ON	AI		26	11
5	1	V	V	V	37	13
6	3	OU	AU		50	9
7	2	N	N	N	60	17
8	3	EI	U	I	77	12

Figure 68 : Elimination des zones en début et fin des voyelles

```
   ENER TYP PITC VOI VSPE DIST1  PHO1 DIST2  PHO2 DIST3  PHO3
37: 582: 0 : 186: 0 : 160:  197 : L  :  208 : N  :  221 : N  :
38: 517: 0 : 186: 0 :  81:  132 : M  :  205 : Z  :  206 : M  :
39: 519: 0 : 183: 0 :  72:  145 : L  :  152 : N  :  155 : N  :
40: 461: 0 : 177: 0 :  64:   99 : M  :  143 : M  :  145 : N  :
41: 434: 0 : 181: 0 :  35:   82 : M  :  134 : M  :  172 : N  :
42: 426: 0 : 181: 0 :  20:   94 : M  :  140 : M  :  160 : N  :
43: 398: 0 : 181: 0 :  38:  100 : M  :  109 : N  :  118 : M  :
                                                    LG=  1
44: 314: 0 : 181: 0 :  84:    0 : B  :  109 : B  :  118 : B  :
45: 221: 0 : 153: 0 :  93:    0 : B  :  109 : B  :  118 : B  :
46: 176: 0 : 161: 0 :  45:    0 : B  :  109 : B  :  118 : B  :
47: 153: 0 : 153: 0 :  23:    0 : B  :  109 : B  :  118 : B  :
48: 136: 0 : 164: 0 :  17:    0 : B  :  109 : B  :  118 : B  :
                                                    LG=  2
49: 336: 0 : 164: 0 : 200:    0 : B  :  109 : B  :  118 : B  :
50: 471: 0 : 164: 0 : 135:  186 : M  :  221 : N  :  229 : Z  :
****************** RECONNAISSANCE DES SEGMENTS VOISES
```

Figure 69 : Semgentation et étiquetage des zones voisées.

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES	NO DEBUT	NB PREL.
1	9	B	2	6
2	3	UN A	9	10
3	9	B	20	6
4	3	UN AI	26	11
5	9	M N	38	6
6	9	B	45	4
7	3	OU AU	50	9
8	4	CH	62	14
9	3	EI U I	77	12

Figure 70 : Reconnaissance phonétique pour la phrase "Madame Boucher"

Remarque : "B" représente la classe des plosives voisées /b, d, g/

dans le segment. Le treillis identifiant le segment est composé des consonnes ayant les plus grands taux par ordre décroissant. Les résultats de la reconnaissance phonétique pour la phrase *"Madame Boucher"* sont représentés sur la figure 70.

V. REDUCTION DU TREILLIS

En comparant les valeurs des scores une à une, on peut négliger certains phonèmes du treillis par rapport à d'autres. En effet, si le score de la trosième consonne dans le treillis est négligeable par rapport à celui de la deuxième alors le segment sera identifié uniquement par les deux premières consonnes du treillis.

De même, si le score de la deuxième consonne du treillis est négligeable par rapport à celui de la première alors le segment sera identifié uniquement par la première consonne du treillis (le même processus est appliqué aux voyelles).

Remarque :

Pour corriger certaines erreurs dues à la sursegmentation, nous comparons les résultats de la reconnaissance des segments des consonnes voisées consécutifs. Si deux segments de consonnes voisées se trouvent en contact et que les premiers éléments de leurs treillis sont identiques alors on fusionne les deux segments. A partir des éléments des deux anciens treillis, le nouveau segment sera identifié en prenant les trois consonnes voisées ayant les plus grands scores.

VI. SCHEMA GENERAL DU MODULE DE RECONNAISSANCE DES CONSONNES VOISEES

Pour rendre l'exposé plus clair, nous avons supposé que les modules de segmentation, d'étiquetage des prélèvements et d'identification des segments des consonnes voisées, se déroulent séquentiellement. En pratique, ces traitements se déroulent simultanément, d'une part pour optimiser le temps de calcul ainsi que la mémoire occupée dans le calculateur, d'autre part, pour corriger certaines erreurs dues à la sursegmentation.

Le schéma général du module de reconnaissance des consonnes voisées est représenté sur la figure 71 où

VSP : indique la variation spectrale entre deux prélèvements consécutifs.

S : seuil de segmentation (fixe)

IDF.SEG : identification du segment courant

FUSION : fusion éventuelle du segment courant et le précédent et identification du nouveau segment.

IDF.PRE : identification du prélèvement.

MAJ SCORES : mise à jour des scores des consonnes voisées.

LG : nombre de prélèvements à reconnaître dans le segment des consonnes voisées.

Sur la figure 72 on a représenté le sonagramme segmenté de la phrase "*Madame Boucher*".

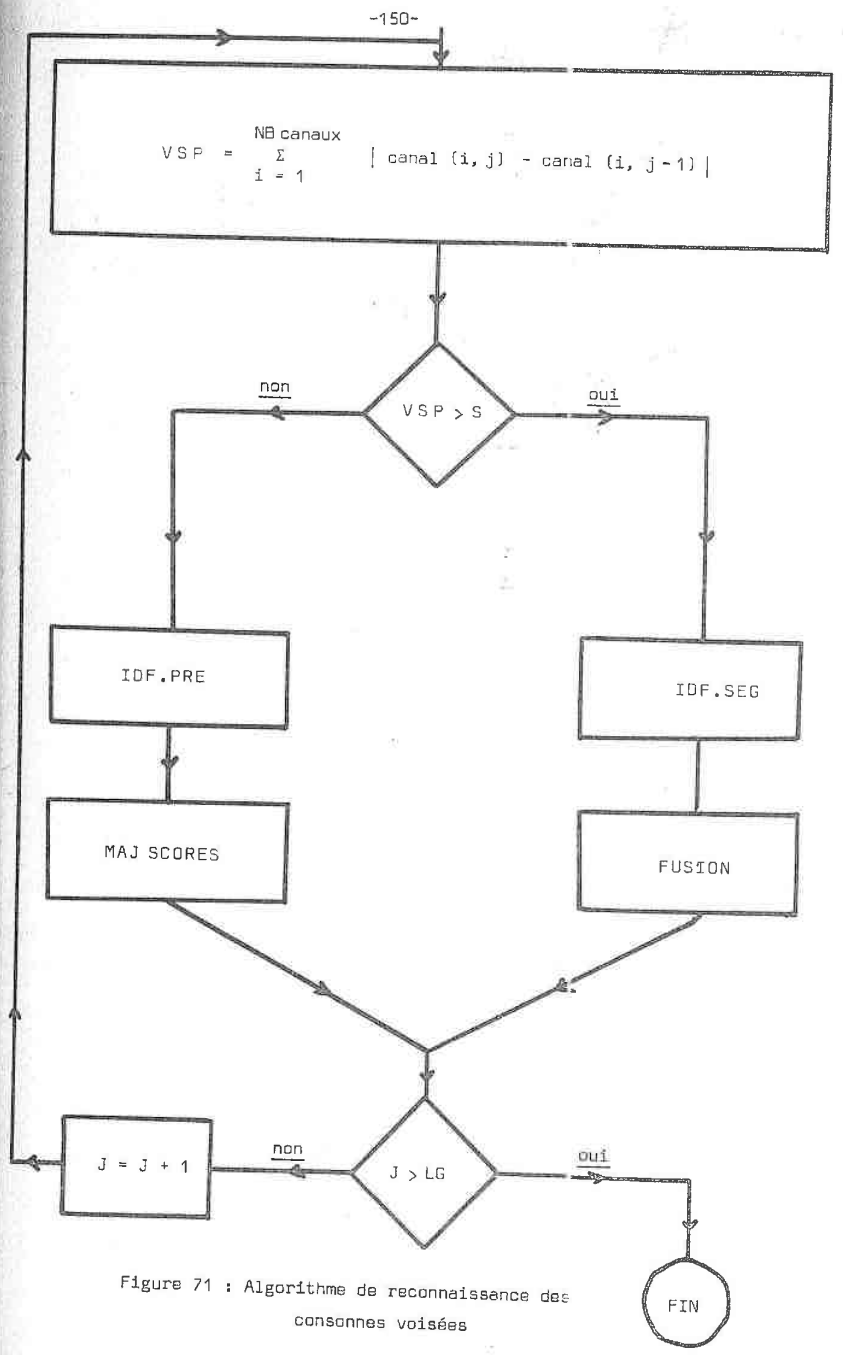
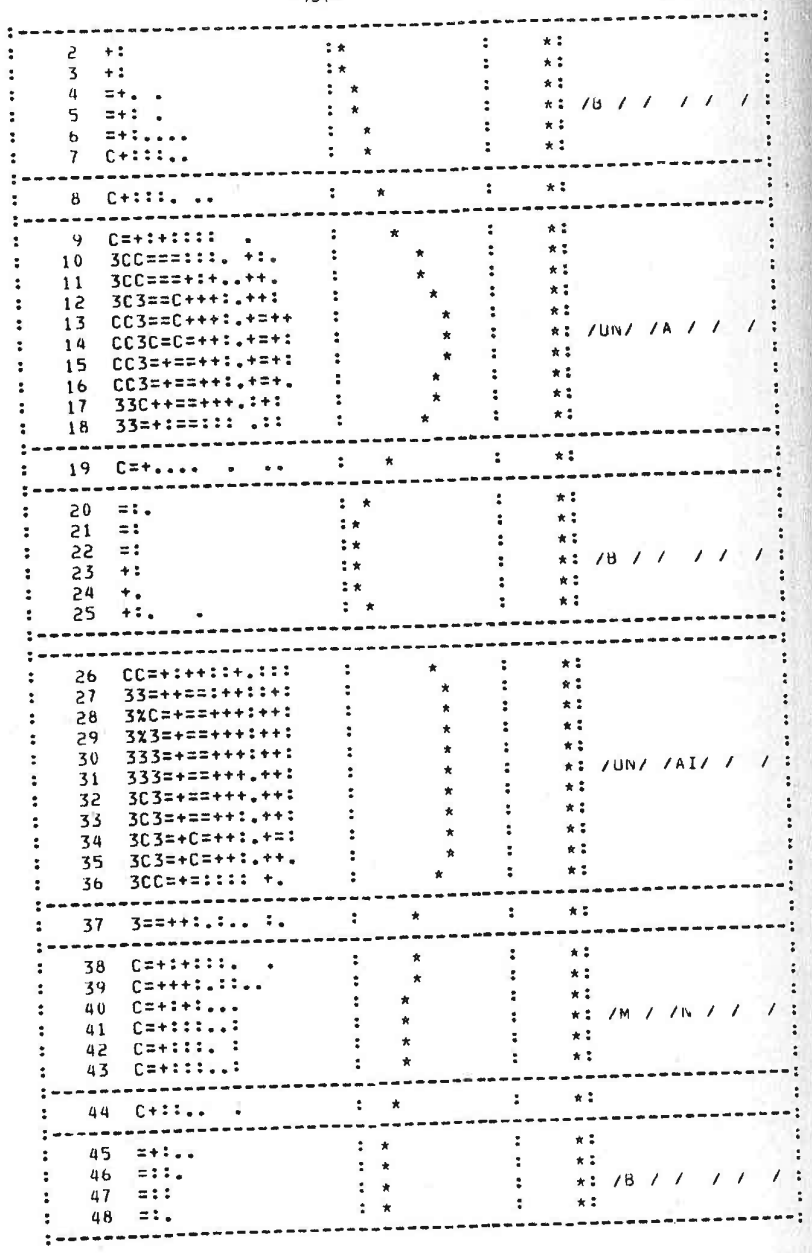


