Université de Nancy I

U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

ADAPTATION AU LOCUTEUR PAR APPRENTISSAGE AUTOMATIQUE APPLICATION À UN SYSTÈME DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

THESE

présentée et soutenue publiquement le **5 juin 1984**

À L'UNIVERSITÉ DE NANCY I

pour l'obtention du diplôme de

DOCTORAT DE 3ème CYCLE EN INFORMATIQUE

par

Christine PISTER

devant la Commission d'Examen :

Président : ..... *J.P. HATON*

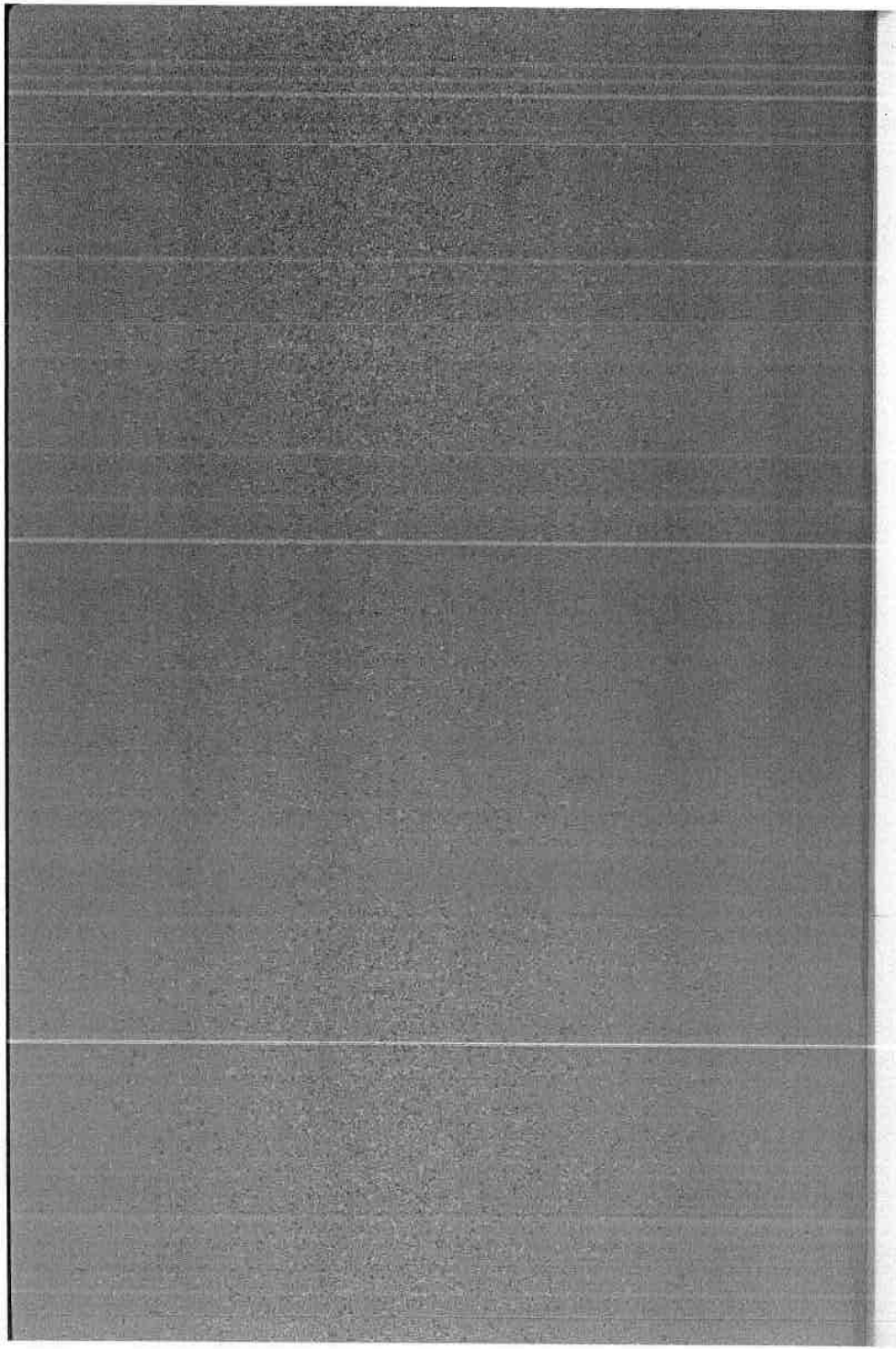
Examinateurs : ..... *M.C. HATON*

F. LONCHAMP

P. MARCHAND

G. PERENNOU

J.M. PIERREL



Université de Nancy I

U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

ADAPTATION AU LOCUTEUR PAR APPRENTISSAGE AUTOMATIQUE APPLICATION À UN SYSTÈME DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

THESE

présentée et soutenue publiquement le **5 juin 1984**

À L'UNIVERSITÉ DE NANCY I

pour l'obtention du diplôme de

DOCTORAT DE 3ème CYCLE EN INFORMATIQUE

par

Christine PISTER

devant la Commission d'Examen :

Président : . . . . . J.P. HATON

Examinateurs : . . . . . M.C. HATON

F. LONCHAMP

P. MARCHAND

G. PERENNOU

J.M. PIERREL

S O M M A I R E

Ce travail a été réalisé au Centre de Recherche en Informatique de Nancy sous la direction de Monsieur le Professeur J.P. HATON. Je le prie de trouver ici l'expression de ma très sincère reconnaissance pour son soutien de toujours et l'intérêt constant avec lequel il a suivi mon travail.

G. PERENNOU, Professeur du CERFIA à l'Université Paul Sabatier de Toulouse m'a fait l'honneur de s'intéresser à ce travail. Je l'en remercie vivement ainsi que d'avoir accepté notre invitation à ce jury.

Je tiens également à remercier Madame M.C. HATON, Maître-assistante à l'Université de Nancy I, Monsieur F. LONCHAMP, Maître-assistant à l'Institut de Phonétique de Nancy II, Monsieur P. MARCHAND, Professeur à l'Université de Nancy II et Monsieur J.M. PIERREL, Professeur à l'Université de Nancy I, d'avoir bien voulu accepter de juger mon travail et de faire partie de mon jury de thèse.

Je veux tout particulièrement témoigner ma profonde gratitude envers Monsieur F. LONCHAMP pour les appuis et les encouragements qu'il m'a prodigués tout au long de cette étude.

Que Monsieur J.L. CASTAGNE trouve également ici l'expression de mes très vifs remerciements pour l'aide et les conseils qu'il m'a prodigués au cours des études statistiques menées durant ce travail.

Qu'il me soit permis d'associer dans une même pensée amicale tous ceux qui au C.R.I.N. ont contribué à la réalisation de ce travail lors des tests d'apprentissage et de reconnaissance. Qu'ils trouvent tous ici l'expression de ma très vive sympathie.

Enfin, que Mademoiselle M. TESOLIN qui a assumé avec diligence la dactylographie de mon manuscrit en soit très sincèrement remerciée.

"Tout progrès technique a une répercussion sur le plan social. Demain un patron, au lieu de dicter ses lettres à sa secrétaire, les dictera à l'ordinateur. Dans une société mal faite cela signifiera accroissement du chômage, dans une société bien faite, accroissement des loisirs et des heures de culture."

Alfred KASTLER,
Prix Nobel de Physique.

INTRODUCTION

Le but principal de la reconnaissance automatique de la parole est de faciliter le dialogue entre l'homme et la machine par la communication orale. Un thème important de la recherche actuelle dans ce domaine est la généralisation de l'utilisation des systèmes de reconnaissance à l'ensemble d'une population importante d'individus, par exemple pour des applications téléphoniques.

La difficulté principale pour les systèmes multilocuteurs est de gérer la grande variabilité intra et inter locuteur de la parole qui se traduit par des variations acoustiques, phonétiques, prosodiques et phonologiques du message à reconnaître.

La recherche d'invariants de la parole n'a pas fourni pour l'instant des paramètres de reconnaissance suffisamment nombreux et pertinents pour obtenir des scores de reconnaissance satisfaisants dans les systèmes fonctionnant pour un nombre élevé d'individus.

C'est pourquoi, dans le cadre d'un système de reconnaissance de mots isolés par approche analytique pour un vocabulaire de grande taille (400 puis un millier de mots) nous avons envisagé une adaptation automatique du système à un nouveau locuteur.

Notre étude comporte quatre chapitres :

Dans le premier chapitre sont présentées quelques notions sur la production de la parole ainsi que les caractérisques phonétiques utiles à la reconnaissance des sons émis par l'homme. Ceci permet de développer l'étude de la variabilité de la parole due à des différences anatomiques, mais aussi à l'environnement linguistique d'un individu.

Le deuxième chapitre présente les différentes approches de la reconnaissance automatique de la parole assorties de leurs comportements face à un changement de locuteur. Le point sur les systèmes multilocuteurs actuels est établi en les divisant en deux grandes classes : les systèmes indépendants du locuteur et les systèmes adaptatifs.

Le troisième chapitre présente l'adaptation au locuteur par apprentissage automatique qui consiste à extraire sans aucune intervention humaine les phonèmes de quelques phrases prédéfinies prononcées par le nouveau locuteur. Un algorithme de cadrage de segments phonétiques décodés à l'aide de traits acoustico-phonétiques peu dépendants du locuteur est à la base de cet apprentissage.

Une analyse des voyelles de quinze locuteurs obtenues par apprentissage automatique fait l'objet du quatrième chapitre. Les études relèvent de la statistique descriptive et de l'interprétation statistique et ont été entreprises en vue de la réalisation de procédure de normalisation et de la génération automatique des formes de références des voyelles d'un locuteur.

CHAPITRE I

INTRODUCTION A LA PAROLE

INTRODUCTION :

Notre approche est d'examiner le mécanisme de production de la parole ainsi que les caractéristiques phonétiques utiles à la reconnaissance des sons émis par l'homme pour indiquer quels sont les aspects du processus en relation avec les caractéristiques anatomiques d'un individu.

Ceci nous permet de développer l'étude de la variabilité de la parole due à des changements anatomiques et physiologiques fonction du temps pour un individu donné, et l'étude de la variabilité de la parole en fonction des différences anatomiques d'un locuteur à l'autre.

D'autre part, la façon dont un locuteur produit certains sons de parole ne dépend pas que de ses caractéristiques anatomiques mais aussi de l'environnement linguistique dans lequel il évolue ; c'est pourquoi nous évoquerons quelques facteurs de variabilité dûs à des habitudes de prononciation.

1.- DESCRIPTION ET FONCTIONNEMENT DE L'APPAREIL VOCAL.

Nous nous limitons à une description générale de l'appareil vocal, juste nécessaire à la poursuite de l'étude des sons émis par l'homme. Pour une description plus détaillée et plus approfondie on pourra consulter parmi de nombreux ouvrages (MARCHAL-80, LIENARD-77).

1.1.- Description

L'appareil vocal comprend les poumons, le larynx, le pharynx, la bouche et le nez. Voir figure I-1.

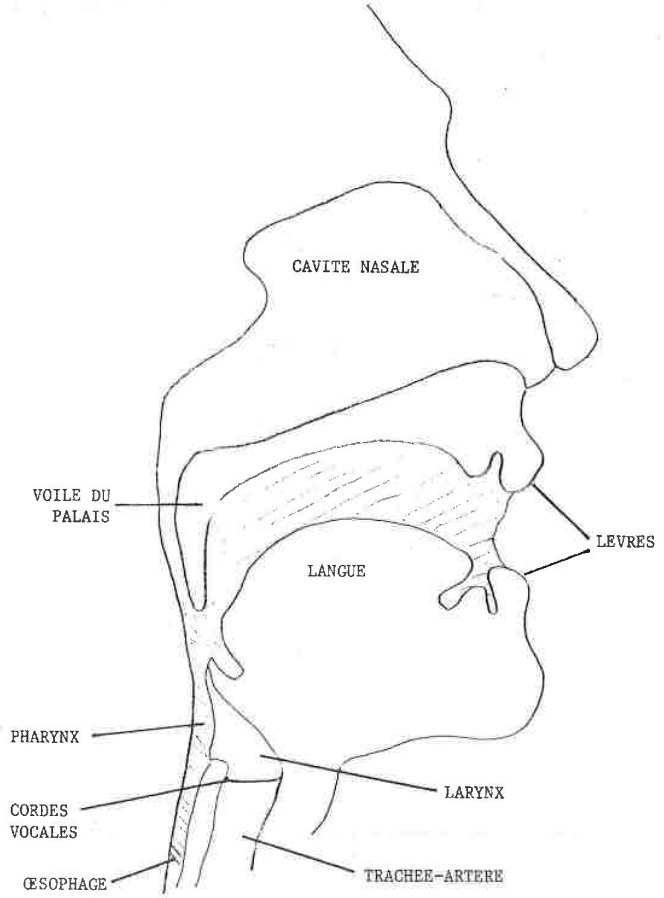


Figure I-1 : L'APPAREIL VOCAL (d'après SUNDBERG-80)

- le rôle principal des poumons, dans la phonation, est de produire un excès de pression et de créer ainsi un flux d'air. L'air passe par la glotte située à la base du larynx entre les cordes vocales qui sont des replis élastiques de la muqueuse tapissant le larynx.

- Le conduit vocal se compose du pharynx, des cavités buccales et des cavités nasales.

Le larynx communique avec le pharynx qui va de la bouche à l'œsophage. La voûte supérieure du pharynx, le voile du palais, règle les échanges avec la cavité nasale.

Quand le voile est relevé, c'est-à-dire pendant l'émission de toutes les voyelles à l'exception des voyelles nasalisées, le passage vers la cavité nasale est clos et l'air s'échappe par la bouche.
La forme du conduit vocal dépend de la position des organes articulatoires qui sont les lèvres, la mâchoire, la langue et le larynx. La partie orale du conduit vocal mesure chez l'homme de 17 à 20 cm,
la femme de 15 à 17 cm,
l'enfant de 10 à 15 cm.

1.2.- Fonctionnement.

* La source des sons de la parole se forme à partir de la rencontre du flux d'air pulmonaire et d'un obstacle. Cette source sera dépendante du débit de l'air et du lieu de l'obstacle.

Les cordes vocales constituent le premier obstacle que va rencontrer le flux d'air. Si la glotte est fermée la forte pression qui règne au-dessous du niveau de la glotte force les cordes vocales à se séparer ; l'air passant entre les cordes vocales engendre une force de Bernoulli qui, grâce aux propriétés mécaniques des cordes, les ferme presque instantanément. La pression augmente alors à nouveau, séparant les cordes vocales et ainsi de suite. C'est ainsi que se forme l'onde de débit d'air, ou onde glottique dont la fréquence est produite par les vibrations des cordes vocales. Ceci est à la base de la formation des sons voisés (sonores), voyelles ou consonnes sonores.

Divers types de voix, registre de poitrine, fausset, etc. peuvent être obtenus en fonction de la pression sous-glottique et des propriétés des cordes vocales. Pour la voix chuchotée , par exemple, les cordes vocales sont peu fortement comprimées l'une contre l'autre et n'entrent pas en vibration.

Un autre obstacle que peut rencontrer le flux d'air est un rétrécissement du conduit vocal. L'air s'échappe alors sous la forme d'un jet tourbillonaire ce qui fournit une source de bruit, caractéristique des consonnes fricatives.

En dernier lieu, l'obstacle au passage de l'air peut être total, produisant une augmentation de la pression d'air en amont du lieu d'occlusion dans le conduit vocal suivie d'un relâchement brusque. Ce phénomène est la source d'un bruit impulsionnel à l'origine des consonnes occlusives.

La production de certains sons de parole met en jeu plusieurs phénomènes, ainsi les consonnes fricatives et occlusives voisées sont le produit de deux sources de son :

la source périodique due aux cordes vocales et la source de bruit continue due à la constitution du conduit vocal.

* La source de son est ensuite modifiée par les cavités de résonance du conduit vocal qu'elle traverse.

La propagation du son à travers le résonateur acoustique, qu'est le conduit vocal, dépend du volume d'air et de la forme du résonateur modifiable par le jeu complexe des organes articulatoires.

La forme des différentes cavités détermine une fonction de transfert qui s'applique à la source du son pour donner le signal de la parole.

La fonction de transfert du conduit vocal détermine les fréquences qui seront renforcées par le phénomène de résonance.

On peut noter que de nombreux modèles mathématiques ont été proposés afin de formuler certaines règles articulatoires utilisées par les phonéticiens pour expliquer la formation des voyelles (MRAYATI-76).

2.- ETUDE DES SONS EMIS.

Introduction :

Une des tâches du phonéticien est, à partir d'un continuum de parole, d'isoler les sons en segments phonétiques et de leur associer un symbole.

La difficulté de cette tâche provient de la multitude de réalisations possibles des sons et de leur caractère transitoire, entraînant une

grande difficulté à établir une classification des sons d'une langue et, par la même, à déterminer l'appartenance d'un son à une classe.

L'élément sonore minimal que va traiter le phonéticien appartiendra à une classe de sons phonétiquement semblables que l'on appelle phonème ; les phonèmes se définissant par des propriétés distinctives.

On peut remarquer que des essais de classification des sons, notamment par la classification ascendante hiérarchique (LELIEVRE-81) ont permis de retrouver la typologie des voyelles et des consonnes établie par les phonéticiens.

FANT (FANT-62) propose non plus une classification de phonèmes mais de segments de son (phones), unités plus petites ou de même dimension que l'unité phonétique, permettant de prendre en considération l'influence du contexte sur les phonèmes.

Mais un des intérêts de la notion de phonème réside dans la possibilité d'une notation phonétique :

Systèmes de notation phonétique que nous utilisons :

- API : alphabet phonétique international
- Système-maison

à voir sur le tableau I-1.

2.1.- Quelques notions d'acoustique .

Un son est une vibration créée par le mouvement d'un corps dans l'air. Une grande partie des sons que nous allons étudier sont des phénomènes quasi-périodiques.

Un son simple est caractérisé par sa fréquence de vibration qui détermine la hauteur du son et par son amplitude qui détermine l'intensité.

Les sons complexes, dont les vibrations périodiques résultent de la superposition de vibrations simples, sont à la base de nombreux sons de la parole.

La décomposition d'un son complexe en une série de sons simples se fait à l'aide de la transformée de FOURIER. Cette transformée est fondée sur le théorème de FOURIER qui indique que tout signal périodique peut être décomposé en une somme de sinusoïdes harmoniques, c'est-à-dire dont les fréquences suivent une progression arithmétique .

API	SYSTEME MAISON	MOT-CLE
a	A	Anne
ɑ	A	âne
e	Ai	fée
ɛ	Ei	fait
i	I	lit
o	AU	sot
ɔ	O	sotte
u	OU	loup
y	U	lu
ø	EU	peu
œ	E	peur
ə	E	le
ɑ̃	AN	lent
ɔ̃	ON	long
ɛ̃	IN	lin
œ̃	UN	l'un
p	P	pou
t	T	tou
k	K	cou
b	B	bout
d	D	doux
g	G	goût
m	M	mou
n	N	nous
ɲ	non étudié	agneau
ŋ	non étudié	camping
f	F	fou
s	S	rosse
ʃ	CH	chou
v	V	vous
z	Z	rose
ʒ	J	joue
l	L	loup
R	R	rue
w	W+	voir
j	I+	bailler
ɥ	U+	huit

Tableau I-1

Systèmes de notation phonétique utilisés.

Ainsi, les cordes vocales en vibration ont une fréquence que l'on appelle fréquence fondamentale F_0 et qui est la fréquence du premier terme du développement de FOURIER.

Le fondamental détermine la hauteur du ton.

La fréquence du 2ème terme, ou harmonique 2 est égale à $2F_0$. Il en va de même pour les autres harmoniques. L'amplitude des harmoniques décroît régulièrement de 12dB par octave. La représentation de ces fréquences en fonction de leur amplitude constitue un spectre de raie (figure I-3).

Les sons de la parole, signaux non-périodiques, sont analysés par l'intégrale de FOURIER pour obtenir alors un spectre continu, comme on le verra dans le chapitre II.

2.2.- Caractéristiques acoustiques et articulatoires des voyelles.

La plupart des systèmes de reconnaissance automatique de la parole, dont ceux que nous présentons au chapitre II travaillent sur les paramètres acoustiques.

Notons toutefois qu'il est possible d'extraire automatiquement des paramètres articulatoires ainsi que l'a montré SHIRAI (SHIRAI-81) et de s'en servir de façon satisfaisante pour la reconnaissance des voyelles dans la parole continue.

Font partie des voyelles orales françaises :
/ɑ/, /a/, /i/, /e/, /ɛ/, /ə/, /ϕ/, /œ/, /ɔ/, /o/, /u/, /y/

Font partie des voyelles nasales françaises :
/ɔ̃/, /ɑ̃/, /ɛ̃/, /œ̃/

* Du point de vue articulatoire on classe les voyelles du français d'après :

. l'aperture (distance minimale entre la langue et la voûte palatine) définissant des voyelles de petite, moyenne ou grande aperture.

. le lieu d'articulation définissant les voyelles antérieures (palatales) et postérieures (vélaires).

. la participation des lèvres qui permet de distinguer les voyelles labiales des non-labiales.

Les voyelles nasales s'articulent comme les voyelles orales dans la cavité buccale, avec en plus le voile du palais qui s'abaisse.

Le trapézoïde des voyelles françaises est donné sur la figure I-3.

* Du point de vue acoustique, les voyelles sont des sons voisés, résultat de l'interaction de l'onde glottique et du conduit vocal.

- Les formants

L'onde glottique est filtrée par le résonateur acoustique qu'est le conduit vocal possédant 4 ou 5 fréquences de résonance principales.

Ceci introduit dans le spectre résultat un certain nombre de maxima d'énergie à certaines fréquences que l'on appelle formants. (voir figure I-2).

Signalons que pour les voyelles nasalisées, la cavité résonante nasale fait apparaître un phénomène d'anti-résonance vis-à-vis des résonances orales.

- Les fréquences des formants

Les fréquences des formants sont fixées essentiellement par la forme du conduit vocal. Si ce conduit long de 17 cm pour un adulte mâle était un cylindre parfait, fermé au niveau de la glotte et ouvert au niveau des lèvres, les 4 premiers formants seraient proches de 500, 1 500, 2 500 et 3 500 Hz. Selon que le conduit vocal est plus long ou plus court ces fréquences sont diminuées ou augmentées.

Chaque formant est associé à une onde stationnaire, c'est-à-dire à une répartition géométrique fixe de variation de pression dont l'amplitude est maximale à l'extrémité de la glotte et est presque minimale au niveau des lèvres.

Le formant le plus grave correspond à $\frac{1}{4}$ de longueur d'onde, c'est-à-dire qu'un quart de longueur d'onde s'ajuste dans le conduit vocal. De la même façon, les second, troisième et quatrième formants correspondent respectivement à $\frac{3}{4}$, $\frac{5}{4}$ et $\frac{7}{4}$ de longueur d'onde.

Dans la parole courante les fréquences des formants varient en fonction de la voyelle que l'on cherche à produire, de l'entourage phonétique de cette voyelle, et du locuteur qui la produit.

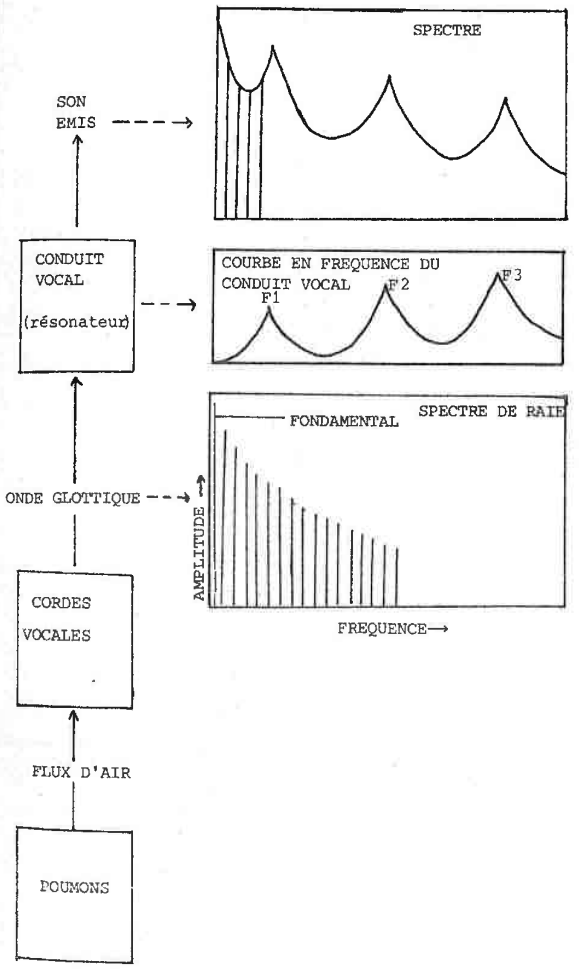


Figure I-2 : PRODUCTION D'UNE VOYELLE (d'après SUNDBERG-80)

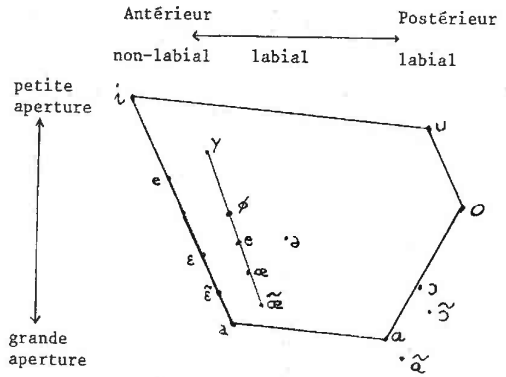


Figure I-3 : TRAPEZOIDE DES VOYELLES FRANCAISES

(d'après CARTON-79)

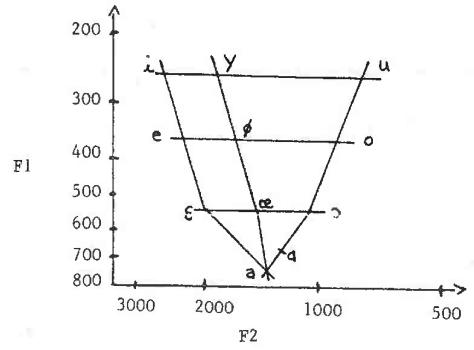


Figure I-4 : TRIANGLE ACOUSTIQUE DU FRANCAIS (d'après P. DELATTRE)

Les voyelles nasales, sur le plan spectral, apparaissent plus instables que les orales correspondantes. On peut se reporter aux conclusions de MRAYATI (MRAYATI-75) en ce qui concerne la présence de formants ainsi que leurs valeurs.

- Le triangle acoustique

La valeur des fréquences des formants permet un classement bien connu des voyelles. Notamment, l'utilisation des 2 premiers formants, pour la différenciation des voyelles orales françaises, se justifie à la vue, dans le plan F1-F2 à échelle logarithmique, de ce qu'on appelle le triangle acoustique des voyelles synthétiques du français (voir figure I-4). La figure I-5 illustre bien le lien qui existe entre la formation articulatoire des voyelles (anglaises) et leurs caractéristiques formantiques.
(Il est clair que ce triangle ne décrit pas la difficile réalité de la variabilité des voyelles).

La prise en compte de voyelles prononcées dans des contextes différents, avec ou sans accent , et dites par des personnes différentes nous amène à définir non plus des points, mais des zones dans le plan F1-F2.

Ainsi on voit sur la figure I-6, rapportée des travaux de PETERSON et BARNEY en 1951, la répartition dans le plan F1-F2 de 10 voyelles anglaises produites par 76 locuteurs comprenant des hommes, des femmes et des enfants.

Il faut noter que toutes ces voyelles ont été prononcées dans le même contexte /h/ - /d/ , cependant on se rend déjà compte de la grande dispersion d'une même voyelle dans le plan F1-F2.

- Classification acoustique des voyelles

La répartition de l'énergie dans le spectre, fortement liée aux formants, permet d'établir une classification phonétique des voyelles à partir de traits distinctifs.

Les systèmes vocaliques sont tous bâtis sur la double opposition grave/aigü, diffus/compact :
Le trait grave/aigü exprime la concentration de l'énergie dans les basses (respectivement hautes)fréquences .
Le trait diffus/compact exprime l'écartement important (respectivement faible) de deux concentrations d'énergie.

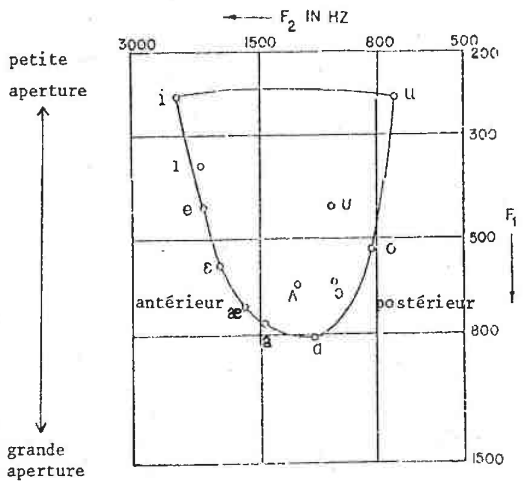


Figure I-5 : CARACTERISTIQUES ARTICULATOIRES ET FORMANTIQUES DES VOYELLES ANGLAISES (d'après BROAD-74)

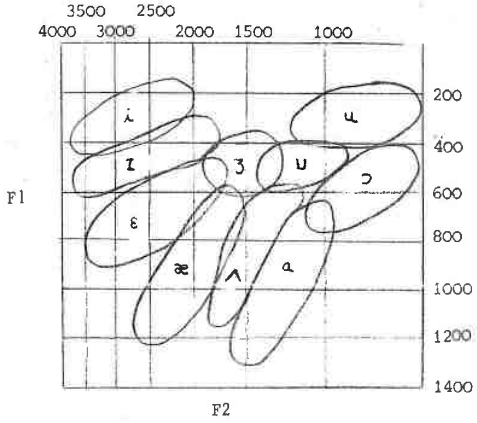


Figure I-6 : REPARTITION DANS LE PLAN F1-F2 DES VOYELLES ANGLAISES PRODUITES PAR 76 LOCUTEURS (d'après PETERSON-52)

D'autres traits distinctifs ont été proposés, tels que bémolisé diésé, nasal.

A partir de ces traits distinctifs on a pu établir des systèmes de classification des voyelles (voir figure I-7).

Certaines recherches sur la transcription phonétique automatique exploitent ces connaissances sur les traits acoustiques des sons, notamment en France (CAELEN-81, ROSSI-81, LAZREK-83, MERCIER-78).

2.3.- Caractéristiques acoustiques et articulatoires des consonnes.

Tandis que les voyelles sont caractérisées acoustiquement par l'absence de bruit audible, et au point de vue articulatoire par un passage d'air libre, les consonnes contiennent des bruits et se prononcent avec une fermeture ou un rétrécissement du passage de l'air. Elles peuvent être voisées dans le cas où le flux d'air provenant des poumons fait vibrer les cordes vocales, ou bien non voisées dans le cas contraire.

Le mode articulatoire, lié aux diverses sources d'excitation du conduit vocal, permet une classification intéressante que nous allons détailler en lui associant les caractéristiques spectrales des consonnes.

Une consonne peut être soit :

- occlusive : par fermeture complète mais momentanée du passage de l'air, suivi d'une explosion d'air. Font partie de la classe des occlusives :
/p/, /t/, /k/ : plosives sourdes.
Au moment de l'occlusion, si celle-ci est totale, il y a absence de signal, puis le spectre au moment de l'explosion (burst en anglais) devient très complexe en raison de la brièveté des bruits impulsionnels.
Précisons que pendant l'identification automatique des plosives sourdes, le burst peut jouer un rôle important, mais que quelquefois on peut trouver des plosives sourdes sans burst.
/b/, /d/, /g/ : plosives voisées.
A l'inverse des plosives sourdes, pendant l'occlusion, les cordes vocales continuent à vibrer. Il y a apparition d'une résonance grave qui produit le buzz ou onde basse fréquence (voir figure I-9).

	a	ɑ	ɔ	ɛ	œ	ɸ	e	o	i	y	u	ɑ̃	ɔ̃	ɛ̃	œ̃
Nasal	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+
Compact	+	+	-	-	-	-	-	-	-	-	-	o	o	o	o
Grave	-	+	+	-	-	+	-	-	-	-	+	+	+	-	-
Complex	o	o	+	+	+	+ −	+ −	+ −	-	-	-	o	o	o	o
Flat	o	o	o	-	+	o	-	+	-	+	o	-	+	-	+

Figure I-7 : Traits acoustiques du système vocalique du français (d'après MALMBERG-72)

	ɲ	m	n	g	k	b	p	d	t	ʒ	ʃ	v	f	z	s	l	R	j	w
nasal	+	+	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
vocalique	+	+	+	—	—	—	—	—	—	—	—	—	—	—	—	+	±	+	+
interrompu	+	+	+	+	+	+	+	+	+	—	—	—	—	—	—	+	±	—	—
continu	+	+	+	—	—	—	—	—	—	+	+	+	+	+	+	+	±	+	+
compact	+	—	—	+	+	—	—	—	—	+	+	—	—	—	—	—	±	+	—
aigu	+	—	+	±	±	—	—	+	+	+	+	—	—	+	+	+	±	+	—
voisé	+	+	+	+	—	+	—	+	—	+	—	+	—	+	—	+	+	+	+

Figure I-8 : Analyse binaire, en traits acoustiques, du système consonantique français (d'après ROSSI-77)

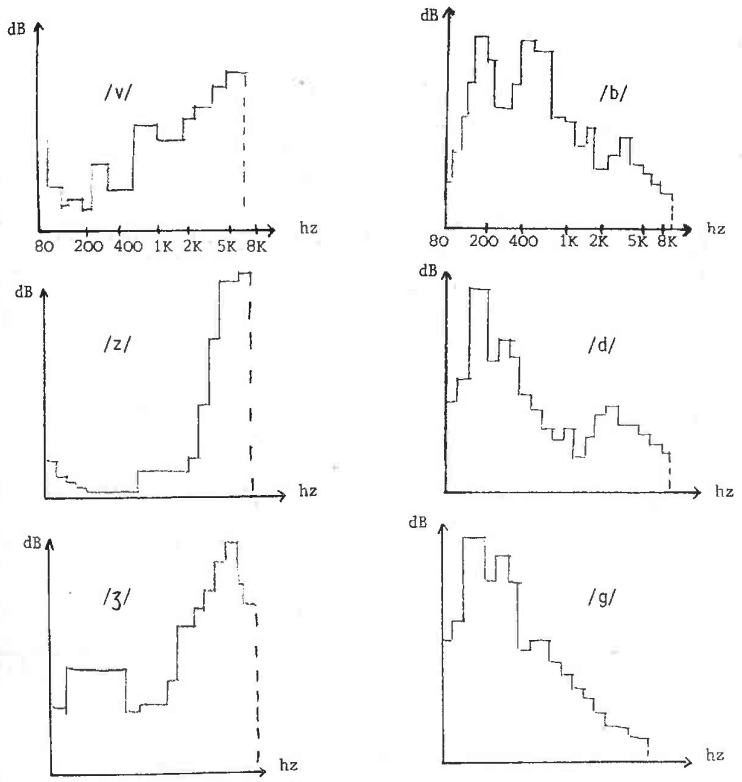


Figure I-9 : SPECTRES EN 1/3 D'OCTAVE DES FRICATIVES ET PLOSIVES VOISEES (d'après MARCHAL-80)

- constrictive : par passage de l'air resserré en un ou plusieurs endroits. Au plan spectral on relève des bruits de friction dans des bandes de fréquences plus ou moins aigües.
Font partie de la classe des constrictives :
/f/, /s/, /ʃ/ : fricatives sourdes.
/s/ a plusieurs pôles de bruit dont le plus important se situe vers les hautes fréquences (≃ 5 000 Hz).
/ʃ/ a un pôle de bruit prépondérant à moyenne fréquence (≃ 2 500 Hz).
/f/ est plus instable et a une concentration de bruit en basses et moyennes fréquences (≃ 1 700 Hz).
/v/, /z/, /ʒ/ : fricatives voisées.
Les fricatives voisées présentent des pôles de bruit semblables à leurs homologues sourdes auxquels s'ajoute en très basse fréquence la contribution du voisement. (voir figure I-9).