Figure 71 : Algorithme de reconnaissance des consonnes voisées



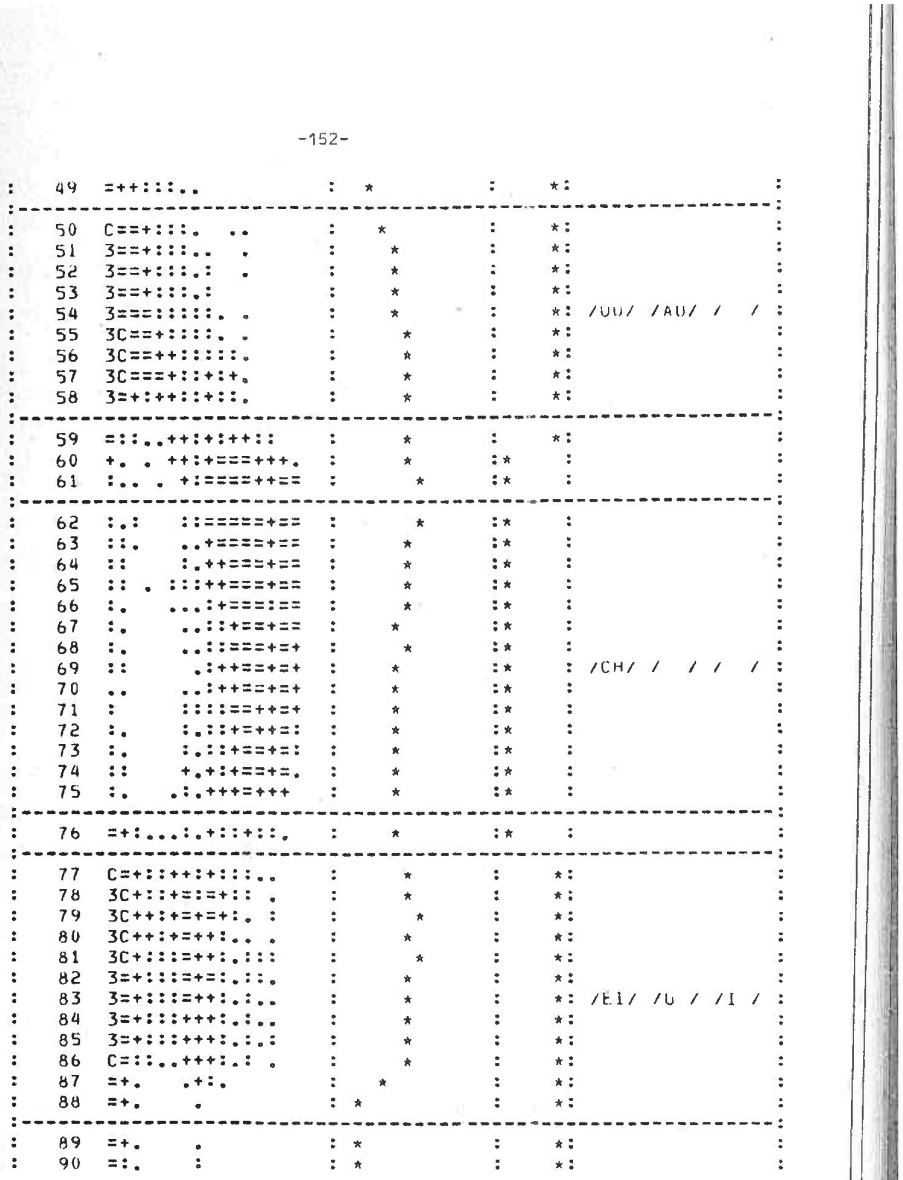


Figure 72 : Résultats de la reconnaissance phonétique pour la phrase *"Madame boucher"*

CHAPITRE VII

RECONNAISSANCE DES CONSONNES SOURDES

CHAPITRE VII

RECONNAISSANCE DES CONSONNES SOURDES

I. INTRODUCTION

Les consonnes voisées sont distinguées des consonnes sourdes à partir des informations fournies par le détecteur de mélodie "Mélographe". Dans les chapitres précédents nous avons identifié les segments voisés. Il s'agit dans ce chapitre d'étudier les zones non voisées.

On peut regrouper les consonnes sourdes en deux classes :

i) les fricatives non voisées : /f/ /s/ /ʃ/

ii) les plosives non voisées : /p/ /t/ /k/.

Dans une zone sourde peuvent exister plusieurs consonnes non voisées. Prenons l'exemple du mot "*explosion*" (figure 73) où les trois consonnes /k/ /s/ /p/ sont en contact.

Le module de reconnaissance des consonnes sourdes se déroule à nouveau en deux étapes :

i) Segmentation

ii) Identification

trois consonnes non voisées consécutives /k/ /s/ /p/

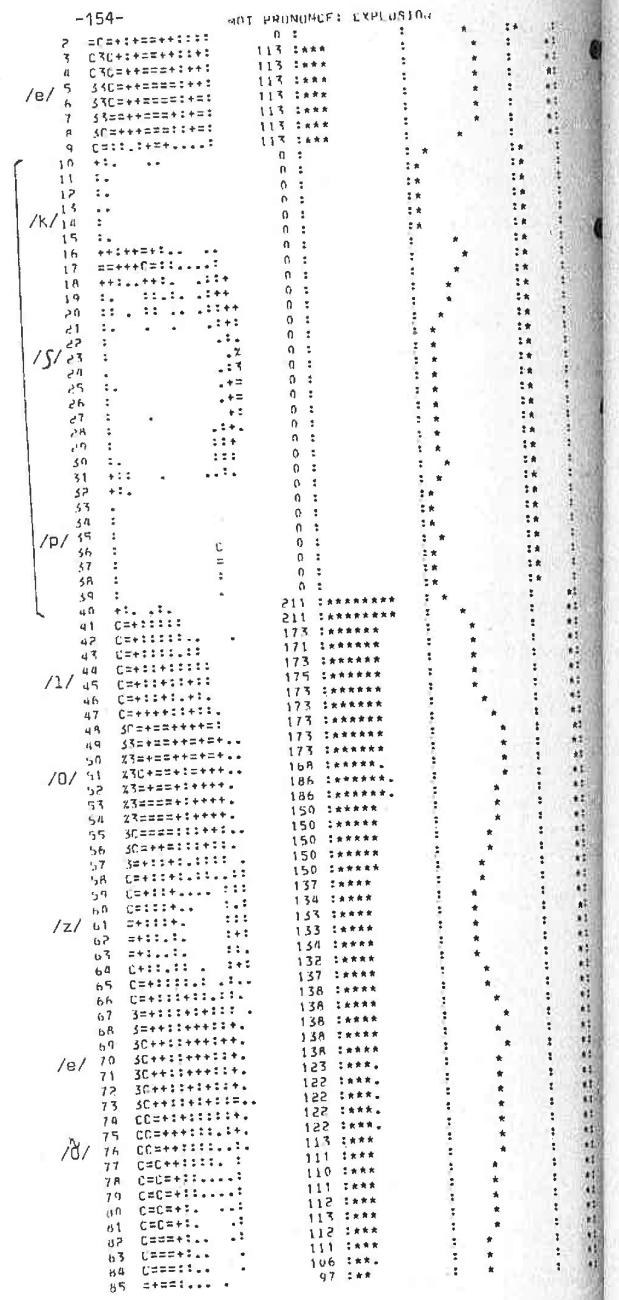


Figure 73 : Sonagramme du mot "*explosion*"

II. SEGMENTATION DES ZONES DES CONSONNES NON VOISEES

La segmentation des zones des consonnes sourdes se fait suivant l'algorithme :

Pour chaque échantillon non voisé

Si $|Var(t) - Var(t+1)| > S$ alors

marque de segmentation

fsi

fpour

où S est un seuil fixe et Var (t) est la variance de l'énergie de chaque canal autour de la valeur moyenne moy(t) :

$$Var(t) = \sum_{j=1}^{NB} (canal(j,t) - moy(t))^2$$

avec $moy(t) = \frac{Energie(t)}{NB}$

t étant le numéro du prélèvement courant.

Sachant qu'on connaît le début et la fin des segments des consonnes sourdes, il s'agit maintenant de les identifier. L'identification des consonnes non voisées s'effectue en deux étapes :

i) étiquetage de chaque prélèvement du segment.

ii) identification du segment à partir de ces étiquettes.

III. ETIQUETAGE DES PRELEVEMENTS DES CONSONNES NON VOISEES

A partir de l'énergie dans les hautes fréquences, nous détectons la consonne /ʃ/, pour localiser la consonne /s/ nous calculons l'énergie dans les hautes fréquences ainsi que la distribution d'énergie dans le spectre.

Les plosives sourdes sont localisées à partir de l'énergie dans les basses fréquences et de la distribution d'énergie dans le spectre.

En calculant certains paramètres fréquentiels nous étiquetons ainsi les prélèvements sans consulter les références phonétiques. Mais, dans les cas d'ambiguïté, on fait appel aux références des consonnes sourdes constituées par la consonne /f/, (dépendante beaucoup du contexte), ainsi qu'à quelques formes particulières de /ʃ/ , /s/ , /p/ /t/ /k/, et de la consonne non voisée /r/ (assourdie par le contexte).

L'étiquetage des prélèvements des consonnes non voisées se fait suivant le schéma de la figure 74 où

. HF représente l'énergie dans les hautes fréquences, dans le système, nous prenons l'énergie dans la bande 2400 - 5000.

. BC étant le centre de gravité de la distribution d'énergie dans le spectre, obtenu par la formule :

$$BC = \frac{\sum_{j=1}^{NB} J * Canal(j, i)}{\sum_{j=1}^{NB} Canal(j, i)}$$

i étant le numéro du prélèvement courant.