- une sonante : par passage de l'air plus facilement franchissable que pour une constrictive, entrainant un phénomène de résonance.
Une sonante est presque toujours sonore.
Font partie de la classe des sonantes :
/m/, /n/, /ɲ/ : nasales.
Elles sont produites par une source périodique et un bruit impulsionnel, l'énergie laryngienne étant dérivée dans le canal par suite de l'occlusion du canal buccal. Il existe en moyenne une résonance à 250 Hz, murmure nasale, celle-ci étant dépendante de la géométrie des cavités nasales du locuteur, et des anti-résonances, plus difficiles à localiser.

/l/, /r/ : liquides.
/l/ sur le plan spectral est fluide, et est très sensible au contexte vocalique.
/r/ possède trois principaux variphones qui sont :

/r/ qui est roulé, propre à certains parlers régionaux
/R/ qui est grasseyé par des battements au niveau du voile du palais
Ces deux variantes sont caractérisées par des occlusions répétées
/ɾ/ est produit sans battement, souvent entre deux voyelles.
Il est clair que le spectre de ces différentes variantes de /r/ est fort différent. Cependant la concentration de l'énergie au centre du spectre entre 1 000 et 1 500 Hz est une caractéristique générale du /r/.

Pour une classification plus fine des consonnes au niveau articulatoire il importe de connaître le lieu de l'articulation et l'organe articulant, ce que nous n'exposons pas ici. On pourra se reporter aux travaux de SHIRAI & al. (SHIRAI-82) pour l'étude de la reconnaissance des consonnes à partir de paramètres articulatoires.

On peut comme pour les voyelles établir une classification des consonnes fondée sur des traits acoustiques (voir figure I-8).

2.4.- Variabilité phonétique.

Il est impropre de considérer un continuum de parole comme un ensemble d'unités fixes et invariables (voyelles ou consonnes) alignées les unes à côté des autres (MALMBERG-79).

Dans une chaîne de son, la production de chaque son est fortement liée à la production des sons voisins. Ainsi, pour un même phonème, les positions articulatoires peuvent être différentes selon son environnement : c'est l'effet de coarticulation.

Certains des traits distinctifs, articulatoires ou acoustiques des phonèmes isolés que nous avons cités dans les paragraphes précédents, demeurent exactes pour les phonèmes en contexte bien que les spectres de ces derniers soient très influencés par leur environnement.

Certains phonèmes, notamment les consonnes y gagnent des caractéristiques supplémentaires (En contexte CV on distingue les plosives "dures" des plosives "douces") (CAELEN-79).

Le problème actuellement est de connaître de manière rigoureuse toutes les influences sur le plan acoustique des phonèmes dans tous les contextes possibles.

En voici quelques exemples :

Pour les voyelles :

. Deux voyelles adjacentes peuvent subir des phénomènes d'assimilation aux traits fermé/ouvert, diffus/ compact
exemple : dans RAOUL, /u/ tend a être ouvert.

. /r/ a une grande influence et tend à compacifier certaines voyelles comme /a/, /o/.

. Les effets des nasales et des fricatives sont faibles ;
Les voyelles aiguës étant les plus influencées par les fricatives (tendance à élever F2).
Voir sur le tableau I-2 les résultats sur les formants F1, F2 de la perturbation de l'articulation des voyelles par un contexte consonnantique.

Pour les consonnes :

Presque toutes les consonnes du français ont tendance à changer de lieu d'articulation selon les voyelles qui les entourent ; les cas extrêmes étant /k/ et /g/.

Ces influences articulatoires ont été groupées de la façon suivante : palatisation, velarisation, labialisation, etc.

Voici quelques résultats sur le plan spectral de ces différentes modifications articulatoires.

- Le bruit caractéristique du /k/ en contexte /k i/ est plus élevé en fréquence, qu'en contexte /k u/
- /ℓ/ est fortement influencé par un contexte inter-vocalique (LAZREK-83)
- /r/ peut subir un assourdissement après une plosive sourde.
- les formants des fricatives varient en fonction des voyelles dans le contexte CV (SOLI-81).

Les modifications que les sons subissent au contact d'autres sons ne sont pas, en général, de nature à changer les qualités essentielles de ceux-ci.

Mais il peut arriver que ces modifications aillent plus loin et changent des qualités plus importantes des sons. Ceci est dû à une certaine "paresse" du locuteur qui cherche à obtenir le maximum d'effet acoustique avec un minimum d'effort articulatoire.

Nous verrons plus loin certaines de ces modifications des mots d'une langue.

	r		L		nasales		plosives-fricatives labiales		palatales		moyenne		
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	
	427	800	350	794	292	680	343	591	330	1020	357	845	u
	517	852	502	1030	485	895	510	867	435	940	492	963	o
	722	1180	620	1360	680	1427	644	1225	572	1500	650	1320	a
	523	1080	444	1396	552	1492	470	1350	430	1480	460	1386	ə
	600	1530	600	1570	645	1530	538	1580	530	1550	600	1570	ɛ
	354	1693	387	1643	335	1756	326	1620	264	1789	325	1734	y
	500	1600	450	1700	466	1990	455	1534	357	1533	410	1720	e
	360	2100	317	1973	268	2066	300	1900	325	2050	322	2080	i

Tableau I-2 : Valeurs moyennes des 2 premiers formants des voyelles en contexte CV (d'après CAELEN-79)

3.- VARIABILITE DE LA PAROLE INTER ET INTRA LOCUTEUR.

Le problème de la variabilité de la parole a, la plupart du temps, été évacué des études systématiques des phonéticiens (BÖE-80.a). Les études et les résultats sur ce sujet proviennent essentiellement de recherches sur la parole qui ont été obtenus par :

- les psychologues pour l'analyse du processus humain de reconnaissance d'un locuteur et pour l'établissement de l'état émotionnel d'un locuteur à partir de mesures faites sur la parole.
- les spécialistes de la production de la parole pour l'établissement des aspects physiologiques d'un locuteur en fonction de mesures effectuées sur sa voix.
- les chercheurs qui s'intéressent à l'identification ou la vérification automatique de locuteurs (CORSI-79) et à la reconnaissance automatique de la parole multilocuteur.
- les spécialistes de la perception de la parole.

Les résultats des travaux d'apprentissage automatique des phonèmes nous ont permis d'étudier les variations des spectres vocodeur (sur lesquels nous effectuons la reconnaissance automatique de la parole) de 13 voyelles de 15 locuteurs, dont 6 femmes et 9 hommes.

Sur les schémas qui suivent, les graphiques représentés permettent de visualiser les mesures obtenues à l'aide du vocoder SLE en portant à l'abscisse i et l'ordonné y_i l'énergie mesurée sur le canal i. Les courbes ou "spectres" sont obtenues par interpolation linéaire entre ces points. La description du vocoder utilisé se trouve dans le chapitre II §.1.1.c.

Les courbes des locuteurs masculins sont représentées par des tirets et celles des locuteurs féminins par des pointillés.

On y remarque la grande variabilité des courbes brutes pour les voyelles /a/ et /u/, /i/ et /y/. (voir les autres voyelles en annexe 1)

Les courbes représentées sur la partie droite sont des essais de normalisation de l'énergie dans les différents canaux, avec,

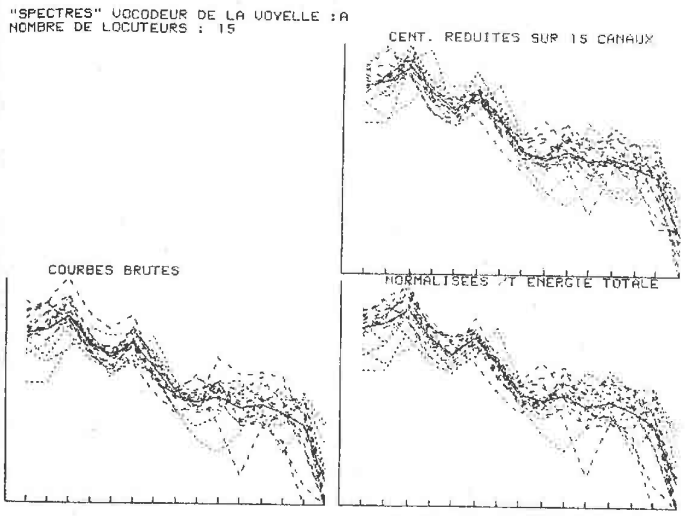
en haut : les énergies centrées réduites à partir de la moyenne et l'écart-type des énergies des 15 canaux,

en bas : les énergies normalisées par rapport à la somme des énergies des 15 canaux.

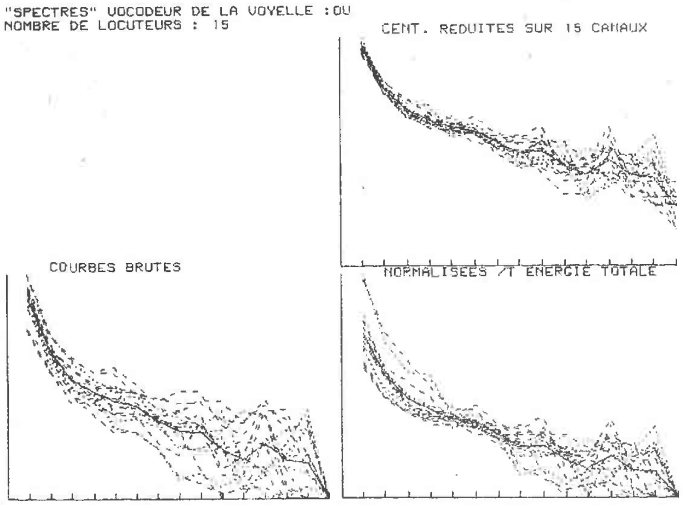
Nous verrons dans le chapitre III l'effet de ces normalisations.

Représentation des sorties VOCODER à partir des 3 types de données utilisées : brutes, normalisées par rapport à l'énergie totale, centrées réduites.

VOYELLE /a/

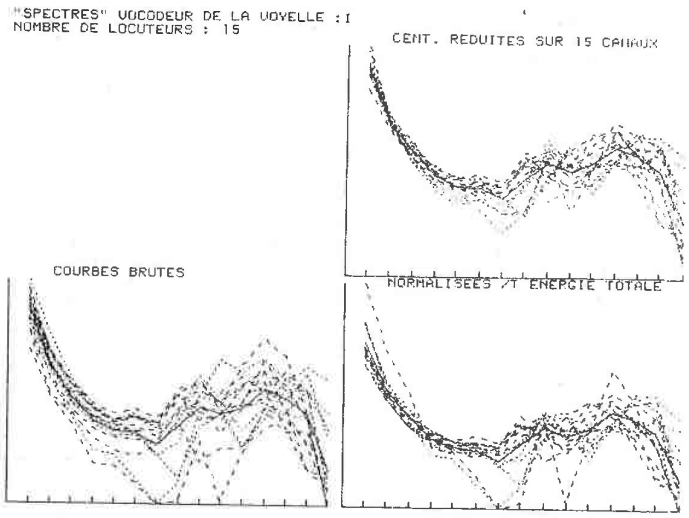


VOYELLE /u/

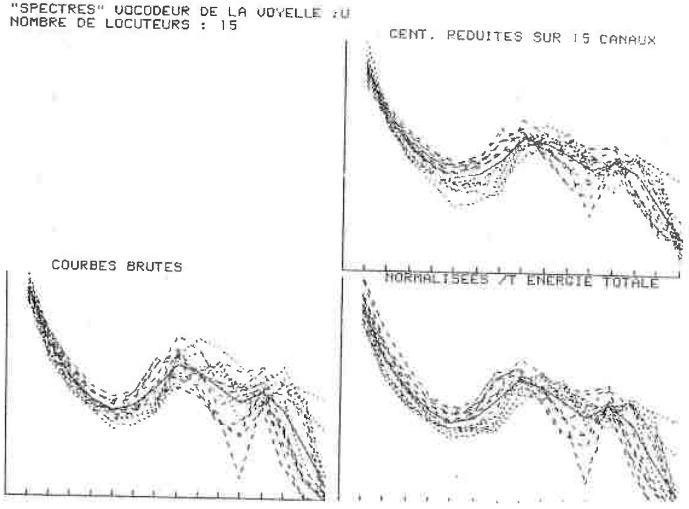


Représentation des sorties VOCODER à partir des trois types de données utilisées : brutes, normalisées par rapport à l'énergie totale, centrées réduites.

VOYELLE /i/



VOYELLE /y/



Dans la suite nous allons évoquer les principales sources de variabilité de la parole, accompagnées si possible des variations acoustiques, phonétiques et prosodiques.

3.1.- Différences anatomiques de l'appareil vocal et influence sur la parole.

Les différences anatomiques sont une des sources de variabilités dont l'influence sur les propriétés acoustiques des sons ont été le plus étudiées (STEVENS-71).

Ces différences, conditionnées par l'âge, le sexe et l'hérédité, atteignent à la fois le système respiratoire, le larynx, le conduit vocal et les cavités nasales.

a) La Fréquence du fondamental :

α) Sources de variabilité de la fréquence du fondamental.

- Les cordes vocales sont probablement responsables de la plus grande cause de variabilité inter-locuteur.

Ainsi la fréquence du fondamental est directement déterminée par la longueur et par la tension des cordes vocales. Cette tension est fonction de l'élasticité, la masse et la forme de celles-ci.

Plus les cordes vocales sont longues et épaisses, plus les vibrations sont lentes. Plus elles sont courtes et minces, plus la fréquence devient grande.

Une humidité excessive ou trop réduite des cordes vocales modifie la tension de celles-ci ; ou bien une excroissance sur une corde vocale peut donner des irrégularités dans la périodicité du signal.

- Le système respiratoire varie énormément d'un individu à l'autre, quant à la dimension et l'élasticité des voies aériennes entraînant une différence dans la résistance de ces dernières et quant à la taille des poumons entraînant une différence de capacité vitale. Tous ces facteurs déterminent la pression sous-glottique d'un individu.

Toutefois la variation du système respiratoire n'a qu'un effet secondaire sur les sons de la parole. Par exemple, plus la pression de l'air dans les poumons est forte, plus les cordes vocales sont amincies et tendues et plus haute est la fréquence de vibration de ces cordes.

Pour un individu, la modification des caractéristiques de ses poumons influence le contour et l'étendue de la fréquence du fondamental utilisée quand il parle (STEVENS-71).

ß) Variation de la fréquence du fondamental.

La vitesse de vibration des cordes vocales varie entre environ 60-70 Hz pour les voix masculines les plus basses et 1 200-1 300 Hz limite supérieure d'un soprano.

La moyenne est de 200 à 300 Hz pour une femme,
et de 100 à 150 Hz pour un homme.

Pour les adolescents, des différences systématiques ont été observées entre 15 et 18 ans.

Il y a des différences considérables dans la fréquence du fondamental et dans sa distribution pour des individus de même sexe et de même âge comme on le voit sur la figure I-10.

Pour un locuteur donné, la fréquence du fondamental, l'étendue et les formes des contours de la fréquence du fondamental pour une phrase donnée sont fortement influencées par son état émotionnel, ce que nous verrons un peu plus loin.

b) Les formants :

α) Sources de variabilité

- Le conduit vocal peut être responsable d'une variabilité considérable inter-locuteur, de par sa taille et par sa forme.

Les fréquences des formants des voyelles orales sont directement liées à la longueur du conduit vocal (cf §.2.2 de ce chapitre).

En prenant les valeurs moyennes des fréquences des formants pour un nombre important de voyelles, on peut trouver une indication sur la longueur moyenne du conduit vocal d'un individu.

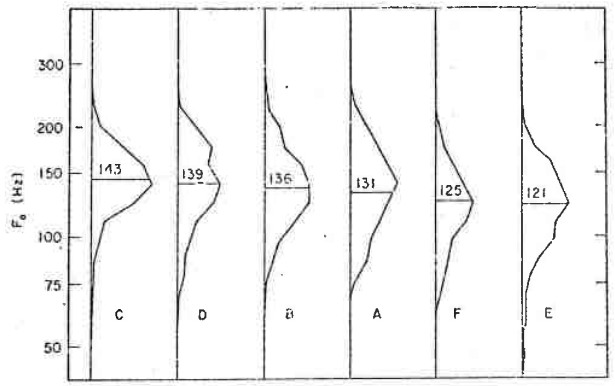


Figure I-10 : DISTRIBUTION DE LA FREQUENCE DU FONDAMENTAL PRODUITES PAR 6 LOCUTEURS MASCULINS (FAIRBANKS-40)

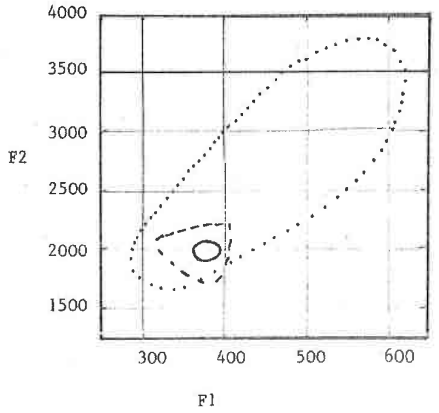


Figure I-11 : VARIATION DE F1, F2 DE LA VOYELLE /I/ EN FONCTION DES CONTEXTES ET DES LOCUTEURS (d'après BROAD-74)

Il est intéressant de noter que pour deux locuteurs distincts ayant des conduits vocaux de longueurs différentes, les variations de longueur de voyelle à voyelle sont les mêmes (ZERLING-80).

Tout changement de la section du conduit vocal déplace la fréquence de chaque formant. La direction du déplacement, vers les hautes ou basses fréquences dépend du point de l'onde stationnaire où à lieu ce changement.

Mais d'après ZERLING des différences systématiques du diamètre du conduit vocal pour deux locuteurs peuvent n'avoir à elles seules aucune influence au plan acoustique, c'est la variation relative de l'aire qui importe

- Les cavités nasales, comme on l'a vu, jouent un rôle important dans la production des voyelles nasales. La taille et la configuration de ces cavités diffèrent considérablement d'un sujet à l'autre faisant varier les caractéristiques acoustiques des sons nasalisés.

β) Variation des formants des voyelles orales.

Pour mesurer l'étendue de cette variation on peut déjà se reporter aux travaux complets pour l'anglais de PETERSON et BARNEY dont nous avons déjà présenté un résultat sur la figure I-6.