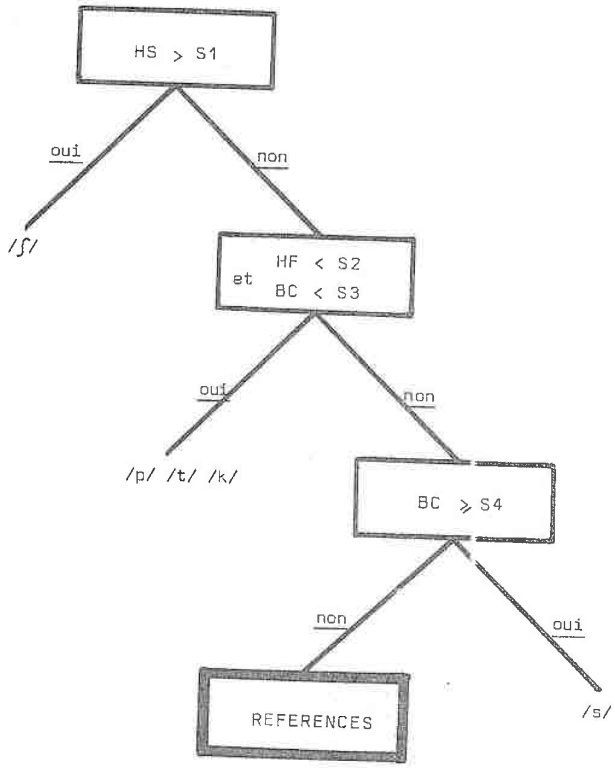


Figure 74 : Reconnaissance des consonnes sourdes

La figure 75 montre un exemple de segmentation et d'identification dans une zone non voisée du mot "EXPLOSION".

	ENER	TYP	PITC	VOI	VSPE	DIST1	PHO1	DIST2	PHO2	DIST3	PHO3	
11:	70:	0 :	0:	0 :	9:	122	: P :	122	: P :	166	: F :	1er segment : /k/
12:	63:	0 :	0:	0 :	0:	0	: P :	122	: P :	166	: P :	
13:	48:	0 :	0:	0 :	0:	0	: P :	122	: P :	166	: P :	
14:	37:	0 :	0:	0 :	0:	0	: P :	122	: P :	166	: P :	
15:	64:	0 :	0:	0 :	0:	0	: P :	122	: P :	166	: P :	
										LG=	1	
16:	527:	0 :	0:	0 :	104:	567	: F :	773	: S :	796	: S :	2ème segment : /s/
17:	686:	0 :	0:	0 :	75:	788	: F :	966	: CH :	976	: S :	
18:	543:	0 :	0:	0 :	127:	0	: S :	966	: S :	976	: S :	
19:	409:	0 :	0:	0 :	51:	0	: S :	966	: S :	976	: S :	
20:	435:	0 :	0:	0 :	41:	0	: S :	966	: S :	976	: S :	
21:	310:	0 :	0:	0 :	47:	0	: S :	966	: S :	976	: S :	
22:	147:	0 :	0:	0 :	74:	0	: S :	966	: S :	976	: S :	
										LG=	2	
23:	203:	0 :	0:	0 :	31:	167	: S :	214	: S :	248	: P :	3ème segment : /p/
24:	189:	0 :	0:	0 :	15:	0	: S :	214	: S :	248	: S :	
25:	187:	0 :	0:	0 :	14:	0	: S :	214	: S :	248	: S :	
26:	206:	0 :	0:	0 :	19:	0	: S :	214	: S :	248	: S :	
27:	169:	0 :	0:	0 :	25:	0	: S :	214	: S :	248	: S :	
28:	173:	0 :	0:	0 :	35:	0	: S :	214	: S :	248	: S :	
29:	173:	0 :	0:	0 :	26:	0	: S :	214	: S :	248	: S :	
30:	202:	0 :	0:	0 :	22:	0	: S :	214	: S :	248	: S :	
31:	264:	0 :	0:	0 :	59:	0	: S :	214	: S :	248	: S :	
32:	122:	0 :	0:	0 :	87:	190	: S :	201	: P :	205	: P :	
										LG=	2	
33:	27:	0 :	0:	0 :	20:	0	: P :	201	: P :	205	: P :	
34:	34:	0 :	0:	0 :	0:	0	: P :	201	: P :	205	: P :	
35:	33:	0 :	0:	0 :	0:	0	: P :	201	: P :	205	: P :	
										LG=	3	
36:	125:	0 :	0:	0 :	90:	0	: S :	201	: S :	205	: S :	
37:	106:	0 :	0:	0 :	17:	0	: S :	201	: S :	205	: S :	
38:	73:	0 :	0:	0 :	33:	84	: P :	88	: P :	170	: S :	
39:	60:	0 :	0:	0 :	15:	77	: P :	81	: P :	157	: S :	
40:	200:	0 :	211:	0 :	23:	77	: P :	81	: P :	157	: S :	
							↑ 1ère étiquette		↑ 2ème étiquette		↑ 3ème étiquette	

Figure 75 : Segmentation et étiquetage des zones sourdes

IV. IDENTIFICATION DES CONSONNES NON VOISEES

L'identification des consonnes sourdes se fait à partir des 3 étiquettes retenues pour chaque prélèvement du segment à reconnaître. Nous affectons le score 3 à la première étiquette, le score 2 pour la deuxième et le score 1 pour la troisième. Le treillis identifiant le segment est constitué des trois consonnes ayant le plus grand score, par ordre décroissant. Le résultat de la reconnaissance phonétique pour le mot "*explosion*" est indiqué sur les figures 76 et 77.

V. REDUCTION DU TREILLIS

Comme précédemment, on peut négliger certains phonèmes du treillis par rapport à d'autres, en comparant les scores entre eux.

Remarque :

Pour corriger certaines erreurs dues à la sursegmentation, nous comparons les résultats de la reconnaissance des segments des consonnes sourdes consécutifs. Si deux segments de consonnes non voisées se trouvent en contact et que les premiers éléments de leur treillis sont les mêmes alors on fusionne les deux segments.

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	3	EI	U	AI	2	6
2	2	N	N	N	10	30
3	1	V	V	V	40	4
4	3	U	O	AI	44	12
5	1	V	V	V	57	9
6	3	EI	AI	U	66	9
7	3	ON	A	OU	75	10

a) Reconnaissance des voyelles

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	3	EI	U	AI	2	6
2	5	P			11	5
3	4	S			18	11
4	5	P			33	3
5	1	V	V	V	40	4
6	3	U	O	AI	44	12
7	1	V	V	V	57	9
8	3	EI	AI	U	66	9
9	3	ON	A	OU	75	10

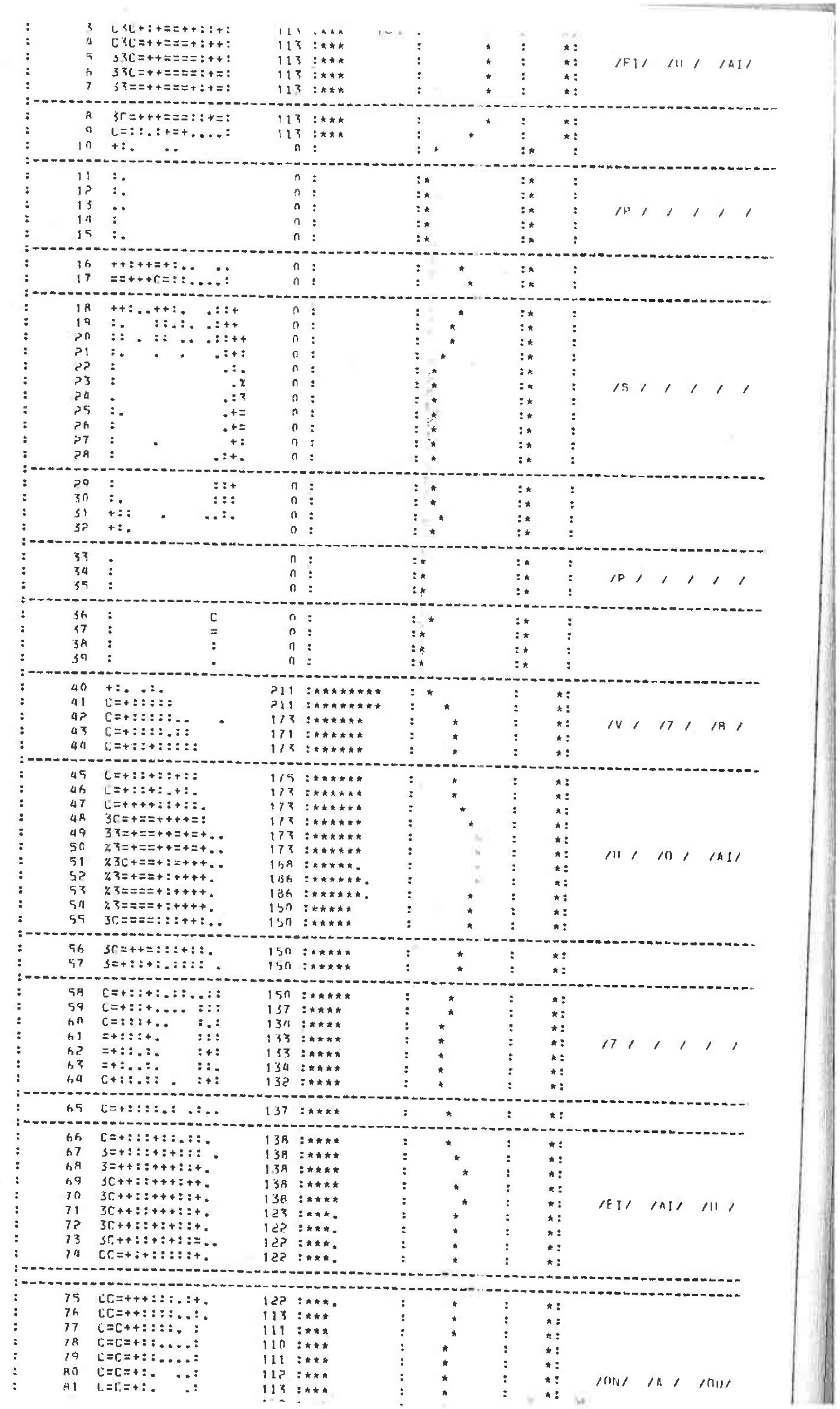
b) Reconnaissance des consonnes sourdes

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	3	EI	U	AI	2	6
2	5	P			11	5
3	4	S			18	11
4	5	P			33	3
5	3	V	Z	B	40	5
6	3	U	O	AI	44	12
7	3	Z			58	7
8	0	EI	AI	U	66	9
9	0	ON	A	OU	75	10

c) Reconnaissance des consonnes voisées

Figure 76 : Etapes du décodage phonétique du mot "*Explosion*" (cf. figure 77)

Remarque : le symbole "P" représente la classe des plosives sourdes /p,t,k/.



...gure 77 : ...iagramme ...gmenté du ...*plosion*"

VI. SCHEMA GENERAL DU MODULE DE RECONNAISSANCE DES CONSONNES NON VOISEES

Pour rendre l'exposé plus clair, nous avons supposé que les modules de segmentation, d'étiquetage des prélèvements et d'identification des segments des consonnes sourdes se déroulent séquentiellement. En pratique, ces traitements se déroulent simultanément, comme ceux correspondants à la reconnaissance des consonnes sonores (cf. paragraphe VI du chapitre précédent).