Sur la figure I-11 on se rend compte de la variabilité des deux premiers formants F1, F2 de la voyelle anglaise /I/. Ainsi, la courbe en trait plein représente la distribution de la voyelle répétée par le même locuteur dans le même contexte ; la courbe en tiret représente la distribution de la voyelle prononcée par le même locuteur dans 576 contextes consonantiques différents ; la courbe en pointillé représente la distribution de la voyelle dans le contexte /h/-/d/ prononcée par 76 locuteurs.

La figure I-12 représente de façon plus détaillée la répartition des trois premiers formants de 12 voyelles françaises produites en parole courante par trois locuteurs dont un féminin (rectangles blancs) et deux masculins (rectangles noirs et grisés).

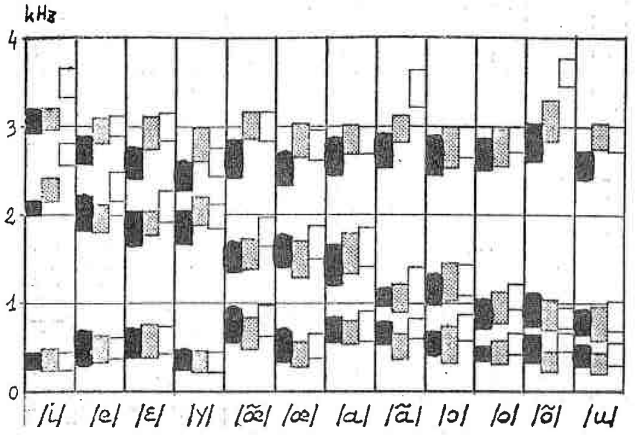


Figure I-12 : VARIATION DES FORMANTS POUR 3 LOCUTEURS (d'après LIENARD-72)

masculin
féminin

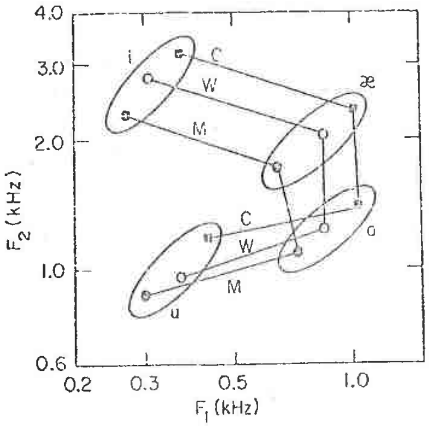


Figure I-13 : VALEUR MOYENNE DE F1, F2 POUR 4 VOYELLES DE LOCUTEURS MASCULINS M, FEMININS W et ENFANTS C (d'après PETERSON-52)

On y remarque la tendance qu'ontles formants "féminins" à être de fréquence plus élevée que les formants respectifs "masculins". De plus les variations entre même formant pour deux locuteurs ne sont pas les mêmes quand on passe d'une voyelle à l'autre.

On peut trouver une autre représentation de ces mêmes caractères sur la figure I-13 pour les formants moyens F1 et F2 de locuteurs masculins, féminins et enfants pour les voyelles /i/, /æ/, /a/ et /u/ dans le contexte /h/-/d/.

La variation de F3 est en général beaucoup moins dépendante, pour un locuteur donné, de la voyelle prononcée. C'est pourquoi la valeur de F3 est pertinente pour la recherche de la longueur du conduit vocal.

Quant aux largeurs de bande des formants, on a remarqué que pour la voyelle /i/, elles présentaient de grandes différences inter-locuteur. Pour certains locuteurs c'est le troisième formant qui a une grande largeur de bande, pour d'autres c'est le second formant.

En reconnaissance automatique de la parole, l'identification des voyelles se faisant très souvent à partir de F1 et F2, de nombreuses techniques de normalisation de formants ont été utilisées (cf. chapitre II) afin de réduire ces variations dans les systèmes multilocuteurs.

Parmi les courbes des locuteurs que nous avons étudiés on retrouve, ce décalage vers les hautes fréquences, des pics d'énergie dans les spectres de voix de femmes, par rapport aux voix d'hommes comme le montre la figure I-14 qui représente les spectres de la voyelle /$\tilde{\varepsilon}$/ pour deux locuteurs de sexe différent.

Ceci se confirme à la vue des spectres vocodeurs moyens homme - femme de la figure I-16.

Cependant, comme le montre la figure I-15 il est possible de trouver des spectres quasiment identiques pour une même voyelle prononcée par deux locuteurs de sexe différent.

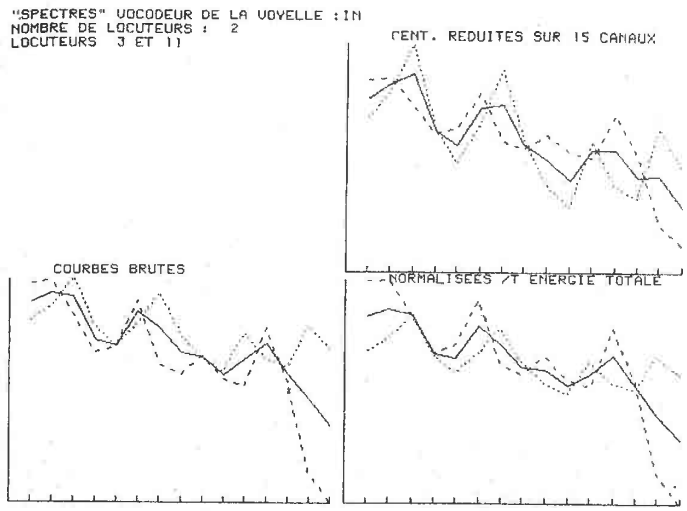


Figure I-14 : voyelle /$\tilde{\varepsilon}$/

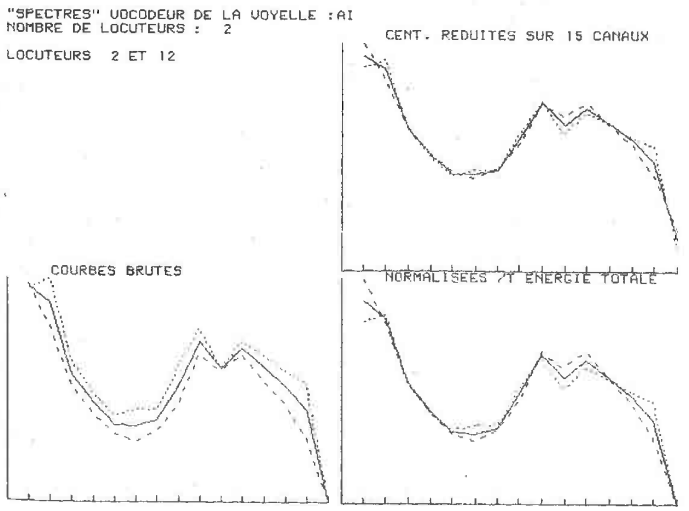


Figure I-15 : voyelle /e/

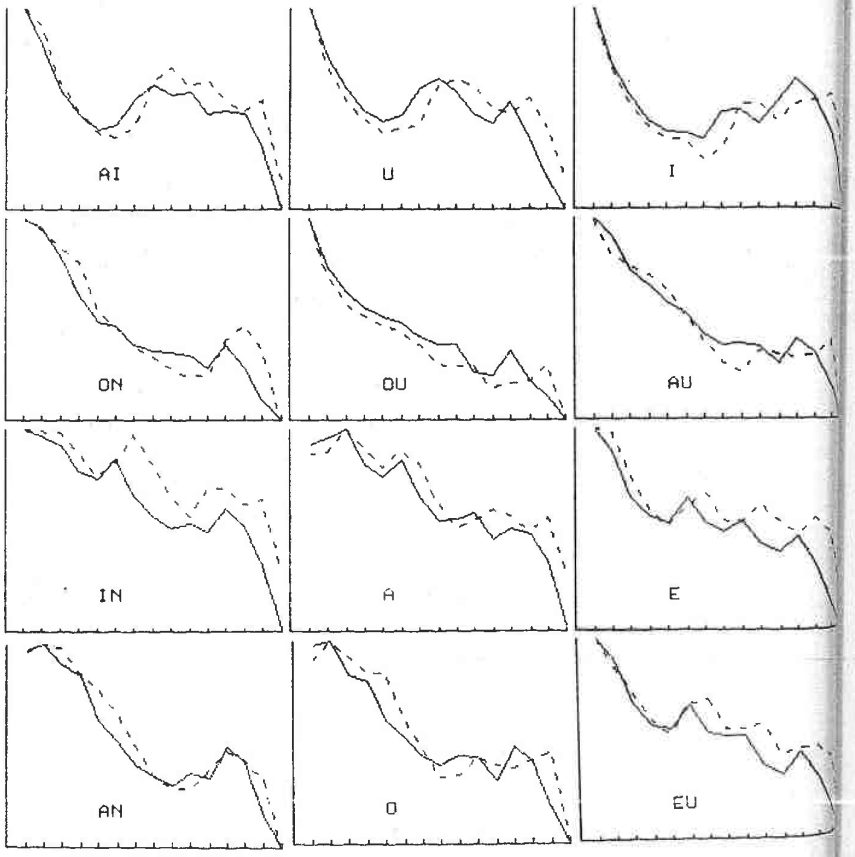


Figure I-16 "SPECTRES" VOCODEUR MOYENS DE ---- 6 femmes
—— 9 hommes

c.- Les consonnes.

La variabilité des consonnes en fonction de leur environnement n'a pas encore été étudiée de manière systématique et satisfaisante pour un seul locuteur.

On imagine alors facilement le peu de résultat concernant la variabilité et les variations acoustiques des consonnes inter-locuteurs. On trouvera quelques résultats dans (STEVENS-71).

3.2.- Différences acquises.

Les habitudes de prononciation font partie des différences acquises, dépendantes de l'origine géographique, du milieu social et culturel du locuteur.

La plupart des informations concernant ce problème proviennent essentiellement de la phonologie, et plus particulièrement de l'étude des altérations phonologiques des phonèmes ou des mots en contexte.

a) Relatives à un individu.

α) Différences articulatoires

Il existe deux aspects des différences articulatoires inter-locuteur :

- celles qui tendent à diminuer au maximum les différences de réalisations acoustiques dues aux différences morphologiques : sorte de compensation articulatoire inter-locuteur
- celles qui dépendent des habitudes ou des comportements phonatoires propres à chaque individu :

Prenons l'exemple du lieu occlusif de [g] en contexte [i] qui est post-palatal pour certains sujets alors qu'il est pré-palatal chez d'autres sujets (ZERLING-80) ; cette différence n'ayant pas un effet notoire sur le plan acoustique.

Par contre il existe des altérations provoquées par l'influence mutuelle des mouvements articulatoires voisins, qui seront plus ou moins importantes selon les locuteurs, et qui modifient considérablement la structure phonétique des mots ou phrases énoncés. En font partie : les assimilations, fusions, anticipations des phonèmes à l'intérieur des mots et les élisions, insertions, substitutions des phonèmes à la jonction des mots.

Les assimilations sont souvent liées à la tendance à l'économie articulatoire et à un manque de coordination des mouvements articulatoires. Il arrive souvent que les vibrations des cordes vocales commencent (ou cessent) trop tôt ou trop tard parce que l'entrée en action des différents organes n'est pas synchrone.

La représentation acoustique sous forme symbolique, accompagnée de la fréquence du fondamental, de la figure I-17, illustre le dévoisement d'une partie de /ʒ/ qu'une locutrice effectue systématiquement lors de la production de [ʒə] .

La figure I-18 illustre le dévoisement plus naturel de /r/ dans le contexte [kr] .

Voici d'autres exemples :

- Dans CHAPDELAINE [ʃapdəlɛnə] le [p] peut être sonorisé ⇒ [b]
- Dans MESSE DE MINUIT [mɛsəd̥əminyi] il peut y avoir élision du [ə] et sonorisation du [s] en [z]
- Dans ABSURDE [absyrd] il peut y avoir assourdissement du [b] en [p]
- Dans MADEMOISELLE [madəm zɛl] [nm] il peut y avoir élision du [ə] et nasalisation du [d]

On a également constaté qu'il existe des locuteurs parlant avec une certaine stabilité, c'est-à-dire qu'ils reproduisent les sons ou les mots de manière relativement semblable. A l'opposé certains locuteurs, sans changer de condition de locution, possèdent une grande variabilité de réalisation de la parole (RABINER-80).

β) Différences prosodiques.

Les deux principaux faits de la prosodie, l'intonation et l'accentuation sont fortement liés à la langue du locuteur.

Toutefois, pour un groupe d'individus parlant la même langue, prononçant la même phrase dans les mêmes conditions on trouve des différences prosodiques (MARTIN-80).

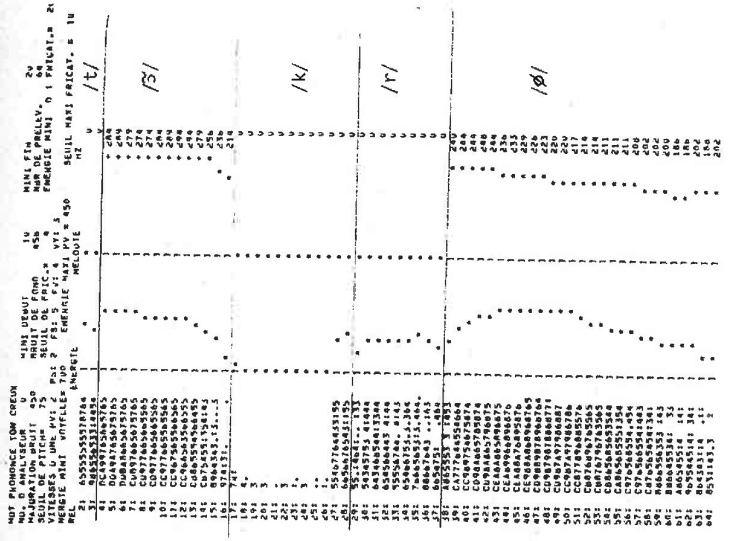


Figure I-18 : Exemple de dévoisement de /r/

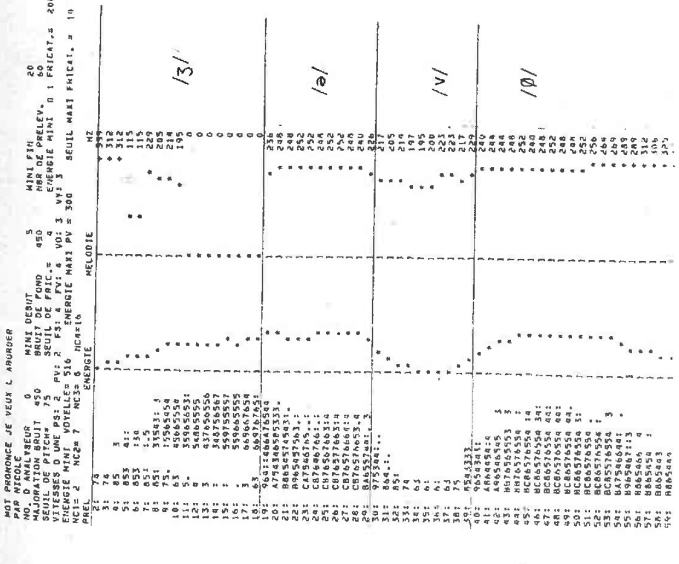


Figure I-17 : Exemple de dévoisement d'une partie de /ʒ/

D'autres caractères prosodiques comme la durée, les pauses, le débit, liés à la vitesse d'élocution, ou même le rythme et l'intensité sonore du message parlé présentent autant de variations qu'il est nécessaire de prendre en compte lors de la reconnaissance automatique de la parole.

Les problèmes de variations de vitesse d'élocution ont été résolus de manière satisfaisante par la programmation dynamique.

La répercussion de la variation de l'intensité d'élocution sur les spectres VOCODEUR est bien illustrée sur les figures I-19 et I-20 où l'on remarque non seulement des différences de répartition de l'énergie dans chaque bande de fréquence mais aussi la différence d'énergie globale des voyelles /u/ et /ɛ/.

Dans les systèmes de reconnaissance automatique de la parole les problèmes d'intensité variable sont, en général, pris en compte par des procédures de normalisation (chapitre II).

b) Relatives à un dialecte.

Les accents régionaux peuvent être un facteur de variabilité important. Citons, par exemple, l'existence d'une voyelle supplémentaire /œ̃/ dans le système phonologique marseillais, par rapport au système phonologique parisien. Voici quelques altérations phonologiques dues aux accents régionaux :

LUNDI	:	[lœ̃di]	ou	[lœ̃ndi]
POTAGE	:	[potaʒə]	ou	[pɔtaʒə]
MARIE	:	[maRi]	ou	[mari]
PERE	:	[perə]	ou	[pɛrə[
SEPTEMBRE	:	[setɑ̃bRə]	ou	[septɑ̃bRə] .

Notons que pour la constitution de la base de données du GRECO on tente d'utiliser dans un premier temps des voix françaises "sans accent".

Figure I-19 : Répercussion de l'intensité d'élocution sur les différents spectres.