Le schéma général du module de reconnaissance des consonnes sourdes est représenté sur la figure 78, où

VAR : la variance de l'énergie de chaque canal autour de la valeur moyenne.

i : numéro du prélèvement courant.

IDF.SEG : identification du segment.

FUSION : fusion éventuelle du segment courant avec le précédent et identification du nouveau segment.

IDF.PRE : identification du prélèvement.

MAJSCORES : mise à jour des scores des consonnes sourdes.

LG : nombre de prélèvements à reconnaître dans le segment.

La reconnaissance phonétique pour la phrase "*est-ce que je peux*" est représentée sur la figure 80.

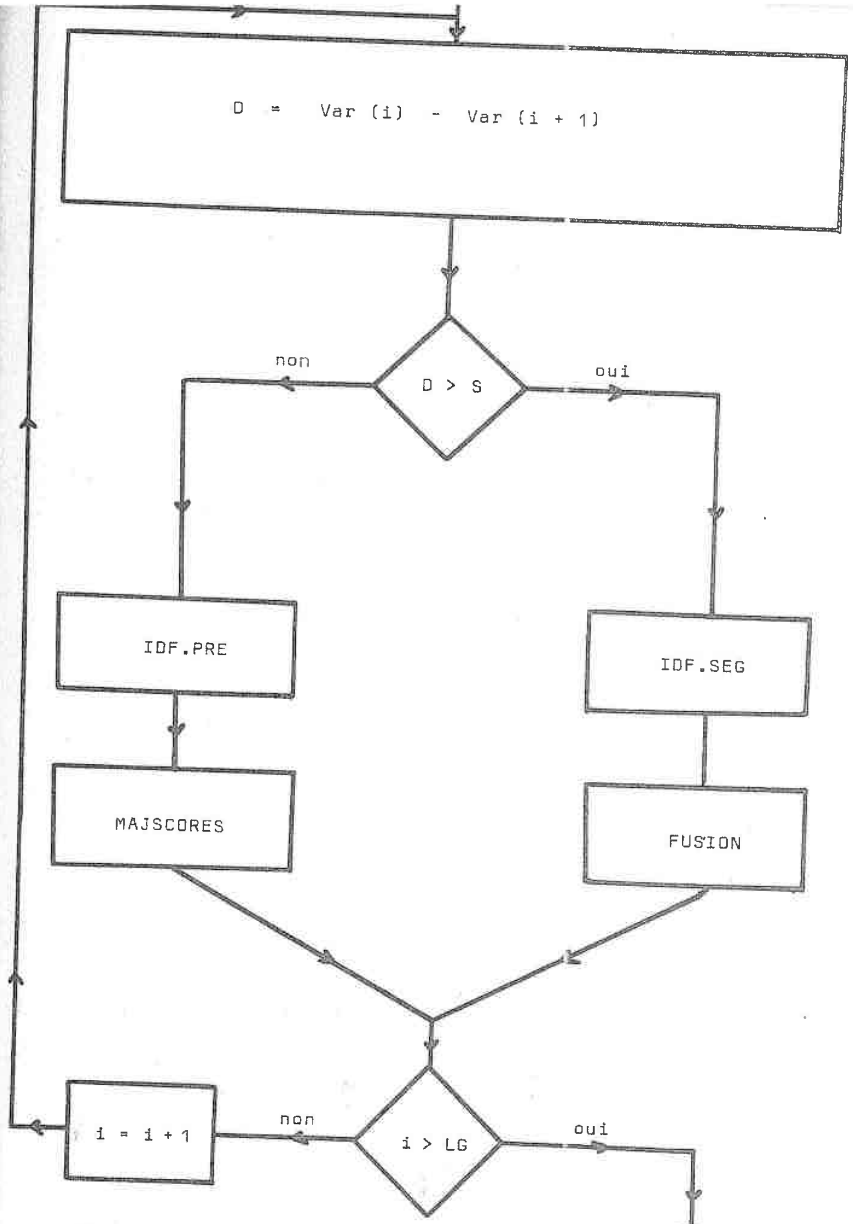


Figure 78 : Algorithme de reconnaissance des consonnes sourdes

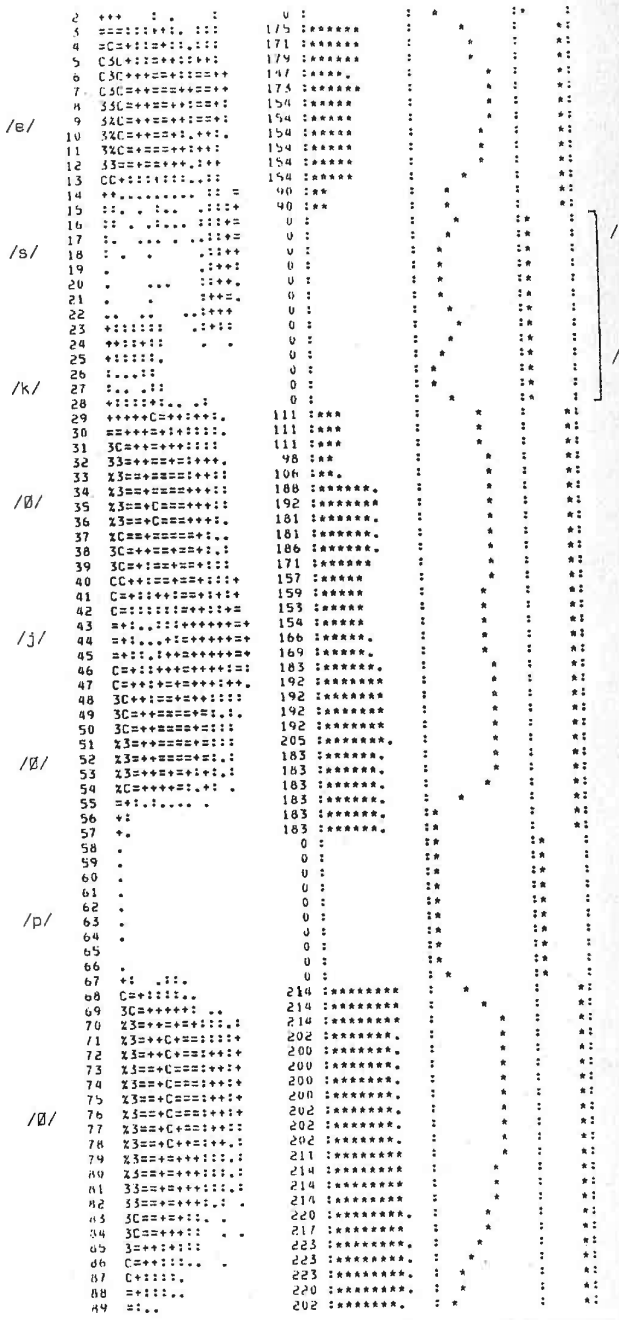


Figure 79 : Sonagramme de la phrase "*Est-ce que je peux*"

DU PHONEME	* MACRO-CLASSE *	TREILLIS DE PHONEMES				* NO DEBUT *	NB PREL.
1	* 3 *	UN	U	EI	*	2 *	10
2	* 1 *	V	V	V	*	13 *	3
3	* 2 *	N	N	N	*	16 *	13
4	* 3 *	EU	U		*	31 *	9
5	* 1 *	V	V	V	*	40 *	8
6	* 3 *	U	EU		*	48 *	7
7	* 1 *	V	V	V	*	55 *	3
8	* 2 *	N	N	N	*	58 *	10
9	* 3 *	EU			*	69 *	17
10	* 1 *	V	V	V	*	86 *	3

a) Reconnaissance des voyelles

DU PHONEME	* MACRO-CLASSE *	TREILLIS DE PHONEMES				* NO DEBUT *	NB PREL.
1	* 3 *	UN	U	EI	*	2 *	11
2	* 2 *	N	N	N	*	16 *	13
3	* 3 *	EU	U		*	31 *	10
4	* 1 *	V	V	V	*	41 *	7
5	* 3 *	U	EU		*	48 *	8
6	* 2 *	N	N	N	*	58 *	10
7	* 3 *	EU			*	69 *	19

b) Elimination de début et fin des voyelles

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	3	UN	U	EI	2	11
2	4	S			17	9
3	0	F	P		26	3
4	3	EU	U		31	10
5	1	V	V	V	41	7
6	3	U	EU		48	8
7	5	P			59	8
8	3	EU			69	19

c) Reconnaissance des consonnes non voisées

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	3	UN	U	EI	2	11
2	4	S			17	9
3	0	F	P		26	3
4	3	EU	U		31	10
5	9	J			45	4
6	3	U	EU		48	8
7	5	P			59	8
8	3	EU			69	19

d) Reconnaissance des consonnes voisées

Figure 80 : Etapes du décodage phonétique de la phrase
"Est-ce que je peux ..."

CHAPITRE VIII

LE POINT SUR LE SYSTEME

CHAPITRE VIII

LE POINT SUR LE SYSTEME

I. MISE EN OEUVRE DU PARALLELISME

La structure générale du système, ainsi que l'organisation de chaque étape permet d'envisager le traitement parallèle des différentes tâches (figure 81). Cette technique accélèrerait les traitements et permettrait ainsi un fonctionnement en temps réel.

De plus, en combinant les deux premières étapes de la figure 81 (l'acquisition et l'extraction des paramètres), nous optimisons le temps de réponse du système.

Nous n'avons pas actuellement envisagé la réalisation de cette technique, mais il est certain qu'un tel système est appelé à être implanté sur une architecture de type multi-micro-processeurs.

Notre système, pour l'instant, est implanté, en Fortran, sur un mini calculateur Mitra 125. Le temps de traitement sur cet ordinateur est du même ordre que le temps d'élocution de la phrase à reconnaître.

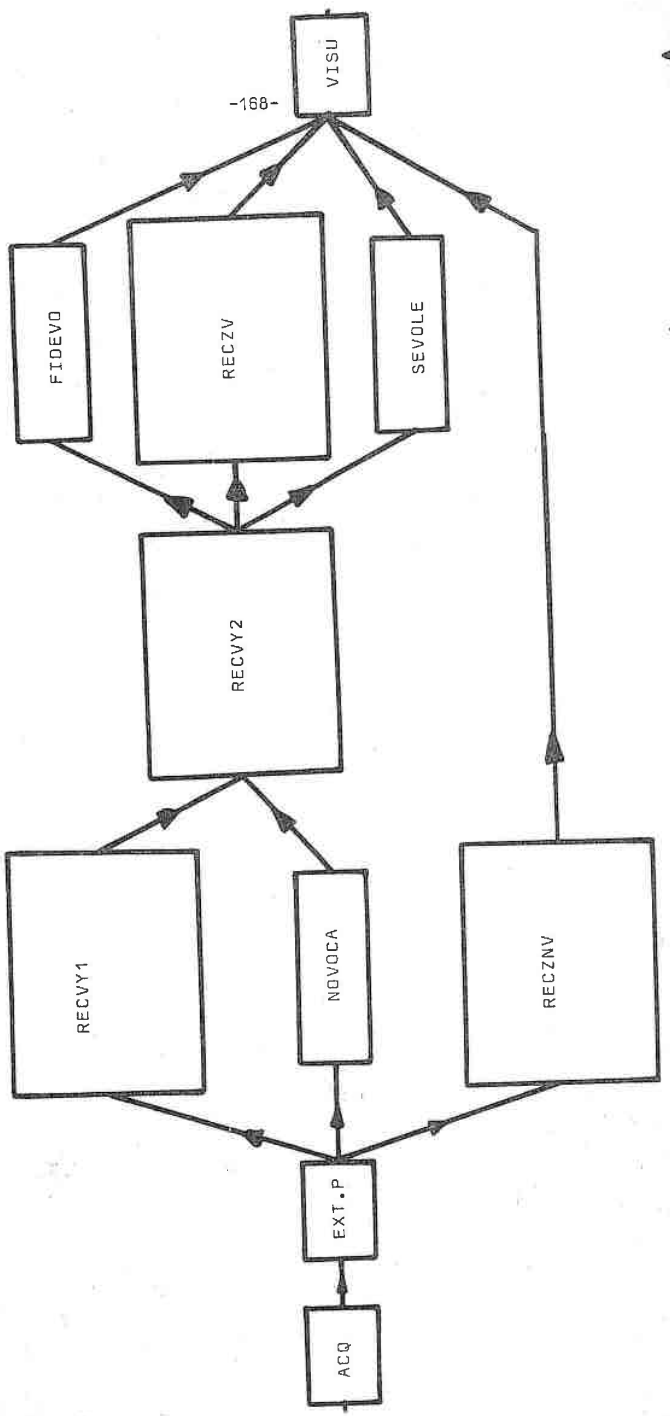


Figure 81 : Mise en oeuvre du parallélisme

ACQ : l'acquisition de la parole.

EXT.P : extraction des paramètres du signal vocal et segmentation de ce dernier en segments voisés et sourds.

RECVY1 : détection et identification des voyelles par la méthode spectrale.

FIDEVO : élimination des zones de transition en début et fin des voyelles.

RECZV : segmentation et identification des segments des consonnes voisées.

SEVOLE : segmentation des segments longs.

RECZNV : segmentation et identification des segments des consonnes non voisées.

VISU : visualisation des résultats de la reconnaissance phonétique.

NOVOCA : détection des noyaux vocaliques dans la courbe d'énergie.

RECVY2 : segmentation et identification des voyelles dans les segments voisés.

La fonction VISU permet, à la demande, de visualiser diverses informations :

- le spectrogramme,
- le spectrogramme segmenté et identifié,
- le spectrogramme segmenté par FTSEG,
- les résultats de la reconnaissance,
- les différentes étapes de la reconnaissance,

et tout cela soit sur écran, soit sur imprimante.

II. ETAPES DE LA RECONNAISSANCE PHONETIQUE SUR UN EXEMPLE

Le système de décodage phonétique procède schématiquement en 6 étapes successives.

1. Acquisition.
2. Extraction des paramètres.
3. Reconnaissance des voyelles par la méthode spectrale.
4. Reconnaissance des voyelles par la courbe d'énergie.
5. Reconnaissance des consonnes voisées.
6. Reconnaissance des consonnes sourdes.

Nous allons examiner les résultats de ces différentes étapes sur la phrase :

"Est-ce que je peux avoir Monsieur Boucher".

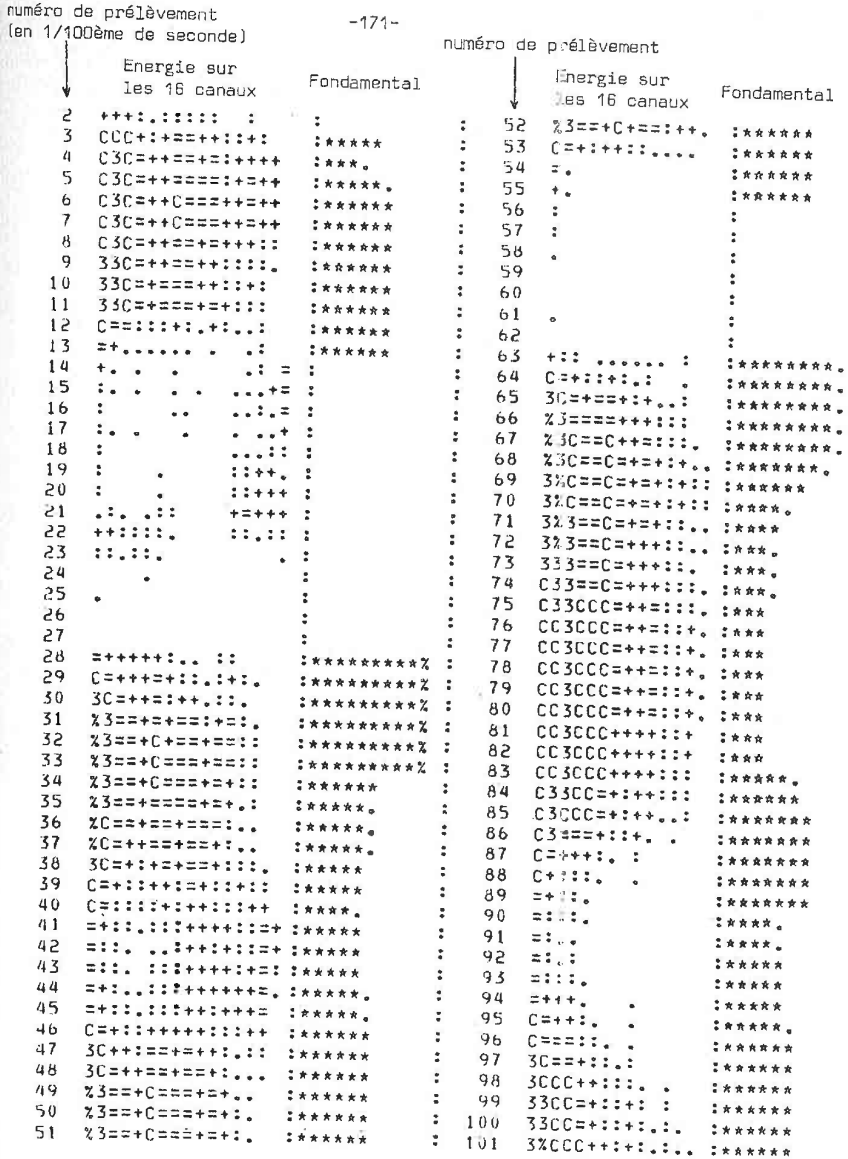
Après l'acquisition et l'extraction des paramètres, nous obtenons le sonagramme codifié de la phrase à reconnaître (figure 82).

A partir de ces différents paramètres (calculés sur les sorties du vocoder), de la courbe d'énergie et de la fonction de segmentation (segmentation par identification) nous localisons et identifions les voyelles sur le signal vocal (figure 83).

Les informations fournies par le détecteur de voisement permettent de séparer les consonnes voisées des non-voisées.

Nous reconnaissons les consonnes sourdes à partir (figure 84) :

- de l'énergie dans les hautes fréquences,
- du centre de gravité de la distribution d'énergie dans le spectre,



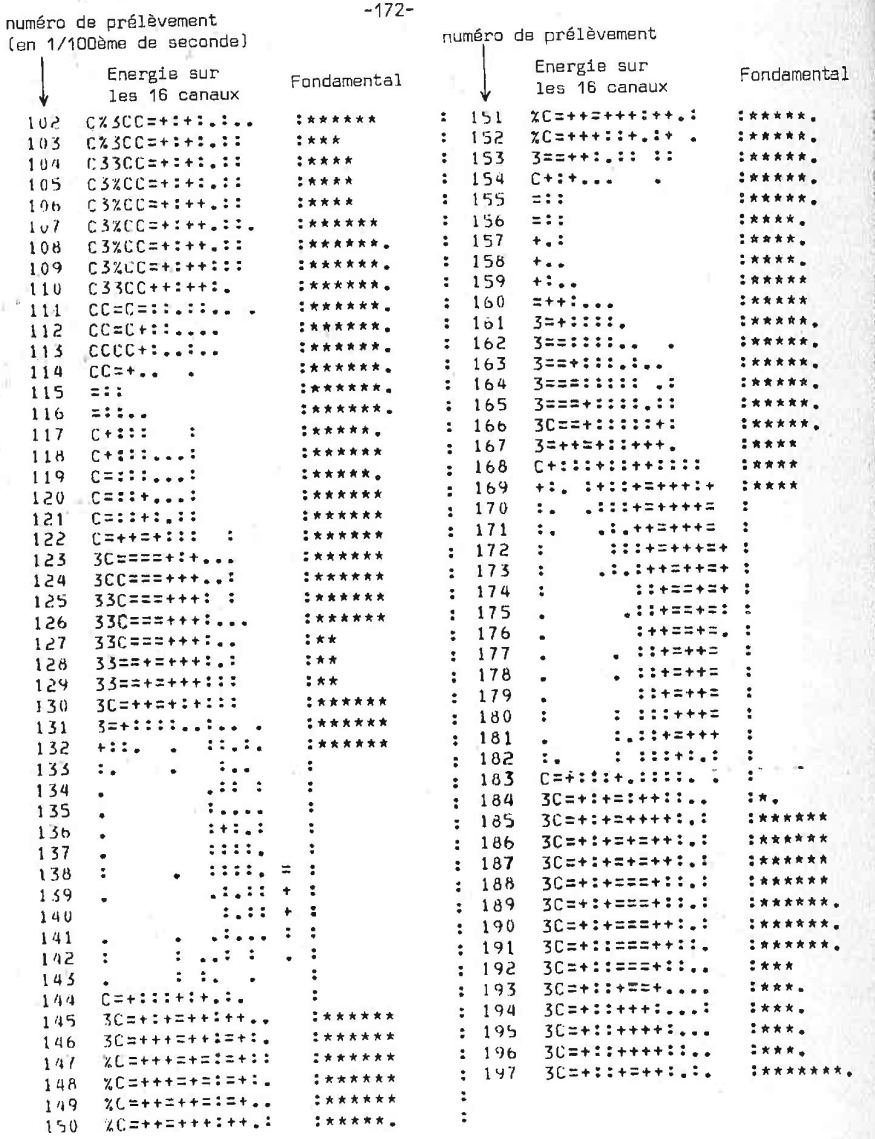


Figure 82 : Sonagramme de la phrase :

"Est-ce que je peux avoir Monsieur Boucher"

N DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	3	EI	O		2	9
2	2	N	N	N	14	14
3	3	EU	U		30	10
4	1	V	V	V	40	7
5	3	EU	U		47	7
6	2	N	N	N	56	7
7	3	EU		A	64	9
8	3	A			73	15
9	1	V	V	V	88	8
10	3	A			96	16
11	1	V	V	V	113	10
12	3	EU			123	7
13	2	N	N	N	133	12
14	3	U	AI	EU	145	10
15	1	V	V	V	155	6
16	3	AI	OU		161	8
17	2	N	N	N	170	14
18	3	AI	U	EI	184	13