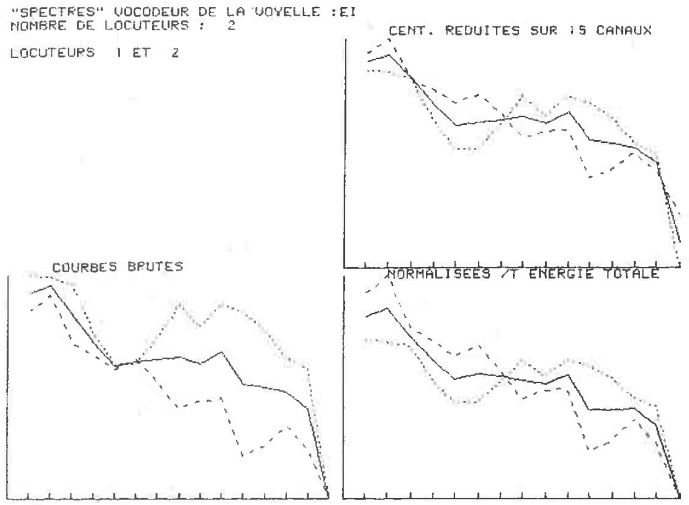
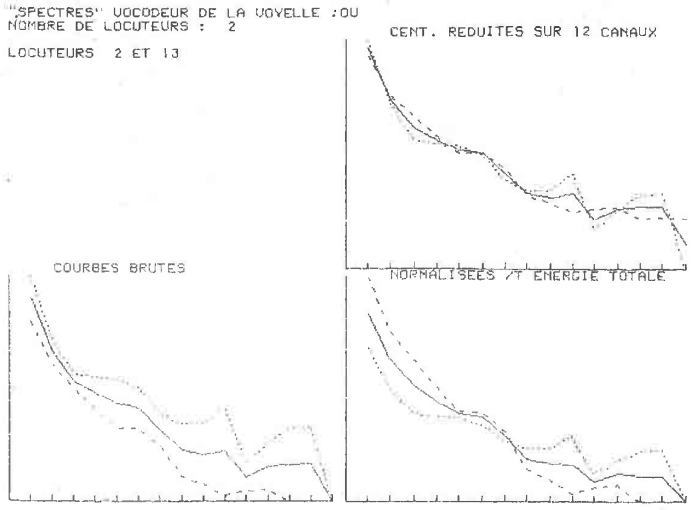


Figure I-20 : Répercussion de l'intensité de l'élocution sur les différents spectres.



c) Relatives à une langue.

Le locuteur natif va être marqué par :

- le système vocalique et consonnantique
- la base articulatoire
- la prosodie

de sa langue, la base articulatoire étant l'ensemble des habitudes articulatoires qui caractérisent une langue. Notons que toute l'articulation française est caractérisée par une tendance antérieure, et dominée par l'articulation labiale, se trouvant ainsi à l'opposé de la langue anglaise.

Pour des détails sur l'accent du français comparé avec l'accent d'autres langues on pourra lire (CARTON-79).

3.3.- Différences conditionnées par une situation particulière.

a) Etat émotionnel.

Un changement de l'état émotionnel du locuteur provoque des modifications anatomiques de courtes durées.

WILLIAMS & al.(WILLIAMS-70) ont montré les effets, sur la fréquence du fondamental, de 5 émotions soient la tristesse, la joie, la peur, la colère, et un état normal.

En général, pour la peur et la colère la fréquence du fondamental augmente et pour la tristesse elle diminue. Ces variations restant tout de même dépendantes du locuteur.

En ce qui concerne les formants, c'est surtout le premier formant de la première syllabe des mots qui est influencé par l'état émotionnel du locuteur.

Bien qu'en reconnaissance de la parole on se limite à un état "neutre" du locuteur, il faut tout de même prendre en considération la déformation de la voix volontaire ou non, d'un locuteur face à une machine.

b) Type de voix adopté.

La voix d'un locuteur dépend également des modifications volontaires qu'il effectue pour faire face à une situation particulière ; par exemple :

- l'augmentation de l'intensité pour les sons criés,
- la diminution de l'intensité et l'absence de vibrations des cordes vocales pour les sons chuchotés,
- la modification du conduit vocal faisant apparaître un "formant de chant" pour la voix de certains chanteurs d'opéra.

c) La maladie.

Evoquons juste ici, quelques maladies qui provoquent une modification passagère du système phonatoire, dont le rhume, la laryngite, pharyngite entraînant la nasalisation des sons, ou une modification de la fréquence du fondamental, etc.

3.4.- Voix pathologiques.

En dernier lieu, il faut tenir compte, dans les sources de variabilité de la parole inter-locuteur, de certaines anomalies physiques ou neurologiques du locuteur.

La surdité, facteur important des troubles du langage est la cause, entre autre, . d'altération de la voix : altération du timbre, voix trop grave, trop aigüe, altérations mélodico-rythmiques,
. d'altérations phonétiques : imprécision des zones formantiques, voyelles trop intenses... (AIMARD-74).

D'autres troubles de la parole, d'origine physique, dont les malformations bucco-faciales, ou les maloclusions labiales affectent la production de la parole.

Dans notre laboratoire, M.C. HATON, après avoir mené une étude sur les voix pathologiques, a développé le système SIRENE, système interactif de rééducation vocale des enfants non-entendants, dans le cadre d'une thèse d'Etat.

CONCLUSION :

Le signal de parole véhicule de nombreuses informations que l'on peut répartir en trois catégories :

- des informations sur le message,
- des informations sur l'identité du locuteur,
- des informations sur l'état du locuteur.

Une tâche à accomplir lors de tout traitement d'un signal de parole est la séparation de ces diverses informations.

Par la suite nous ne nous intéresserons qu'aux deux premières catégories d'informations, considérant uniquement des locuteurs dans un état "normal".

Il nous reste à tenir compte, pour faire de la reconnaissance automatique de la parole multilocuteur, des différentes variations du message parlé que nous venons d'exposer dans ce premier chapitre et que l'on peut classer ainsi :

. Les variations acoustiques : variations spectrales et variation de la fréquence du fondamental des sons.
. Les variations phonétiques : variations des phonèmes en contexte.
. Les variations prosodiques : variation de la fréquence du fondamental et variations d'intensité et de durée.
. Les variations phonologiques : variation des mots ou groupe de mots prononcés.

C H A P I T R E II

LES SYSTEMES DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE MULTILOCUTEUR

INTRODUCTION.

Dans la première partie de ce chapitre nous présentons différentes techniques et méthodes utilisées jusqu'à présent pour résoudre le problème de la reconnaissance automatique de la parole.

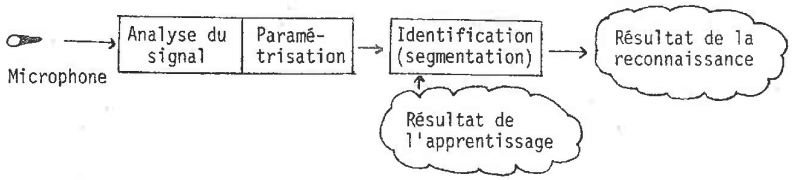
Cette description est assortie si possible du comportement de ces systèmes face à un changement de locuteur.

Ceci permet d'aborder les solutions proposées actuellement pour les systèmes de reconnaissance de la parole multilocuteur dont nous faisons le point dans la deuxième partie du chapitre.

1.- LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE.

Schématiquement l'organisation générale d'un système de reconnaissance de la parole comporte trois phases principales :

- l'analyse et la paramétrisation du signal
- la reconnaissance
- l'apprentissage



L'analyse et la paramétrisation constituent la phase de représentation du signal sous forme de données numériques.

La reconnaissance consiste à identifier les formes élémentaires (mots, phonèmes, spectres à court terme) à l'aide desquelles on identifie si nécessaire des formes moins élémentaires.

L'identification est quelquefois accompagnée d'une phase de segmentation.

L'apprentissage consiste ici à mémoriser : les formes à reconnaître, ou la description de ces formes par des règles (structurelles, contextuelles ou autres) ou encore les paramètres de la reconnaissance.

1.1.- Analyse du signal et paramétrisation

a) Codage de l'onde sonore.

Dans les systèmes de reconnaissance automatique de la parole le capteur est un microphone qui permet d'étudier sous forme électrique les variations de la pression acoustique du signal vocal.

L'analyse de la parole consiste alors à extraire du signal acoustique un faible nombre de paramètres pertinents. Le signal de la parole à la sortie du microphone est une courbe temporelle très complexe puisqu'elle fait figurer des parties quasi-périodiques (sons-sonores) et des segments fortement bruités (sons sourds) (figure II-1).

La paramétrisation constitue une réduction de la représentation (numérique) du signal, nécessaire pour des traitements efficaces et rapides de reconnaissance en vue d'applications en temps réel.

Cette réduction est rendue possible, sans altérer la représentation du signal, par la redondance naturelle du signal vocal pour coder le sens du message (HATON-74).

Notons que la réduction de la redondance peut effacer certaines informations caractérisant le locuteur.

De nombreuses méthodes d'analyse et de paramétrisation du signal sont actuellement utilisées.

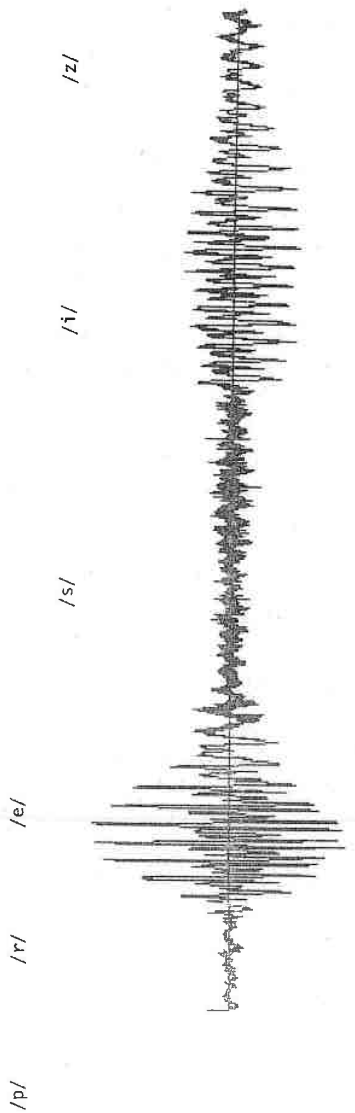


Figure I-1 : Le signal de la parole

On trouvera de manière détaillée la description de la plupart de ces méthodes dans (SCHAFER-74, GUEGEN-76). On les sépare fréquemment en deux catégories bien que cette distinction soit discutable : les méthodes d'analyse temporelle et les méthodes d'analyse fréquentielle.

Ces méthodes ont été guidées soit par une approche de modélisation du processus de production du signal vocal (LPC, Cepstres) soit par une approche de modélisation du système d'audition, ou bien par une approche spectrale conventionnelle.

Tout traitement sur le signal nécessite une première phase de codage de l'onde sonore, effectuée par un échantillonage. Pour être assuré de pouvoir reconstituer exactement le signal original à partir du signal échantilloné il faut (théorème de SHANNON) que, si la largeur de bande du signal vocal (signal à bande limitée) est F, la fréquence d'échantillonage de ce signal soit égale à $2F$.

La plupart des sons de la parole peuvent être étudiés à partir d'une fréquence d'échantillonage comprise entre 8 KHz et 20 KHz .

b) Méthodes d'analyse temporelle.

Ces méthodes fournissent des paramètres à partir de traitements effectués uniquement sur la représentation temporelle du signal (BAUDRY-80, GROCHOLEWSKI-84).

De nombreuses mesures sur cette représentation ont été élaborées :

- mesures des pics qui permettent la détection du pitch et de sa période,
- mesures de l'énergie qui permettent de segmenter les parties voisées des parties non voisées, et de séparer les zones de silence des parties non voisées,
- mesures du nombre de passage par zéro pour déterminer si un segment est voisé ou non.

Les fréquences des deux premiers formants étant calculées à partir de la densité des passages par zéro du signal et de sa dérivée.

La fonction d'autocorrélation appliquée au signal vocal permet de détecter sa périodicité et d'estimer la période du fondamental.

On peut également ranger l'analyse par prédiction linéaire dans les méthodes temporelles puisqu'elle ne demande pas de passage direct dans le domaine fréquentiel pour déduire les paramètres caractéristiques du signal.

C'est une des plus puissantes techniques d'analyse pour estimer les paramètres de base de la parole.

Cette méthode s'appuie sur le modèle de production de la parole dans lequel intervient une source d'excitation et la fonction de transfert du conduit vocal que l'on exprime mathématiquement par :

$$s_n = - \sum_{k=1}^{p} a_k \, s_{n-k} + v_n \quad ,$$

où les a_i représentent les paramètres de la fonction de transfert du conduit vocal et où v_n représente l'excitation.

Si on prédit la valeur du signal $\tilde{s}_n$ à l'aide des p valeurs précédentes du signal de façon linéaire on a :

$$\tilde{s}_n = \sum_{k=1}^{p} \alpha_k \, s_{n-k}$$

l'erreur de prédiction étant $e_n = s_n - \tilde{s}_n$.

A partir des valeurs $s_n, \ldots s_{n-k} \ldots$ du signal on essaie par un système d'équations de trouver les α_k qui minimisent l'erreur de prédiction.

On estime alors que les α_k sont égaux aux paramètres a_k que l'on appelle coefficients de prédiction linéaire.

Pour résoudre le système d'équation on utilise les méthodes de covariance ou d'autocorrélation.

C'est à partir des coefficients de prédiction linéaire que l'on extrait les propriétés du signal vocal :

- estimation du spectre (à partir de 12 coefficients sur la figure II-2),
- estimation de la fréquence des formants,
- détection du fondamental.

c) Méthodes d'analyse fréquentielle.

L'analyse fréquentielle de la parole se ramène aux opérations de la transformée de Fourier et n'a d'intérêt que si elle s'applique à une période stable du signal vocal, donc sur une période assez courte.

Le spectre à court terme du signal $f(t)$ se calcule à partir d'une fenêtre $h(t)$ qui permet d'isoler une portion du passé récent de $f(t)$:

$$F(\omega, t) = \int_{-\infty}^{t} f(\tau)\, h(t-\tau)\, e^{-j\omega\tau}\, d\tau \quad .$$

La quantité $|F(\omega, t)|^2$ est le spectre de puissance à court terme.

On peut représenter $f(t)$ par une succession de spectres à court terme où les variations du signal apparaissent bien. (Voir exemple sur la page suivante).

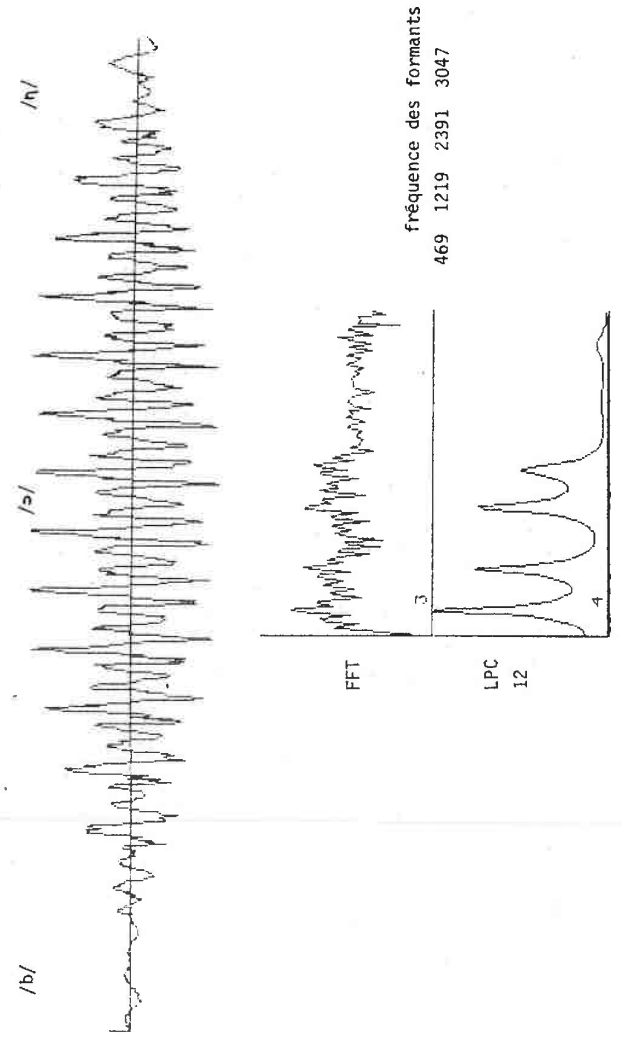
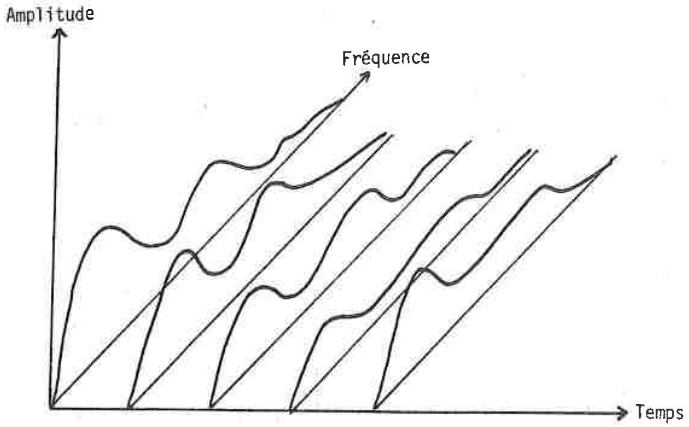


Figure II-2 : LPC à 12 coefficients sur /ɔ/
FFT sur /ɔ/



SPECTRES A COURT TERME

Une des fenêtres donnant de bons résultats est la fenêtre de Hamming.

Deux méthodes pour implémenter l'analyse de Fourier à court terme sont généralement utilisées :

- Les bancs de filtres analogiques ou numériques.

Ils sont composés d'une série de filtres passe-bande. Si les bandes passantes sont soigneusement choisies, l'ensemble des sorties des filtres constitue une bonne approximation du spectre de puissance à court terme du signal (Principe du VOCODER).

Dans notre système nous utilisons l'analyseur d'un VOCODER numérique à canaux, programmable, c'est-à-dire que l'on peut modifier par programme le nombre de filtres et les paramètres définissant chaque filtre.

Actuellement, nous utilisons 16 canaux qui fournissent, toutes les 10 ms, l'énergie du signal dans les différentes bandes de fréquence dont la répartition est décrite sur le tableau II-1.

Caractéristiques des filtres du VOCODER (SLE) utilisé :

N° CANAL	BANDE PASSANTE	LARGEUR DE BANDE
1	350	200
2	550	200
3	750	200
4	950	200
5	1 175	250
6	1 450	300
7	1 750	300
8	2 050	300
9	2 350	300
10	2 650	300
11	2 950	300
12	3 300	400
13	3 700	400
14	4 100	400
15	4 700	800
16	5 900	1 600

Tableau II-1

- Utilisation de la FFT : Transformée de Fourier Rapide

A partir du signal échantillonné, la transformée discrète de Fourier (DFT) permet de définir N valeurs du spectre à l'aide de N échantillons $x_i(n)$ de $f(t)$

$$X_i(\omega) = \sum_{n=0}^{N-1} x_i(n)\, w(n)\, e^{-j\omega n} \qquad (w(n) : \text{fenêtre})$$

La FFT est un algorithme très performant pour évaluer les expressions de la DFT.