Figure 83 : Reconnaissance des voyelles

N DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	NB PREL.
1	3	EI	O		2	9
2	4	S			15	4
3	5	P			24	4
4	3	EU	U		30	10
5	1	V	V	V	40	7
6	3	EU	U		47	7
7	5	P			57	6
8	3	EU		A	64	9
9	3	A			73	15
10	1	V	V	V	88	8
11	3	A			96	16
12	1	V	V	V	113	10
13	3	EU			123	7
14	4	S			134	10
15	3	U	AI	EU	145	10
16	1	V	V	V	155	6
17	3	AI	OU		161	8
18	4	CH			171	11
19	3	AI	U	EI	184	13

Figure 84 : Reconnaissance des consonnes sourdes

NO DU PHONEME	MACRO-CLASSE	TREILLIS DE PHONEMES			NO DEBUT	Nb PREL.
1	3	EI	O		2	9
2	4	S			15	4
3	5	P			24	4
4	3	EU	U		30	10
5	9	J			41	4
6	3	EU	U		47	7
7	5	P			57	6
8	3	EU		A	64	9
9	3	A			73	15
10	9	B			88	8
11	3	A			96	16
12	9	M	N		118	4
13	3	EU			123	7
14	4	S			134	10
15	3	U	AI	EU	145	10
16	9	B			155	5
17	3	AI	OU		161	8
18	4	CH			171	11
19	3	AI	U	EI	184	13

Figure 85 : Reconnaissance des consonnes voisées

- des formes de références,
- de la fonction de segmentation (variation de l'énergie de chaque canal autour de la valeur moyenne).

La reconnaissance des consonnes voisées est représentée sur la figure 85, elle se fait à partir :

- des formes de références,
- de la fonction de segmentation (variation spectrale entre deux prélèvements successifs),
- de l'énergie dans les hautes fréquences,
- du centre de gravité de la distribution d'énergie dans le spectre.

III. EXEMPLES DE DIALOGUE DANS LE SYSTEME MYRTILLE I

L'objet de notre travail a été essentiellement d'extraire, du signal vocal, les phonèmes qui seront utilisés par les traitements linguistiques dans un système de compréhension automatique de la parole continue. Pour pouvoir tester les choix effectués à ce niveau, le résultat de la reconnaissance phonétique est fourni au système de compréhension de langage MYRTILLE I [PIER-75] [PIER-81], dans le cadre de l'application actuelle, simulant un standard téléphonique automatique.

On donne ci-dessous, quelques exemples de transcription obtenue pour des phrases, lors d'un dialogue avec le système :

DEBUT DE SESSION

DEBUT DE COMMUNICATION

STANDARD AUTOMATIQUE DE MYRTILLE I. JE VOUS ECOUTE.

---> COR :

A IN
S P F
FI U AI
M N B
OU
A
M N B
AI FI EU
S
U EI FU
B
AI T FI
P
AN ON A

(Passez-moi Monsieur Dupont)

JE VOUS PASSE M. DUPONT.

VOUS AVEZ EN LIGNE M. DUPONT.

FIN DE COMMUNICATION

DEBUT DE COMMUNICATION

STANDARD AUTOMATIQUE MYRTILLE I. JE VOUS ECOUTE.

---> COR :

U	AI	FU
S		
EU	U	EI
B		
OU	AI	
CH		
EI	U	AI
S		
FI	AI	T
B		
OU	AU	
P		
N	R	M
AI	I	FI

(Monsieur Boucher, s'il-vous-plaît)

VOUS AVEZ BIEN DEMANDE M. BOUCHER

---> COR :

EI EU I

(oui)

VOUS AVEZ EN LIGNE M. BOUCHER

FIN DE COMMUNICATION

DEBUT DE COMMUNICATION

STANDARD AUTOMATIQUE DE MYRTILLE I. JE VOUS ECOUTE

---> COR :

O	OU	
R	N	Z
U	FI	AI
V	Z	B
O	UN	A
B		
A		
AU	AI	OU
S		CH
U	EI	FU
B		
FI	AI	U
R		
A	AN	

(Pourrais-je avoir Monsieur DURAND ?)

VOUS AVEZ BIEN DEMANDE M. DURAND ?

---> COR :

EI EU I

(oui)

LE POSTE DE M. DURAND EST OCCUPE. VOULEZ-VOUS PATIENTER ?

---> COR :

N	L	
AN	ON	A
M	N	
EI	EU	
V	Z	B
S		
I	AI	
J		
EU	I	
R	M	
A		
P		
U	AI	EI
L	N	
EU	U	
V	Z	B
AI	I	EI

(Non merci je rappellerai)

FIN DE COMMUNICATION

DEBUT DE COMMUNICATION

STANDARD AUTOMATIQUE DE MYRTILLE I. JE VOUS ECOUTE.

---> COR :

J
EU I
B
OU U
B
AI EU EI
L N
EU EI AI
P
A
L
AI EU EI
A UN
M B
FU
S
U EI
B
EI U AI
P
A AN

(Je voudrais parler à MONSIEUR DUPONT)

JE VOUS PASSE M. DUPONT.

VOUS AVEZ EN LIGNE M. DUPONT.

FIN DE COMMUNICATION

DEBUT DE COMMUNICATION

STANDARD AUTOMATIQUE DE MYRTILLE I. JE VOUS ECOUTE.

---> COR :

J
EU EI U
B
OU AU
U AI EI
L N
EI EU I
N M
A UN U
B
UN ON
B M N
AU OU
AI I
CH
AI EI U

(je voudrais Madame Boucher)

DESOLE JE N'AI PAS COMPRIS. VEUILLEZ REPETER, JE VOUS ECOUTE.

---> COR :

J
EU EI U
B
OU O AU
V Z B
AI I EU
L N
EU AI
N M L
UN U
B
A UN U
M N
B
OU O
CH
AI I EI

(je voudrais Madame Boucher)

JE VOUS PASSE MADAME BOUCHER

VOUS AVEZ EN LIGNE MADAME BOUCHER

FIN DE COMMUNICATION

DEBUT DE COMMUNICATION

STANDARD AUTOMATIQUE DE MYRTILLE I. JE VOUS ECOUTE.

---> COR :

AN	A	
N	L	
AI	O	AU
B		
O	AN	A
J		
OU	O	
J		
EU	EI	U
B	M	
O	OU	
B		
Z	R	L
AI	EU	EI
L	N	
EU	EI	U
P		
AN	A	O
S		
P	S	
B		
EI	EU	
S		
AN	A	
AN	A	OU
B	N	
F	AI	EI

(Allo bonjour, je voudrais le poste 202)

VOUS AVEZ EN LIGNE LE POSTE 202

FIN DE COMMUNICATION

FIN DE SESSION

IV. RESULTATS EXPERIMENTAUX

Le système a été expérimenté sur environ 1 000 phonèmes formant cinquante phrases dans le cadre de l'application actuelle du système MYRTILLE I, simulant un standard téléphonique automatique. Ces tests ont eu lieu en salle d'ordinateur où, en plus du bruit des ventilateurs intégrés au calculateur, nous trouvons beaucoup de bruit ambiant (perforatrice, rotation des disques du calculateur, imprimante).

La reconnaissance phonétique est réalisée à partir des formes de références stockées au cours d'une phase d'apprentissage. Dans ce système, nous avons 35 références pour les voyelles (figure 86), 14 pour les consonnes voisées (figure 87) et 7 pour les consonnes sourdes (figure 88).

On peut constater, sur les exemples du paragraphe précédent, un certain nombre d'erreurs dans les chaînes reconnues. Cependant, en regroupant certains phonèmes, on peut corriger ces erreurs et par suite obtenir un taux de reconnaissance très élevé :

classe de phonèmes		taux de reconnaissance
{ɛ , e}	⟶	96 %
{ə , Ø}	⟶	96 %
{$\tilde{\varepsilon}$, $\tilde{\text{œ}}$}	⟶	98 %
{$\tilde{\text{a}}$, $\tilde{\text{o}}$}	⟶	92 %
{u , O , ɔ}	⟶	98 %
{p , t , k}	⟶	96 %
{b , d , g}	⟶	96 %

PHONEMES MOYENS
==========

	IDF	TYP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	GA	FO	BD	DS	CE
1	A	5	99	100	110	103	91	72	60	47	54	54	40	42	42	35	10	0	12	10	4	-98	226
2	O	5	114	107	90	67	55	54	47	44	38	31	27	32	19	6	0	0	89	-23	3	-70	272
3	I	5	106	80	59	48	41	44	42	62	61	54	53	52	53	37	0	0	16	-46	66	-115	299
4	EU	5	110	105	76	62	55	64	55	55	48	32	23	48	25	29	0	0	56	-34	1	-80	314
5	EI	5	101	93	64	50	42	43	41	58	52	48	33	32	42	38	0	0	21	-37	39	-101	290
6	OU	5	112	87	80	57	49	49	43	39	30	17	18	23	6	0	0	0	106	-32	-13	-47	246
7	AN	5	101	88	92	88	61	49	41	33	20	17	39	46	38	3	0	0	60	-8	-4	-37	156
8	UN	5	93	79	89	70	57	69	66	51	47	28	27	45	45	37	0	0	11	-4	0	-76	268
9	ON	5	106	113	113	93	69	62	55	48	44	39	42	50	36	33	35	0	2	6	12	-48	244
10	U	5	103	77	62	50	45	46	52	54	45	20	6	31	4	4	0	0	95	-40	-17	-66	279
11	U	5	108	88	68	59	53	59	73	72	69	53	53	48	22	29	0	0	57	-39	54	-123	352
12	A	5	93	102	101	74	69	68	52	41	39	45	29	32	26	29	0	0	38	8	5	-84	249
13	EI	5	108	98	73	56	50	52	61	78	67	72	53	53	57	44	1	0	6	-35	75	-139	352
14	EU	5	110	99	72	59	54	66	51	53	42	32	38	36	15	18	0	0	77	-38	16	-75	303
15	A	5	96	103	94	72	61	65	48	42	43	23	30	36	28	27	1	0	40	-2	-7	-65	254
16	EU	5	108	99	76	60	58	69	55	62	44	32	46	44	5	4	0	0	99	-32	20	-77	318
17	A	5	88	96	99	80	74	56	43	34	38	27	12	19	13	5	0	0	70	11	-35	-65	196
18	A	5	91	89	104	94	91	82	60	49	50	37	24	56	54	47	0	0	-10	13	-29	-88	234
19	I	5	98	69	53	43	37	40	43	58	47	51	42	46	19	26	0	0	53	-45	56	-99	270
20	AI	5	107	91	70	56	46	53	70	67	63	60	43	53	24	21	0	0	62	-37	57	-123	359
21	AU	5	118	107	90	68	57	56	48	44	40	33	25	27	27	13	0	0	78	-28	1	-74	278
22	U	5	103	80	66	56	50	51	72	75	57	46	41	43	22	21	0	0	60	-37	37	-103	328
23	EU	5	107	92	69	56	52	70	54	54	48	32	44	42	16	3	0	0	88	-38	24	-81	315
24	EU	5	103	87	67	56	51	56	65	66	68	48	47	38	3	24	0	0	76	-36	44	-116	334
25	O	5	120	107	96	78	61	60	52	55	48	47	46	40	34	25	0	0	61	-23	32	-95	280
26	I	5	95	76	59	48	44	45	41	61	70	54	66	66	56	35	0	0	4	-36	77	-124	294
27	OU	5	112	88	82	67	49	51	44	41	30	21	16	24	6	0	0	0	106	-30	-12	-51	234
28	AN	5	106	102	109	100	78	62	50	46	40	37	55	63	49	27	0	0	30	3	14	-78	207
29	UN	5	105	90	93	79	66	75	67	48	47	44	41	55	49	43	19	0	-6	-11	19	-71	276
30	EI	5	106	88	63	51	45	48	64	59	57	51	44	42	5	10	0	0	91	-42	50	-108	321
31	AI	5	104	83	61	49	42	44	45	47	41	49	30	37	5	0	0	0	99	-43	37	-91	267
32	EI	5	93	102	105	77	61	66	83	91	68	80	90	79	77	71	67	0	-122	11	110	-81	348
33	AU	5	107	93	85	58	46	46	40	33	29	12	1	18	0	0	0	0	107	-22	-32	-41	235
34	A	5	96	95	111	99	94	75	58	47	48	70	47	60	60	58	55	0	-77	15	22	-64	223
35	U	5	107	84	71	60	57	58	78	89	67	62	48	46	35	31	0	0	41	-36	54	-129	365

Figure 86 : Formes de références des voyelles

PHONEMES MOYENS
==========

	IDF	TYP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	GA	FO	BD	DS	CE
1	B	5	68	43	30	7	10	0	1	0	0	0	0	0	0	0	0	0	68	-37	-10	0	97
2	B	5	84	52	38	23	24	19	13	9	6	2	1	4	0	2	0	0	82	-45	-20	-8	137
3	J	5	78	64	47	35	29	32	36	54	40	50	50	67	62	56	54	0	-94	-31	71	-36	236
4	Z	5	89	75	55	45	41	55	35	17	13	27	3	11	16	34	13	0	26	-34	-10	-27	195
5	L	5	99	69	51	43	37	46	31	26	42	37	6	20	27	16	0	0	56	-47	6	-60	227
6	R	5	80	85	81	68	63	46	28	15	35	19	16	33	18	16	0	0	46	1	-27	-55	153
7	R	5	87	87	73	80	53	41	31	29	43	24	33	27	11	11	0	0	65	-14	4	-67	159
8	M	5	84	62	45	34	31	31	31	30	29	2	0	1	0	0	0	0	84	-35	-26	-31	186
9	N	5	91	62	45	37	32	32	20	19	19	20	0	0	0	5	0	0	86	-46	-12	-39	170
10	L	5	102	74	55	48	44	56	39	27	44	35	39	28	15	14	0	0	73	-46	30	-60	249
11	L	5	90	65	51	43	45	46	31	23	44	40	19	9	0	0	0	0	90	-39	13	-85	216
12	N	5	93	70	51	42	36	43	35	19	17	38	0	7	0	0	0	0	93	-42	2	-55	195
13	N	5	96	72	54	47	43	43	42	33	38	41	0	2	0	0	0	0	96	-41	-2	-79	229
14	M	5	93	67	52	45	45	42	32	21	22	0	0	11	4	11	0	0	78	-40	-45	-22	188

Figure 87 : Formes des références des consonnes voisées

PHONEMES MOYENS

	IDF	TYP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	GA	FO	BD	DS	CE
1	CH	2	30	8	0	0	1	0	6	10	39	45	52	66	56	53	57	14	-150	-29	97	-27	95
2	CH	2	35	19	6	7	8	5	15	18	47	51	57	64	60	53	59	39	-176	-24	90	-35	119
3	P	2	23	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	23	-20	0	0	23
4	P	2	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	-22	0	0	22
5	S	2	37	13	4	1	1	0	2	1	0	0	0	0	12	19	19	51	-64	-33	-1	19	51
6	S	2	37	12	7	2	4	0	0	0	0	0	0	0	28	31	27	20	-69	-30	-4	27	46
7	F	2	37	27	27	22	17	15	13	9	8	4	0	0	0	0	0	0	37	-10	-13	-12	67

Figure 88 : Formes des références des consonnes sourdes

Les tableaux I et II fournissent respectivement les matrices de confusion (en pourcentage) pour les voyelles et les consonnes. Ces matrices correspondent à plusieurs prononciations des phonèmes dans des contextes différents par un seul locuteur masculin.

Dans le tableau II, "b" représente la classe des plosives voisées /b , d , g/ et "p" la classe des plosives sourdes /p , t , k/. On ne fait pas pour l'instant de distinction à l'intérieur de ces classes.

Phonème → Reconnu
Prononcé ↓

	a	ɛ	e	i	ə	ø	u	y	o	ɔ	ɑ̃	ɔ̃	ɛ̃	œ̃
a	94												3	3
ɛ		50	46	4										
e		46	50	4										
i		2	2	96										
ə		2	2		60	36								
ø		2	2		36	60								
u							80		10	10				
y		5	5	10				80						
o							10		50	40				
ɔ							10		40	50				
ɑ̃	8										56	36		
ɔ̃	8										36	56		
ɛ̃													60	40
œ̃													40	60

Tableau I
Matrice de confusion de voyelles

Phonème → Reconnu
Prononcé ↓

	r	l	b	m	n	v	z	j	f	s	ʃ	p
r	50	30		10	10							
l	30	50		10	10							
b			96				4					
m		5	5	65	25							
n		5	5	25	65							
v			60			40						
z							90	10				
j							5	95				
f									60	20		20
s										90	10	
ʃ										5	95	
p										4		96

Tableau II
Matrice de confusion de consonnes

CONCLUSION

La conception du système de reconnaissance phonétique de la parole continue présentée dans cette thèse, s'appuie sur un modèle mixte original mettant en jeu deux processus complémentaires fonctionnant en parallèle :

- une identification centiseconde,
- une segmentation des noyaux vocaliques et idenfication des segments.

La prise en compte simultanée des résultats de ces deux processus permet de corriger certaines erreurs de reconnaissance, dues à une mauvaise segmentation.

La reconnaissance centiseconde identifie des tranches du signal vocal de 10 ms, puis ensuite regroupe les segments identiques.

Notre système est fondé sur le phonème en tant qu'unité de reconnaissance, sans pour autant négliger la notion de noyau syllabique, tant dans la phase de segmentation que d'identification.

L'analyse acoustique est effectuée par un vocoder à canaux, avec détection de fondamental. L'identification des voyelles, des consonnes voisées et des consonnes sourdes s'effectue séparément sur la base d'indices déterminés à partir des données fournies par vocoder.

La fonction de segmentation du signal vocal, en phonèmes, est choisie en fonction de la nature de la zone à segmenter : la détection des voyelles est réalisée par la fonction : "segmentation par identification",

la segmentation des zones voisées par la variation spectrale, et celle des zones sourdes par la variance de l'énergie de chaque canal autour de la valeur moyenne.

La prononciation en contexte d'un phonème modifie ses caractéristiques. Ces variations sont aussi liées au mode d'élocution adopté (rapide, lent, ...). Pour tenir compte de ces altérations, nous avons développé une méthode qui détecte les noyaux vocaliques, à partir de la courbe d'énergie, puis segmente et identifie les voyelles.

Notre système représente l'actuel niveau phonétique, chargé de fournir un treillis de phonèmes, des systèmes de compréhension de langages MYRTILLE I et MYRTILLE II.

Nous obtenons, actuellement, un taux de reconnaissance de 90 % pour les voyelles ainsi que pour les plosives et les fricatives /j/ /z/ /s/ /ʃ/, celui des autres phonèmes avoisine 75 %, pour un locuteur après apprentissage. Le taux de reconnaissance moyen est donc de l'ordre de 80 % pour un locuteur, en considérant au maximum trois phonèmes candidats pour chaque segment identifié. On peut l'améliorer en utilisant, en plus des paramètres issus du vocoder, les informations suivantes :

- les paramètres calculés à partir du signal temporel (nombre de passages par zéro, longueur curviligne, nombre de maxima ...)
- les paramètres calculés à partir d'un analyseur spectral (F.F.T., L.P.C., Cepstre ...) dans des portions du signal bien précises :

. après une zone de silence, pour étudier le "burst" et par suite séparer les différentes plosives.

. dans les segments correspondants aux liquides, pour étudier la position et le suivi des formants.

- les règles issues du système expert en décodage phonétique (en étude dans notre laboratoire depuis 1982). Voici quelques exemples de règles applicables dans un système de reconnaissance des phonèmes :

(1) Si concentration d'énergie aux environs de 250 Hz alors
sonore

(2) Si (1) et énergie globale forte alors
voyelles, sonantes

(3) Si (1) et énergie faible alors
plosives, sonantes

(4) Si (1) et pas d'énergie visible dans le haut du spectre alors
b, d, g, v

(5) Si (1) et F1 bas et décrochement avec les formants des voyelles contigues alors
m, r, ʁ, l

(6) Si la durée d'un silence est comprise entre 50 ms et 150 % de la durée vocalique moyenne alors une occlusive.

Parallèlement à ces améliorations d'autres aspects sont à l'étude tels que l'automatisation de l'apprentissage, la détermination des seuils, l'adaptation aux nouveaux locuteurs, etc ...

BIBLIOGRAPHIE

[ALIN-72] ALINAT P.
"Reconnaissance des phonèmes en temps réel en vue de réaliser des liaisons à faible débit. Application à la sténotypie automatique"
Rapport DRME n° 198/71, octobre 1972.

[BAUD-72] BAUDRY M., DUPEYRAT B. FRANK C.
"Reconnaissance automatique de la parole. Etude de la segmentation"
Actes des 3èmes journées du GALF. Lannion, mai-juin 1972.

[BAUD-74] BAUDRY M., DUPEYRAT B.
"Temporal Analysis of the Voice Signal Automatic Recognition of Spoken Word"
2nd Inter Joint Conf. On Pattern Recognition Copenhague, August 1974.

[BAUD-78] BAUDRY M.
"Etude du signal vocal dans sa représentation amplitude - temps. Algorithmes de segmentation et de reconnaissance de la parole".
Thèse d'Etat, Paris VI, juin 1978.

[BELL-78] BELLISSANT C.
"Contribution à l'analyse et la reconnaissance automatique de la parole"
Thèse d'Etat, Université de Grenoble, 1978.