Le spectre résultat de la FFT (figure II-2) est modulé par la fréquence du fondamental.

Afin d'extraire de ce spectre des informations plus pertinentes, telles que les formants, le spectre lissé, le fondamental, on peut effectuer une analyse cepstrale.

L'analyse cepstrale s'appuie sur l'hypothèse que la production de la parole peut être considérée comme résultat de la convolution d'une fonction d'excitation avec la réponse impulsionnelle du conduit vocal.

Soit $V(\omega)$ le spectre de la source vocale,

$H(\omega)$ le spectre de la fonction de transfert du conduit vocal,

$S(\omega)$ le spectre du signal,

$$S(\omega) = V(\omega).H(\omega) .$$

Cette méthode permet de séparer les deux composantes de la convolution en passant par une somme des deux composantes.

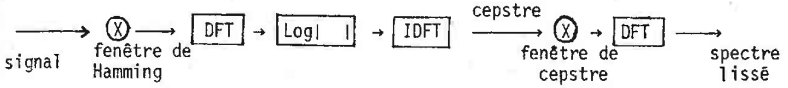
Le calcul du cepstre se fait donc en appliquant la fonction logarithme aux résultats de la transformée de Fourier, après quoi on effectue une transformée de Fourier inverse.

La présence ou l'absence d'un pic important dans le cepstre renseigne sur le caractère voisé ou non voisé du son et de plus, fournit

une bonne indication sur la fréquence du fondamental.

L'enveloppe spectrale du conduit vocal, obtenue par une nouvelle transformée de Fourier, fait nettement apparaître, sur le spectre lissé, les formants.

Les étapes du cepstre :



Sur la figure II-4, on peut distinguer les résultats des différentes étapes :

- FFT,
- Spectre d'excitation,
- Cepstre,
- Spectre lissé par cepstre.

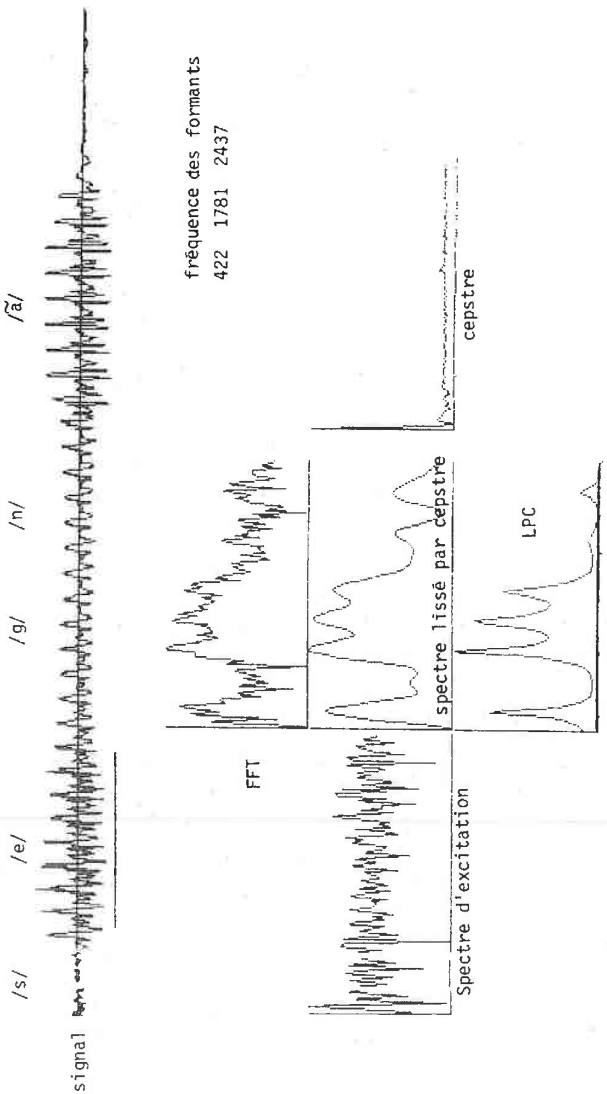
d) L'analyse du signal vis-à-vis de l'adaptation au locuteur.

Une manière d'aborder l'adaptation au locuteur est de faire appel à des paramètres permettant d'appliquer nos connaissances acoustiques, phonétiques ou même articulatoires. Ceci nécessite en général que l'analyse du signal sépare les influences relatives au conduit vocal de celles de la source d'excitation.

Les techniques d'analyse du signal telles que FFT ou VOCODER présentent l'inconvénient de ne pas faire cette séparation.

A l'inverse, les techniques de modélisation telles que la prédiction linéaire ou l'analyse cepstrale réalisent une analyse de portée générale en s'appuyant sur les caractères spécifiques du système de phonation (GUEGEN-76).

Figure II-4 : Les résultats des différentes étapes de l'analyse cepstrale sur /e/



La grande diversité des locuteurs susceptibles d'utiliser un système de reconnaissance nécessite une analyse du signal adaptée à tous les types de voix. Par exemple pour les femmes, il faut une analyse du signal permettant de détecter une énergie dans les très hautes fréquences.

On peut remarquer que les spectres VOCODER sur lesquels nous avons travaillé présentent quelques imperfections dont notamment :

- une mauvaise analyse du signal dans les très hautes fréquences : absence fréquente d'énergie à 5 000 Hz,
- la présence de faible énergie dans les basses fréquences pour les phonèmes sourds,
- le peu de précision dans la répartition de l'énergie dues au nombre restreint de canaux utilisés.

1.2.- Les méthodes de reconnaissance.

Introduction

L'élaboration d'une méthode de reconnaissance pour la parole est déterminée par une suite de choix qui dépendent des objectifs visés et des contraintes techniques imposées.

Ces objectifs peuvent être classés de la sorte :

- reconnaissance de mots isolés - de mots connectés - de parole continue,
- reconnaissance monolocuteur - multilocuteur,
- reconnaissance de petits vocabulaires (250 mots) - grands vocabulaires,
- reconnaissance de vocabulaire de petite complexité - grande complexité (beaucoup de mots monosyllabiques, etc.) (ROSENBERG-82),
- reconnaissance en temps réel.

Les méthodes de reconnaissance peuvent suivre deux approches différentes :

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

Dans l'approche globale (ou acoustique) le message à identifier (le plus souvent le mot) est une forme insécable qu'il s'agit de reconnaître dans un ensemble de formes de références. L'approche globale ne permet d'effectuer de la reconnaisance que de mots isolés et pour des vocabulaires de petites ou moyennes tailles. Cette dernière contrainte étant due d'une part à la taille mémoire relativement limitée face à la mémorisation importante des formes acoustiques de tous les mots du vocabulaire, et d'autre part au respect d'un temps de réponse assez court incompatible avec le temps de comparaison nécessaire pour un grand nombre de mots.

Dans l'approche analytique (ou phonétique) le message est considéré comme une suite d'éléments à reconnaître. Ces éléments peuvent être des phones, phonèmes, diphonèmes, syllabes, etc. . L'approche analytique permet la reconnaissance de mots isolés pour de très grands vocabulaires (500 à 1 000 mots ou plus) car on ne mémorise qu'un nombre restreint d'éléments (50 à 100) indépendants de la taille du vocabulaire. Elle permet également la reconnaissance de la parole continue.

Schéma récapitulatif (HATON-82)

Approche	MOTS ISOLES	PAROLE CONTINUE
globale	Reconnaissance de mots, petits vocabulaires.	Localisation de mots dans des phrases.
analytique	Reconnaissance de mots, grands vocabulaires.	Localisation de mots, reconnaissance et compréhension de phrases.

a) Approche globale.

L'approche globale fait appel aux techniques classiques de reconnaissance de formes ; nous en exposons brièvement quelques unes.

- Les méthodes de corrélation : on calcule un taux de corrélation entre la forme à identifier et les formes prototypes (pattern matching). En ce qui concerne la reconnaissance de mot on pratique toujours une normalisation temporelle entre le mot à identifier et les mots prototypes et ceci souvent à l'aide d'une méthode de comparaison dynamique (DI MARTINO-84). Notons que dans ces systèmes les formes prototypes qui sont une représentation acoustique des mots sont fort dépendants du locuteur ayant effectué l'apprentissage.

- Les méthodes paramétriques de partition de l'espace des formes :

une forme est décrite par N paramètres. L'identification d'une forme se fait en déterminant l'appartenance de cette forme à une classe de formes parmi plusieurs classes, chaque classe pouvant être figurée par un nuage de points dans l'espace des paramètres.

Le but de ces méthodes est de trouver des fonctions discriminantes définissant des hypersurfaces séparatrices des nuages de points.

Les systèmes de reconnaissance de mots qui utilisent ce type de méthode réalisent souvent un regroupement (clustering) qui consiste d'une part à condenser les nuages correspondant à une classe donnée et d'autre part à éloigner les uns des autres les nuages de classes différentes.

b) Approche analytique.

Dans l'approche analytique il y a plusieurs problèmes à résoudre :

1) segmentation du message en unités
2) identification de ces unités
3) reconnaissance du message.

α) Segmentation et identification

En parole, la première difficulté consiste à définir l'unité de segmentation : phone, phonème, diphonème, syllabe. Chacune d'elles présente des avantages et des inconvénients, la plus utilisée étant le phonème.

Le décodage acoustico-phonétique permet de réaliser les phases 1) et 2). Dans certains systèmes ce découpage en deux phases n'est pas toujours distinct.

Le décodage acoustico-phonétique a été abordé de différentes façons, dont on en trouve une description dans (ZUE-81).

On peut retenir essentiellement deux classes de méthodes de décodage acoustico-phonétique qui prennent en compte ou non l'influence de l'environnement des phonèmes sur leurs réalisations acoustiques :

- Les méthodes indépendantes du contexte.

En général elles exploitent les techniques usuelles de reconnaissance des formes :

La segmentation est effectuée à l'aide de paramètres, tels que l'énergie du signal, la variation spectrale, le voisement, la densité de passage par 0 qui, en fonction de leurs variations, permettent de trouver les frontières entre segments.

Pour l'identification on utilise une méthode du type de celles que nous avons exposées pour la reconnaissance globale de mot : analyse discriminante, ou "pattern matching" avec normalisation fréquentielle, que l'on applique à des segments phonétiques. On conserve alors autant de prototypes (formes de référence) qu'il y a de segments différents.

La représentation des prototypes sous forme de spectres, calculés à partir de LPC, FFT ou les sorties d'un VOCODER, rend ces systèmes entièrement dépendants du locuteur ayant réalisé l'apprentissage des prototypes (HATON-81).

Il existe toutefois des méthodes permettant l'identification des unités à partir d'indices et de traits phonétiques et qui ne tiennent pas compte de l'environnement de ces unités.

Le système de décodage acoustico-phonétique, réalisé au CRIN par LAZREK M. dans le cadre de sa thèse de troisième cycle, s'appuie sur un modèle mixte d'identification : identification spectrale,
identification par les indices et les traits.

- Les méthodes dépendantes du contexte :

. *Segmentation* :

L'ensemble des paramètres et des seuils utilisés pour effectuer la segmentation est dépendant d'un premier contexte phonétique facilement détectable. Par exemple dans le système de reconnaissance phonétique asynchrone de MYRTILLE II les paramètres de segmentation sont différents pour les zones voisées, non voisées et les zones transitoires. Signalons l'utilisation d'un ensemble de règles syntaxiques contextuelles qui permet une segmentation relativement indépendante du locuteur dans le système ARIAL II (DOURS-81).

. *Identification* :

Les méthodes dépendantes du contexte nécessitent le plus souvent l'extraction de traits acoustico-phonétiques et ceci en s'appuyant sur un ensemble de connaissances sur la nature acoustique de différentes unités à identifier.

On détermine ensuite l'identité de l'unité en utilisant des techniques statistiques ou bien à l'aide de règles heuristiques ou de règles syntaxiques.

La description et l'identification des unités phonétiques par des traits ont l'avantage de diminuer l'influence des différences acoustiques dues aux locuteurs.

Exemple de traits phonétiques généralement utilisés :
ouvert/fermé ; aigu/grave ; doux/strident ; nasal/oral ;....

Un des problèmes majeurs actuels, comme nous l'avons vu dans le chapitre I, reste l'étude systématique de l'influence de tous les contextes phonétiques possibles.

L'importance de la prise en compte du contexte pour un meilleur décodage des voyelles dans un système de reconnaissance analytique des voyelles fondé sur les indices acoustiques et les traits phonétiques est clairement exposée dans (ROSSI-81).

Dans le même ordre d'idée, le système expert en cours d'élaboration dans notre laboratoire (CARBONELL-84), simulant l'expert phonéticien capable de lire un spectrogramme de parole, fait une large place aux phénomènes phonétiques et phonologiques contextuels explicités sous forme de règles contextuelles.

- Perspectives cognitives.

Une nouvelle approche de prise en compte des problèmes de segmentation-identification automatique de la parole continue est de tenter de reproduire dans la mesure du possible le processus humain. Actuellement, des psycholinguistes (MARLSEN-WILSON-80, LE NY-80) décrivent le fonctionnement humain de compréhension du discours continu de la manière suivante : la compréhension du discours par les individus comporte des traitements multiples conduits en parallèle. Notamment les traitements de segmentation de la chaîne parlée, de l'identification et de désambiguïsation des morphèmes se font en parallèle avec les traitements phonologique, syntaxique et sémantique.

LE NY et DENIS en tirent la conclusion suivante :
"Lorsque l'on étudie les phénomènes de perception d'unités significatives dans le discours, plus précisément l'identification de chaines de constituants de degré inférieur au mot (phonèmes ou syllabes) en tant que

signifiants, ou encore la segmentation du discours continu en de telles unités significatives, on doit accorder une très grande importance aux activités sémantiques préparatoires à chacun de ces actes que constitue l'identification d'une unité".

Notons que cette démarche a été suivie dans le système MYRTILLE II développé au CRIN par J.M. PIERREL (PIERREL-81) uniquement pour des traitements de niveau "supérieur" au décodage acoustico-phonétique.

Dans le domaine cognitif il serait sans doute intéressant de connaître le processus humain d'adaptation au locuteur ; les travaux menés en recherche cognitive devraient nous aider à améliorer nos systèmes de reconnaissance.

β) Reconnaissance du discours continu

L'approche analytique permet seule la reconnaissance du discours continu.

Or, dans l'approche analytique, la chaîne d'unités phonétiques résultat est toujours entachée d'erreurs et de plus cette source d'information est très insuffisante à la reconnaissance du message.

C'est pourquoi les systèmes de reconnaissance, et plus encore de compréhension, font appel à des sources d'informations de différents niveaux que l'on peut répartir ainsi :

- les informations propres à la parole et à la langue :
  - l'acoustique,
  - la phonétique,
  - la phonologie,
  - la prosodie.
- les informations propres au langage de l'application envisagée :
  - le lexique,
  - la syntaxe,
  - la sémantique,
  - la pragmatique.

On trouvera dans (PIERREL-81) un essai de classification de ces sources d'informations et les diverses stratégies d'utilisation de ces informations.

Une façon d'élaborer un système de reconnaissance multilocuteur par approche analytique est de représenter les différentes informations concernant les variations de la parole intra et inter-locuteurs que nous avons exposées dans le chapitre I.

- Ainsi les informations acoustico-phonétiques permettent la paramétrisation du reconnaisseur acoustico-phonétique en fonction du locuteur.

- Le lexique contient une représentation acoustico-phonétique de référence du mot, ainsi que les différentes altérations phonologiques possibles.

- Les informations prosodiques propres au locuteur sont pour le moment inexploitées.

1.3.- L'apprentissage.

L'apprentissage est une phase très importante de la reconnaissance, et de cette phase dépendent en grande partie les résultats du système.

Distinguons plusieurs façons de réaliser l'apprentissage qui sont parfois complémentaires :

- la mémorisation des formes de référence,
- l'apprentissage des paramètres du système de reconnaissance pendant la phase de mise au point,
- l'apprentissage des règles.

La mémorisation des formes de références, mots, phonèmes, etc., doit présenter certaines qualités dont :

. La fiabilité et la robustesse des formes apprises pour tenir compte de la dérive de la voix, des conditions d'acquisition...

. Un nombre de références pas trop élevé pour respecter le temps de réponse du système de reconnaissance.

. Un nombre de locutions pas trop contraignant pour le (ou les) locuteur réalisant l'apprentissage.

. Assurer un bon score de reconnaissance.

On trouvera dans (RABINER-80) une comparaison de différentes méthodes d'apprentissage pour les systèmes de reconnaissance de mots isolés. Entre autres les méthodes de "moyenage" de locutions, ou les méthodes de clustering qui ont l'avantage de fournir des systèmes indépendants du locuteur.

L'apprentissage des paramètres du système est généralement effectué par un opérateur qui au vu des résultats de reconnaissance corrige au fur et à mesure les réglages du système, jusqu'à obtenir un taux de réussite suffisamment élevé.

Ce sera, par exemple, pour l'analyse discriminante l'ajustement des fonctions de décisions et des fonctions discriminantes, ou pour les méthodes heuristiques l'ajustement des seuils de décision et le changement du calcul des paramètres.

A la base de l'apprentissage de l'ensemble des règles d'un système, il y a généralement les connaissances acquises par un observateur, qui sont ensuite complétées pendant la phase de mise au point : création, suppression ou modification des règles. La description des formes par un ensemble de règles possède l'avantage d'être relativement peu dépendante du locuteur. Toutefois certains paramètres du système (indices, attributs des formes élémentaires...) gardent une certaine dépendance vis-à-vis du locuteur.

Dans les systèmes mono-locuteurs, l'opérateur étant souvent le locuteur, on court le risque de voir le locuteur s'adapter au système de reconnaissance jusqu'à compenser certaines lacunes du système.

Pour les systèmes multilocuteurs, on souhaite se dispenser de la présence d'opérateur ; c'est pourquoi une éventuelle phase d'apprentissage des formes de référence ou des paramètres devra être réalisée de façon entièrement automatique.

2.- LE POINT SUR LES SYSTEMES DE RECONNAISSANCE DE LA PAROLE MULTILOCUTEUR.

On peut distinguer deux grandes classes de systèmes multilocuteurs : les systèmes indépendants du locuteur et les systèmes qui s'adaptent à tout nouveau locuteur. Théoriquement ces systèmes fonctionnent pour un nombre illimité d'individus. Rangeons également dans les systèmes multilocuteurs ceux qui peuvent fonctionner pour n locuteurs, ces derniers devant être connus et reconnus par le système. Les différentes approches de réalisation d'un système multilocuteur que nous allons décrire dans la suite sont répertoriées sur la figure II-5.

2.1.- Systèmes n-locuteurs.

Ils nécessitent une première phase d'apprentissage et de mise au point du système pour chaque nouveau locuteur. On associe aux résultats de cette phase (formes de référence, paramètres) un mot-clé propre au locuteur (en général la locution d'un mot identifiant la personne).

Avant toute phase de reconnaissance, on passe par l'identification du locuteur à partir du mot-clé ce qui permet de mettre à jour tous les paramètres du système.

L'inconvénient majeur et évident de ces systèmes est la nécessité de stocker l'ensemble des paramètres qui croît proportionnellement au nombre de locuteurs. Ceci limite donc le nombre de locuteurs et/ou la taille du vocabulaire.

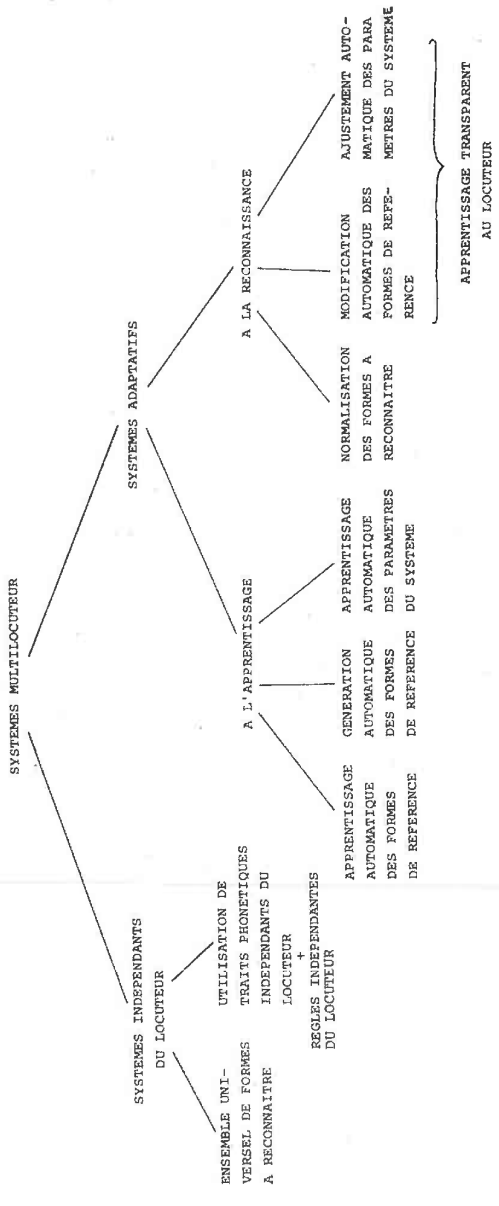


Figure II-5 : LES DIFFERENTES APPROCHES DE REALISATION D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE MULTILOCUTEUR.

La fiabilité de la reconnaissance du mot-clé peut facilement être assurée par des demandes de validation.

Le système PRDY3 de reconnaissance de mots isolés par programmation dynamique mis au point dans notre laboratoire réalise la reconnaissance de chiffres pour un ensemble de locuteurs.

Ces systèmes sont actuellement à la base de la réalisation de certaines applications qui nécessitent une commande vocale telles que le montage de paquet dans les centres de tri ou dans les aéroports.

2.2.- Les systèmes indépendants du locuteur.

La plupart des systèmes indépendants du locuteur sont des systèmes qui, pendant la phase d'apprentissage, tentent de mémoriser tous les représentants possibles des formes à reconnaître (mot, phonème, spectre, etc.) des différents locuteurs afin de former un ensemble universel.

La réduction de cette ensemble de formes, sans diminuer les performances du système a fait l'objet de nombreuses études.

Ainsi il est possible en vue de l'identification des voyelles indépendantes du locuteur de ne mémoriser les formes des différents allophones des voyelles que d'un nombre limité de locuteurs présentant des caractéristiques différentes, en l'occurrence des accents différents (GUPTA-78).

Mais ce choix des références à mémoriser pourra être optimisé, en qualité et en quantité, par des méthodes automatiques de regroupement ("clustering") selon certains critères, ou par des méthodes plus strictes de recherche d'un ensemble de locuteurs à partir d'une typologie des locuteurs.

En général on applique les techniques de "clustering" pour la reconnaissance de mot isolé par approche globale, de petits ou moyens vocabulaires.

On fait prononcer plusieurs fois chaque mot du vocabulaire par un grand nombre de locuteurs (> 20) . Il s'agit ensuite de regrouper ces différentes répliques d'un mot en classes. Plusieurs techniques de classification ont été développées à partir de méthodes classiques d'analyse de données (RABINER-79) et améliorées pour un vocabulaire de 129 mots (WILPON-82).

On constitue l'ensemble des mots de référence en conservant pour chaque classe plusieurs représentants (au moins 5).

Pendant la phase de "clustering", ou la phase de reconnaissance la distance entre deux mots est le plus souvent calculée par un algorithme de programmation dynamique afin de réduire les différences temporelles.

Citons en France le système SYRIL, mis au point au centre de recherche de la C.G.E. (BRIANT-83), fonctionnant sur 35 mots, et le système dont l'étude fait l'objet de la thèse de 3ème cycle de Pascal DIVOUX, au C.R.I.N..

Actuellement, certains travaux (SUGAMURA-83) conduisent à la reconnaissance de mots isolés pour grands vocabulaires généralisés à la reconnaissance indépendante du locuteur.

L'originalité de ces travaux provient du fait que les pseudo-phonèmes de référence, ainsi que la description des mots du dictionnaire (comme suite de pseudo-phonèmes), sont générés automatiquement, à partir de l'ensemble des mots d'apprentissage de 264 locuteurs. (Les pseudo-phonèmes sont déterminés par une technique de clustering sur les différents échantillons extraits de ces mots. Le nombre de représentants d'un mot est limité en ne conservant que les mots ayant le plus grand nombre de proches voisins).

En dernier lieu, citons une approche récente de résolution de ce problème par la recherche d'invariants acoustiques de la parole menée par ROSSI et al. (ROSSI-83) sur les voyelles, et fondée sur les variations d'énergie dans le spectre à partir des données d'un vocodeur à 14 canaux. Cette recherche d'invariants a été également entreprise, et généralisée pour les consonnes, par FONSALE (FONSALE-84) qui définit des traits phonétiques à partir des propriétés du signal indépendantes du locuteur : variation du signal, rapport d'énergie, utilisation de bandes de fréquence assez large.

Actuellement parmi quelques systèmes de reconnaissance de mots isolés indépendants du locuteur, qui sont commercialisés, citons le système de la firme VERBEX qui reconnaît un petit vocabulaire (chiffres plus quelques commandes) par voie téléphonique et qui s'adapte dynamiquement au bruit et à la voix du locuteur pendant la communication. Ce système a été mis au point à partir d'une énorme base de données de mots multi-locuteurs.

2.3.- Les systèmes adaptatifs.

A l'inverse des systèmes indépendants dont la phase d'apprentissage est faite une fois pour toutes, et où la phase de reconnaissance ne varie pas d'un locuteur à l'autre, les systèmes adaptatifs sont des systèmes qui tentent d'améliorer leurs résultats en fonction du nouveau locuteur qui se présente.

L'adaptation du système peut se faire de diverses façons, qui dépendent notamment des techniques de reconnaissance mises en oeuvre dans le système, et de la paramétrisation du signal vocal.

Nous proposons une classification des systèmes adaptatifs en fonction du moment où s'effectue l'adaptation :
à l'apprentissage ou pendant la reconnaissance, les deux n'étant bien sûr pas exclusifs. (voir figure II-5)

a) Adaptation à l'apprentissage.

L'adaptation automatique pendant la phase d'apprentissage ne présente de difficulté et d'intérêt que pour les systèmes s'appuyant sur une approche analytique. Dans un premier temps cette adaptation nécessite souvent une phase d'acquisition automatique de toutes les formes de référence du nouveau locuteur. En outre, ceci permet de réaliser une certaine éducation du locuteur face au système.

α) Apprentissage automatique des formes de références.

L'apprentissage automatique des formes de références présente deux intérêts :

- Il facilite la phase d'apprentissage d'un nouveau locuteur en évitant un étiquetage manuel (donc peut-être la présence d'une personne qualifiée).
- Il facilite l'acquisition des formes (le plus souvent des phonèmes) d'une grande quantité de parole naturelle nécessaire à des études statistiques, à grande échelle, de la variation de la parole inter et intra-locuteurs; Ces études fournissant la connaissance de base nécessaire pour effectuer la reconnaissance de la parole continue multi-locuteur.

Cet apprentissage automatique qui consiste le plus souvent à segmenter et à étiqueter automatiquement les phonèmes d'une phrase, connue à l'avance, est toujours réalisé par l'alignement des segments de la description standard de la phrase avec ceux de la phrase énoncée à l'aide d'algorithmes de cadrage.

Il existe différentes méthodes, dont on peut tenter de faire une classification selon la description standard de la phrase à cadrer et selon l'algorithme de cadrage utilisé.

- De nombreuses méthodes effectuent leur cadrage sur une transcription phonétique sous forme de chaîne de symboles qui représentent soit des phonèmes (LECORRE-79, NEEL-83) soit des segments définissant des classes de phonèmes (WAGNER-81). Dans notre système, la transcription phonétique "idéale" est décrite par une chaîne de traits (acoustico-phonétiques) définissant également des classes de phonèmes.

A chaque symbole est, bien sûr, associée une représentation interne sous forme d'un ensemble de paramètres acoustiques ou phonétiques.

Ces chaînes, fixées a priori, peuvent prendre en considération certaines altérations phonologiques et quelques effets de coarticulation comme les élisions, insertions, liaisons entre mots... ou bien encore les altérations dues aux conditions d'acquisition comme nous l'avons prévu dans notre système.

Toutefois, elle ne reflète pas la réalisation acoustique exacte sur laquelle on souhaite réaliser le cadrage du texte prononcé.

C'est pourquoi une nouvelle approche pour l'apprentissage automatique voit actuellement le jour, qui consiste à cadrer le signal de la parole sur une version synthétisée du texte.

Cette version étant générée à partir d'une transcription donnée en utilisant la synthèse par règles (BRIDLE-83, LENNIG-83).

L'avantage de cette représentation est l'existence d'une forte relation entre la chaîne de symboles et la réalisation acoustique de celle-ci.

Pour améliorer encore la ressemblance des deux phrases à cadrer il serait intéressant de prendre en compte dans la synthèse certaines caractéristiques acoustiques du locuteur.

- Parmi les algorithmes de cadrage on trouve ceux qui nécessitent la construction d'une matrice de comparaison ou de coïncidence, à l'aide d'une mesure de ressemblance entre des phonèmes de références et les phonèmes résultats d'un module de reconnaissance de phonèmes. L'algorithme consiste alors à rechercher le meilleur chemin dans cette matrice en

suivant certaines stratégies : favoriser la diagonale, en tenant compte de certaines altérations (LECORRE-79), effectuer une recherche en faisceaux pour conserver en parallèle les trois meilleurs chemins (NEEL-83).

Le problème principal de ces systèmes reste la constitution du dictionnaire de phonèmes de référence initial : on est confronté au problème "de la poule et de l'œuf", l'algorithme de cadrage nécessitant l'utilisation de phonèmes de références de qualité.

Une méthode qui permet d'éviter ce problème consiste à effectuer le cadrage sur des segments obtenus à partir d'indices acoustico-phonétiques indépendants du locuteur. Ainsi WAGNER (WAGNER-81) compare les résultats d'une segmentation grossière que l'on peut considérer comme indépendante du locuteur (fondée sur des paramètres tels que le voisement, l'énergie), avec la chaîne de la transcription phonétique représentant des classes de phonèmes (voyelle, fricative voisée, etc.). Pour ce faire il applique un algorithme de programmation dynamique (utilisant une distance très simple qui reflète certaines règles acoustico-phonétiques). Un second algorithme de programmation dynamique permet un étiquetage plus détaillé : la comparaison, étant calculée sur la base de dérivées de fonctions (d'énergie, formantique), rend cette étiquetage relativement indépendant du locuteur.

Nous avons choisi de rechercher les segments de la phrase énoncée à l'aide de paramètres indépendants du locuteur (dérivées de la fonction d'énergie, voisement, durée minimale des segments, rapport d'énergies). Le cadrage s'effectue alors de gauche à droite par comparaison des segments, au moyen d'une matrice de confusion et en parcourant en parallèle la réalisation acoustique de la phrase énoncée. L'algorithme permet des retours arrières ainsi que la modification dynamique de certaines paramètres de la procédure de cadrage.

Notons que tous ces systèmes ont naturellement, au niveau de la segmentation-identification, des performances inférieures aux systèmes de décodage acoustico-phonétique mono-locuteur, ceci étant dû à la qualité médiocre des phonèmes de références, ou aux indices indépendants du locuteur insuffisants pour effectuer une très bonne identification. Toutefois ce handicap doit être rattrapé par la connaissance exacte du décodage "idéal" de la phrase prononcée.

Enfin, les systèmes qui utilisent la représentation synthétisée du texte ne sont pas confrontés aux problèmes de segmentation et d'identification de phonèmes, mais à des problèmes d'alignements temporels des échantillons des 2 phrases à cadrer ; ce qui en général est résolu par des algorithmes de programmation dynamique (BRIDLE-83, LENNIG-83).

Comme l'emplacement et l'identité des segments sont connus dans la phrase synthétisée, il suffit alors de transférer cette identité sur les échantillons alignés de la phrase naturelle.

β) Génération automatique des formes de références.

La génération automatique des formes de références a pour but de réduire au minimum la phrase d'apprentissage, c'est-à-dire de faire prononcer au nouveau locuteur un sous-ensemble de mots ou phonèmes aussi restreint que possible.

Généralement, des études statistiques préalables permettent de générer par estimation l'ensemble complet des formes propres au locuteur.

C'est ainsi que S. FURUI (FURUI-80) estime, pour chaque nouveau locuteur, tous ses phonèmes de références à l'aide d'un ensemble de règles de transformation appliquées sur quelques phonèmes d'apprentissage propres au locuteur. Ces règles de transformation sont obtenues par une analyse statistique (méthode d'analyse en régression multiple) d'un grand ensemble de phonèmes multilocuteurs, qui permet d'établir des relations indépendantes des locuteurs entre les différents phonèmes.

Dans le même but, Y. GRENIER (GRENIER-80) utilise l'information contenue dans une seule phrase. Une application optimale, d'une phrase standard sur la phrase propre au locuteur, est recherchée comme composition de projections, calculées par l'analyse canonique des corrélations.

Cette même application permet de transformer les références standards en références adaptées au nouveau locuteur.

γ) Ajustement du système.

En dernier lieu, citons les adaptations qui ont pour but d'optimiser, selon les systèmes, soit les fonctions de décision, soit les fonctions de séparation de données propres au locuteur obtenues automatiquement pendant la phase d'apprentissage.

Ainsi, dans le système de reconnaissance phonétique de MYRTILLE II, la segmentation est adaptée au locuteur lors de l'énonciation d'une première phrase sur laquelle on calcule les moyennes et écart-types (considérés comme dépendants du locuteur) des différents paramètres qui interviennent dans la définition des fonctions de segmentation.

Dans le système KEAL (GERARD-81) l'adaptation est réalisée en déterminant, à partir de l'ensemble des phones d'apprentissage du locuteur, les coefficients optimaux des fonctions linéaires de séparation de cet ensemble par un algorithme d'approximation stochastique.

b) Adaptation pendant la reconnaissance.

α) Modification des formes à reconnaître.

- Normalisation des fréquences formantiques.

Il est admis, à présent, que les deux formants F_1, F_2 seulement, sont nécessaires pour identifier correctement les différentes voyelles stables d'un locuteur (PETERSON-55). C'est pourquoi les recherches de normalisation de formants ont principalement porté sur F_1, F_2 . On trouvera répertoriés dans (BOE-80.b) un certain nombre de travaux entrepris dans le domaine, accompagnés d'évaluations de certaines procédures de normalisation.

Parmi de nombreux travaux citons GERSTMAN (GERSTMAN-68), qui à partir des données de PETERSON et BARNEY, réduit à la même échelle les fréquences de F_1 et les fréquences de F_2 de chaque locuteur en fonction de F_1 min et F_1 max (respectivement F_2 min et F_2 max) de toutes les voyelles du locuteur. Seulement 3 voyelles sont nécessaires pour trouver ces fréquences minimales et maximales. WEINSTEIN et al. (WEINSTEIN-75) dans le cadre du projet ARPA se sont inspirés des travaux de GERSTMAN pour la normalisation des formants de voyelles stable en introduisant de nouveaux facteurs de normalisation, différents pour les hommes et les femmes.

Contrairement aux exemples cités ci-dessus, H. WAKITA propose une normalisation ne faisant pas intervenir de reconnaissance a priori sur le locuteur (WAKITA-77). La normalisation est effectuée à partir de la longueur du conduit vocal du nouveau locuteur et de la fonction d'aire, lesquelles sont estimées directement sur l'onde acoustique de la parole. En faisant l'hypothèse que pour une même voyelle la forme du conduit vocal ne change pratiquement pas d'un locuteur à l'autre mais que seule la longueur du fait de différences anatomiques varie, les fréquences des formants normalisés sont définies comme les fréquences de résonnance d'un tube acoustique présentant les formes du conduit vocal du locuteur, normalisées par la longueur du conduit vocal.

Citons encore LOBANOV (LOBANOV-71) qui normalise par rapport à la valeur moyenne et l'écart-type de chaque formant pour l'ensemble des voyelles, et NEAREY (NEAREY-77) qui propose une relation permettant de passer des formants d'une voyelle d'un locuteur, à ceux de la voyelle d'un autre locuteur.

- Normalisation fréquentielle.