[BELL-78-b] BELLISSANT C.
"Un algorithme de segmentation de la parole"
Congrès AFCET-IRIA, Février 1978, Châtenay-Malabry.

[CAEL-79] CAELEN J.
"Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique".
Thèse d'Etat, Toulouse, 1979.

[CAEL-81] CAELEN J., CAELEN G.
"Indices et propriétés dans le projet ARIAL II".
Actes du séminaire "Processus d'encodage et de décodage phonétiques", Toulouse, Septembre 1981.

[CARB-83] CARBONELLE N., HATON J.P., LONCHAMP F., PIERREL J.M.
"Elaboration d'un système expert pour le décodage phonétique automatique de la parole".
11th int. Cong. of Acoustics, Symposium de Toulouse, 15-16, juillet 1983.

[CART-79] CARTON FERNAND
"Introduction à la phonétique du français"
Collection << ETUDDES >>

[CAST-78] CASTELAZ P.F., NIEDERJOHN R.J.
"A comparison of linear prediction, FFT, zero-crossing analysis techniques for vowel recognition"
Conference record of ICASSP, 541-545, 1978.

[COOL-65] COOLEY J.W., TUKEY J.W.
"An algorithm for the machine calculation of Complex Fourier series"
Math. Comp. 19, 1965.

[DELL-73] DELL FRANÇOIS
"Les règles et les sons, introduction à la phonologie générative".
Collection SAVOIR, HERMANN.

[DEMO-76-a] DE MORI R.
"Experiments in the automatic segmentation of continous speech".
Acoustica 34, 158-166, 1976.

[DEMO-76-b] DE MORI R., LAFACE P., PICCOLO E.
"Automatic extraction and description of syllabic features in continous speech".
IEEE. trans. ASSP, n° 24, 1976.

[DIMA-81] DI MARTINO J.
"Normalisation temporelle par programmation dynamique".
Rapport de D.E.A., Université de Nancy 1, 1981.

[DOUR-74] DOURS D., FACCA R.
"Méthode de segmentation et d'analyse par traitement direct du signal vocal. Application à la classification et à la reconnaissance des voyelles et des consonnes".
Thèse 3ème cycle, Toulouse, Novembre 1974.

[DUPE-75] DUPEYRAT B.
"Reconnaissance de la parole. Méthode des passages par zéro du signal. Reconnaissance automatique des voyelles isolées".
Thèse de 3ème cycle, Université Paris VI, 1975.

[FLAN-72] FLANAGAN J.L.
"Speech Analysis, Synthesis and Perception".
Second Edition, Springer-Verlag, 1972.

[GOLD-75] GOLDBERG H.G.
"Segmentation and labeling of speech : a comparative performance evaluation".
Ph. D. Thesis, Carnegie Mellon University, 1975.

[GREN-78] GRENIER Y.
"Speaker identification from linear prediction". Proceeding of the 4th IJCPR, 1019-1021, 1978.

[HATO-74-b] HATON J.P.
"Contribution à l'analyse, la paramétrisation et la reconnaissance de la parole".
Thèse d'Etat, Université de Nancy 1, 1974.

[HATO-78] HATON J.P., PERENNOU G.
"Reconnaissance et compréhension de la parole".
Cours de l'Ecole d'Eté Informatique de l'A.F.C.E.T., Namur, Belgique, juillet 1978.

[HATO-79] HATON J.P., SANCHEZ C.
"An Experimental System for Acoustic-Phonetic Decoding of Continuous Speech"
ICASSP, April 1979, Washington, D.C.

[HATO-81-a] HATON J.P., SANCHEZ C.
"Méthodes synchrone et asynchrone en reconnaissance phonétique de la parole".
Actes du séminaire "Processus d'encodage et de décodage phonétiques", Toulouse, 24-25 septembre 1981, pp 144-155.

[HATO-81-b] HATON J.P.,MARI J.F., PEROT J.Y., PIERREL J.M., SABBAGH S., SANCHEZ C.
"MYRTILLE II : A system for pseudo-naturel language understanding.
12èmes J.E.P. Montréal, 1981.

[JACO-63] JACOBSON R.
"Essai de linguistique générale"
Edition de minuit, 1963.

[JELI-81] JELINEK F.
"Self-Organized Continuous Speech Recognizer"
NATO ASI "Automatic Speech Analysis and Recognition", Bonas, july 1981.

[KLAT-80] KLATT D.
"SCRIBER and LAFS two new approaches to speech analysis"
Trends in speech recognition, W. LEA, editor, Prentice Hall, 1980.

[LAZR-81] LAZREK M.
"La segmentation de la parole en phonèmes"
Rapport de D.E.A., Université de Nancy 1, 1981.

[LINE-73] LIENARD J.S.
"Normalisation fréquentielle de la parole"
Actes des 4èmes journées du GALF, Bruxelles, mai 1973.

[MALM-72] MALMBERG B.
"Phonétique française"
Hermods Malmo Suède, End Edition, 1972.

[MART-75] MARTIN P.
"Intonation et reconnaissance automatique de la parole"
8ème J.E.P., GALF, Toulouse, mai 1975.

[MART-76] MARTIN T.B.
"Practical Applications of Voice Input to Machines"
Proceeding of the I.E.E.E, Vol. 64, n° 4, April 1976.

[MASS-75] MASSARO D.W.
"Understanding language : An information processing analysis of speech perception, readin and psycholinguistics".
A. Press, New-York, London, 1975.

[MELO-82] MELONI H.
"Contribution à la recherche sur la reconnaissance automatique de la parole continue"
Thèse de doctorat d'état ès sciences, Université d'Aix-Marseille II, février 1982.

[MERC-82] MERCIER G.
"Acoustic Processing and Phonetic Analysis in Continuous Speech Recognition".
"Automatic Speech Analysis an Recognition"
J.P. HATON Editor, REIDEL, 1982.

[MICL-81] MICLET L. et Al.
"Reconnaissance phonétique par prédiction linéaire, l'expérience de l'ENST".
12ème j.E.P. GALF, Montreal, mai 1981.

[PERE-79-a] PERRENOU G.
"Reconnaissance de mots isolés dans le cas d'un grand vocabulaire"
Congrès AFCET, RF-IA, Toulouse 1979.

[PETE-51] PETERSON E.
"Frequency detection and speech formant".
JASA n° 23, 1951.

[PIER-75] PIERREL J.M.
"Contribution à la reconnaissance automatique du discours continu".
Thèse de Doctorat de Spécialité, Université de Nancy 1, novembre 1975.

[PIER-81] PIERREL J.M.
"Etude et mise en oeuvre de contraintes linguistiques en compréhension automatique du discours continu"
Thèse d'Etat, Université de Nancy 1, mars 1981.

[PINS-63] PINSON E.N.
"Pitch - Synchronous Time - Domain Estimation of Formant Frequences and Bandwiths".
JASA, Vol. 35, 1963.

[REDD-66] REDDY D.R., VICENS P.
"A procedure for segmentation of connected speech".
J. Audio Engr. Soc. 16, 4, 1968.

[REGE-82] REGEL PETER
"A module for acoustic-phonetic transcription of fluently spoken german speech".
IEEE, Transaction on acoustics, speech, and signal processing, Vol. ASSP-30, n° 3, june 1982.

[RODE-77] RODET X.
"Analyse du signal vocal dans sa représentation amplitude-temps. Synthèse de la parole par règles".
Thèse d'Etat, Paris VI, juin 1977.

[ROSS-75] ROSSI M.
"Les contraintes phonologiques dans un système de reconnaissance de la parole".
6ème J.E.P. GALF, Toulouse, 1975.

[ROSS-77] ROSSI M., DI CRISTO A.
"Proposition pour un modèle d'analyse de l'intonation".
8ème J.E.P., Aix en Provence, 1977.

[SAKA-79] SAKAI T.
"Automatic mapping of acoustic features into phonemic labels"
Nato Advanced Study Institute, Bonas, 1979.

[SAMB-75] SAMBUR M.R.
"Selection of acoustic features for speaker identification"
IEEE Trans, Vol. ASSP 23, n° 2, 169-176, 1975.

[SANC-78] SANCHEZ C., MESSENET G., HATON J.P.
"Etude et utilisation des indices acoustiques et des traits pour la segmentation et la reconnaissance phonétique de la parole".
9èmes journées d'études sur la parole du Groupe de la communication parlée, Lannion, juin 1978.

[SCHA-70] SCHAFER R.W., RABINER L.R.
"System for automatic analysis of voiced speech".
JASA, vol. 47, n° 2, 1970.

[SCHA-75] SCHAFER R.W., RABINER L.R.
"Parametric representation of speech".
D.R. REDDY Ed., Speech Recognition.
Invited Paper of the IEEE Symposium, New-York : Academic Press, 1975.

[SHIR-83] SHIRAI K., KOBAYASHI T.
"Considerations on articulatory dynamics for continuous speech recognition".
Proc. ICASSP 1983.

[SHOU-80] SHOUP J.
"Phonological aspects of speech recognition"
Trends in speech recognition, W. LEA, editor, Prentice Hall, 1980.

[STRA] STRAKA G.
"Album phonétique"
Les presses de l'université, Laval.

[VAIS-80] VAISSIERE J.
"Speech recognition programs as models of speech perception"
In the cognitive representation of speech T. MYERS, J. LAVER and C. ANDERSON, editors, Noth Holland Publishing company, in press, 1980.

[WOLF-72] WOLF J.J.
"Efficient acoustic parameters for speaker recognition"
JASA, vol. 51, n° 6, 2044-2056, 1972.

NOM DE L'ETUDIANT : LAZREK Mohamed

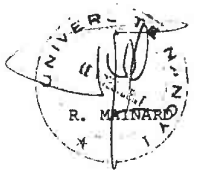
NATURE DE LA THESE : Doctorat 3ème cycle en Informatique

VU, APPROUVE

ET PERMIS D'IMPRIMER

NANCY, le - 8 JUIN 1983 905

LE PRESIDENT DE L'UNIVERSITE DE NANCY I,



RESUME

Le décodage phonétique, c'est-à-dire le passage de l'onde acoustique de la parole à une suite discrète de symboles phonétiques, constitue un problème majeur en reconnaissance automatique de la parole continue. Les données fournies par ce décodage étant ensuite prises en compte par les niveaux linguistiques, il est indispensable d'améliorer la qualité de la chaîne phonémique obtenue.

Nous décrivons dans cet exposé l'étage de décodage acoustico-phonétique utilisé à Nancy dans les systèmes MYRTILLE. Cet étage présente la particularité de s'appuyer sur un modèle mixte à la fois synchrone ("centiseconde") et asynchrone (avec segmentation).

L'analyse acoustique est effectuée par un Vocoder à canaux avec détection de fondamental. L'identification des voyelles, des consonnes voisées et des consonnes sourdes s'effectue séparément sur la base d'indices déterminés à partir des données fournies par Vocoder.

MOTS - CLES

Décodage phonétique.

Onde acoustique de parole.

Reconnaissance synchrone.

Reconnaissanca asynchrone.