MATSUMOTO et al. (MATSUMOTO-79) ont proposé une normalisation spectrale par décalage fréquentiel du spectre de la voyelle à identifier, pendant la comparaison aux spectres des voyelles de référence. Le décalage fréquentiel optimal, fournissant la distance minimale entre deux spectres, est obtenu par une technique de programmation dynamique qui élimine des différences linéaires dues à des différences de longueur de conduit vocal et quelques différences non linéaires entre hommes et femmes.

En vue d'adapter les spectres des voyelles d'un nouveau locuteur aux spectres des voyelles de référence du système, JASCHUL J. (JASCHUL-79) calcule dans un premier temps les paramètres de différentes fonctions de transformation et ceci à partir de l'ensemble des voyelles propres au locuteur. Il détermine entre autres une fonction de pondération spectrale pour éliminer les différences d'amplitude spectrale et différents types de fonctions d'alignement fréquentiel : par décalage des régions formantiques ou par comparaison dynamique.

Cette approche, dont l'inconvénient majeur est de nécessiter un apprentissage complet des voyelles pour tout nouveau locuteur, a tout de même permis d'évaluer un bon nombre de fonctions de transformation pour l'adaptation au locuteur.

β) Ajustement des paramètres du système

Presque tous les algorithmes de classification utilisent un ensemble de seuils dans les processus de décision. L'utilisation d'un ensemble de seuils fixes conduit à un grand nombre de problèmes par le fait que beaucoup de seuils sont dépendants du locuteur et du temps (dérive de la voix).

Pour éliminer cette difficulté, SAMBUR et al. (SAMBUR-75) ont établi une procédure d'auto-normalisation des paramètres dans laquelle certains seuils sont obtenus à partir de mesures de ces paramètres (moyenne, écart-type, etc.) faites directement sur l'échantillon de parole à reconnaître. Cette technique a été appliquée à la reconnaissance de chiffres américains, en utilisant un ensemble de paramètres simples et robustes pour classer les différents phonèmes dans six grandes catégories.

γ) Modification des formes de références

Citons, en dernier lieu, dans le projet ARPA, le système de reconnaissance de la parole HARPY (LOWERRE-77) (avec un vocabulaire de 1000 mots) qui apprend dynamiquement les paramètres dépendants du locuteur pendant que ce dernier utilise le système.

Pour ce faire, un nouveau locuteur débute avec un ensemble de formes de références (phones), chacune étant la moyenne de formes de nombreux locuteurs.

A la reconnaissance, chaque phone correctement reconnu est moyenné avec la forme de référence correspondante pour générer une nouvelle forme de référence qui se "rapproche" donc du phone propre au locuteur.

Le "correctement reconnu" nécessite malheureusement que le système connaisse à l'avance la phrase qu'il a à reconnaître.

CONCLUSION.

L'approche des systèmes indépendants du locuteur par construction d'un ensemble de formes de références "universel" semble avoir trouvé ses limites dans les petits vocabulaires. Par contre la recherche de traits acoustico-phonétiques indépendants du locuteur ainsi que l'utilisation de règles générales valables pour un grand nombre de locuteurs permettent la réalisation de systèmes de reconnaissance pour de grands vocabulaires ou pour la parole continue.

Jusqu'à présent, dans les systèmes adaptatifs, les efforts ont essentiellement porté sur les problèmes de variabilité acoustique (normalisation fréquentielle) phonétique (normalisation formantique, utilisation de phones tenant compte de l'influence du contexte) et prosodique (normalisation d'amplitude et normalisation temporelle).

Toutefois les variations phonologiques des mots, en fonction de l'accent ou en fonction du contexte, ont déjà été à l'étude dans le système HARPY qui permet d'encoder la variabilité dialectal dans son lexique et de prendre en compte les variations contextuelles des mots par des règles de jonctions de mots.

En conclusion, nous pensons qu'une bonne approche de réalisation d'un système multilocuteur, passe par la connaissance a priori de la variation des données traitées par le système de reconnaissance.

C'est pourquoi nous avons voulu aborder ce problème par l'examen d'un grand nombre de données multilocuteurs obtenues facilement par le programme d'apprentissage automatique que nous exposons dans le chapitre III.

CHAPITRE III

APPRENTISSAGE AUTOMATIQUE DES FORMES DE REFERENCES DE PHONEMES

INTRODUCTION

Dans ce chapitre nous présentons le système d'apprentissage automatique de phonèmes que nous avons réalisé. Ce système, qui peut fonctionner pour un grand nombre d'individus, fournit rapidement, en minimisant les contraintes pour le locuteur, les formes de références utilisées par le système de reconnaissance de mots isolés centi-seconde de notre laboratoire (voir figure III-1).

Nous vérifions l'efficacité des phonèmes ainsi appris en comparant les performances du décodage acoustico-phonétique centi-seconde selon les formes de références utilisées.

1.- LE SYSTEME DE RECONNAISSANCE DE MOTS ISOLES CENTI-SECONDE

1.1.- Description.

Ce système (PERRIN-81) permet la reconnaissance en ligne d'un vocabulaire de 400 mots isolés, appuyé sur une méthode de reconnaissance analytique.

Il se divise en 2 grandes parties :

- un module acoustico-phonétique dont les résultats sont sous la forme d'une chaîne de phonèmes à réponse multiple (1 à 3 phonèmes par réponse),
- un module de reconnaissance de mot qui à partir de la chaîne de phonèmes et de la transcription des mots contenus dans un lexique réalise l'identification du mot.

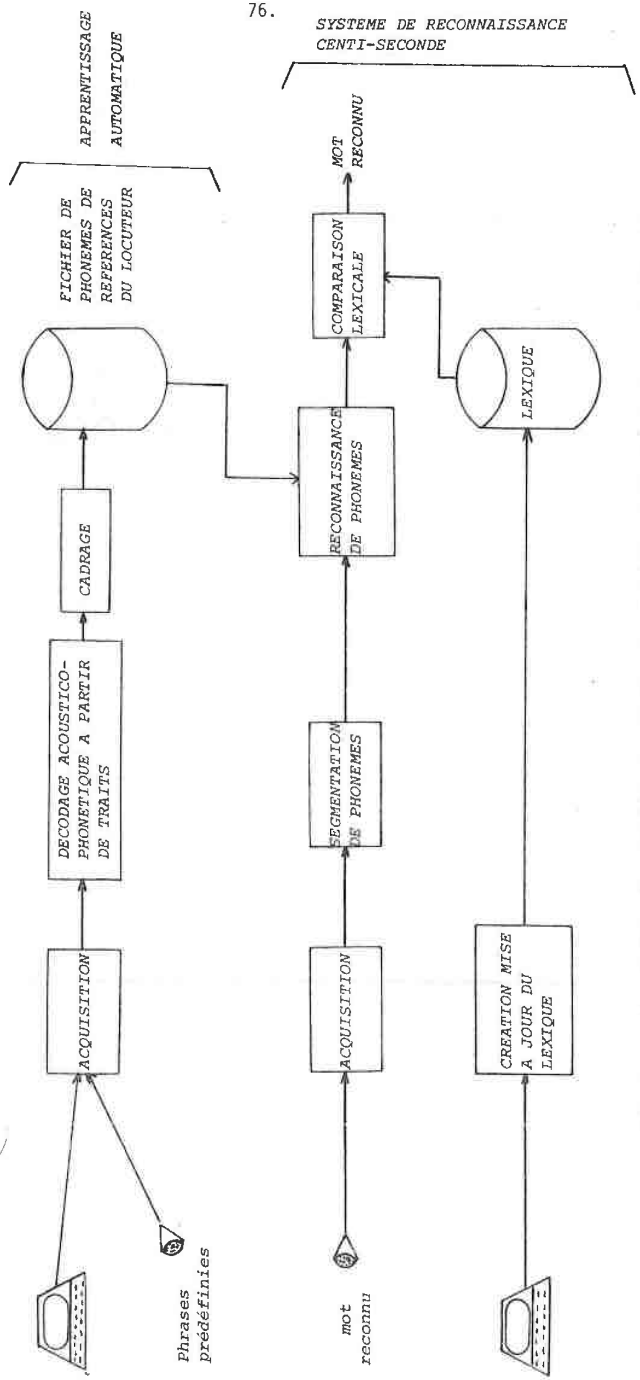


Figure III-1 : Utilisation de l'apprentissage automatique de phonèmes pour le système de reconnaissance de mots centi-seconde.

a) Le matériel utilisé.

L'ensemble des programmes de ce système fonctionne sur mini-calculateur MITRA 125 (mode 15) avec une mémoire centrale de 32 K mots de 16 bits. L'organe de prétraitement du signal vocal est un vocoder à filtres numériques SLE. Le signal vocal, capté au microphone est transmis au vocoder qui délivre tout les centièmes de seconde les valeurs de l'intensité du signal dans les 16 bandes de fréquence s'étendant de 250 à 6500 Hz. La description des 16 filtres utilisés ainsi que quelques caractéristiques de spectres résultats se trouvent dans le chapitre II.

Aux données du vocoder s'ajoute à chaque instant d'échantillonage la valeur de la fréquence du fondamental fournie par un détecteur de mélodie "Mélographe" du CNET. L'obtention en temps réel de ces données est réalisée par un programme qui calcule dans le même temps :

- l'énergie du signal, considérée comme la somme des énergies relevées sur chaque bande de fréquence.
- cinq autres valeurs représentant la répartition de l'énergie dans les bandes de fréquence.

Ainsi, le $i^{ième}$ prélèvement d'un mot, correspondant à 10 ms de parole, est représenté par :

- $E_i(1\text{-}21)$:
  - de E(1) à E(16) : énergies dans les 16 canaux
  - de E(17) à E(21) : fonctions de répartition de l'énergie dans les canaux : GA, FO, BD, DS, CE (LAZREK-83)

avec :

$$GA \text{ (grave/aigu)} = E(1) - \sum_{j=13}^{16} E(j)$$

$$FO \text{ (fermé/ouvert)} = E(3) - E(1)$$

BD (bémolisé/diésé) = E(10) + E(11) - E(5)

DS (doux/strident) = E(15) - E(9) - E(10)

CE (compact/écarté) = E(1) + E(2) + E(6) + E(7) + E(8) + E(9) - 2E(4)

- $MELO_i$: la valeur du pitch

La figure III-2 illustre la représentation d'une phrase à l'aide d'un spectrogramme résultat.

b) Le décodage acoustico-phonétique centi-seconde.

L'algorithme de transcription phonétique comporte deux phases :

- Pendant la première phase, pour chaque prélèvement de parole on effectue :

. La détermination du type de phonème auquel appartient le prélèvement, en fonction du voisement (son voisé ou non voisé),

. La recherche des trois phonèmes de référence les plus proches du prélèvement parmi les phonèmes de même type que le prélèvement. La distance entre le prélèvement i du mot et le phonème k de référence est calculée par :

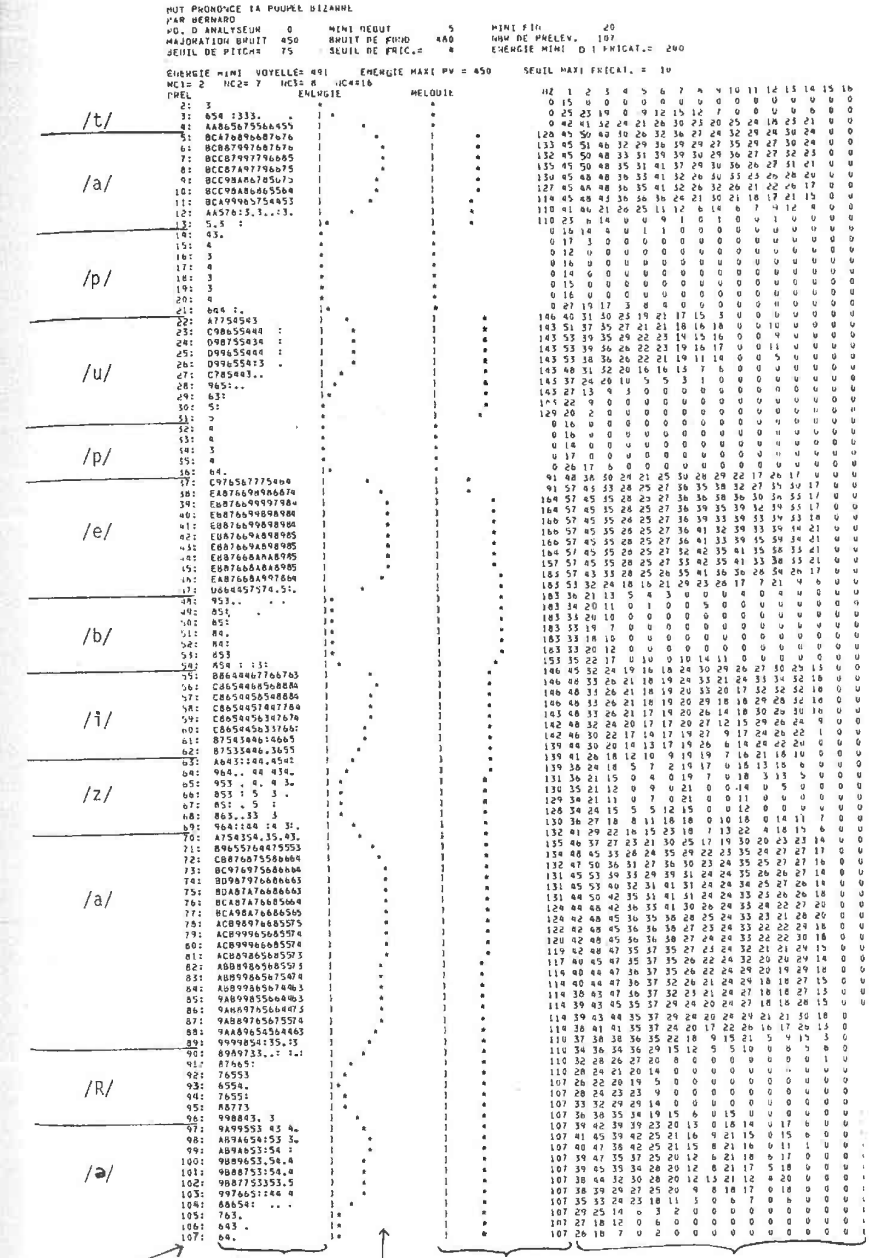
$$\text{distance}(\text{prél}_i, \text{réf}_k) = \sum_{j=1}^{21} |E_i(j) - E_k(j)| \quad .$$

Les phonèmes de références sont acquis au cours d'une phase d'apprentissage manuel que nous exposons plus loin.

. Le calcul de la variation spectrale entre ce prélèvement et le prélèvement précédent :

$$\text{vspe}(\text{prél}_i) = \sum_{j=1}^{21} |E_i(j) - E_{i-1}(j)|$$

Figure III.2. : Représentation de la phrase "TA POUPEE BIZARRE"

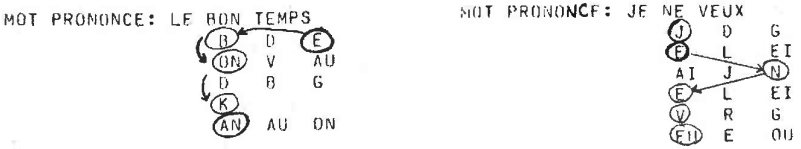


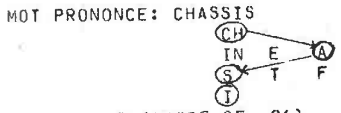
- La deuxième phase est celle du découpage en segments de la suite de prélèvements en fonction des types et de la variation spectrale , suivi de l'identification des segments par un processus de décision majoritaire parmi les 3 phonèmes retenus à chaque centi-seconde.

La chaîne phonémique résultat :

Chaque élément de la chaîne phonémique, représentant le mot prononcé, contient les 3 phonèmes les plus probables parmi 35 phonèmes possibles qui sont répertoriés dans le tableau I-1.

Exemples de chaînes phonémiques et de chemins résultats





J'AI RECONNU "CHASSIS " (SCORE DE 96)

c) Le module de reconnaissance de mots

Le lexique contient la transcription phonétique des mots et permet de tenir compte des erreurs de décodage phonétique telles que confusion, insertion, omission, ainsi que des altérations phonologiques dues au locuteur telles que substitution, insertion, élision. Notons que la représentation des mots du lexique ne prévoit qu'un seul phonème de substitution et qu'un seul phonème d'insertion pour chaque phonème de la transcription. Ceci paraît insuffisant dans le cas de l'utilisation du système généralisée à de nombreux locuteurs, donc à des possibilités de substitutions ou insertions multiples.

L'algorithme de reconnaissance recherche dans la chaîne phonémique un chemin qui décrit la représentation d'un des mots du lexique. Le mot dont le chemin fournit un taux de ressemblance suffisamment élevé et supérieur aux autres est considéré comme le mot reconnu. Voir les exemples de chemins sur la page précédente.

1.2.- Performances du système.

a) Au niveau de la reconnaissance de phonèmes.

Les transcriptions phonétiques obtenues sont de qualité moyenne tant sur le plan de la segmentation (sur-segmentation, sous-segmentation) que sur le plan de l'identification. Ceci est en partie dû à l'ensemble volontairement limité des phonèmes de références (au maximum 40) qui ne peut pas représenter les phonèmes dans différents contextes et à la représentation imparfaite des "spectres" du Vocoder.

Une autre raison de l'insuffisance de la méthode de reconnaissance centi-seconde est qu'elle ne peut pas identifier les sons transitoires et qu'elle ne tient pas compte de l'environnement des phonèmes pendant leur identification. Notons que 65 % des voyelles à reconnaître apparaissent dans le choix proposé au module de reconnaissance de mots. Les consonnes sont globalement bien reconnues à l'intérieur de classes telles que : (B, D, G), (P, T, K), (F, S, CH), (V, Z, J), (N, M) ce qui est assez cohérent.

b) Au niveau de la reconnaissance de mots.

Comme on l'a vu précédemment, la recherche lexicale permet de "rattraper" les cas d'erreurs prévisibles du décodage phonétique, à l'aide d'informations contenues dans le lexique.

Mais l'algorithme de reconnaissance est également adapté à certaines erreurs non prévues dans le lexique. En moyenne, la reconnaissance de mots atteint 65 % de réussite pour un vocabulaire de 400 mots et de complexité importante. Ces performances augmentent largement lorsque la qualité du décodage phonétique s'améliore (en particulier avec les travaux actuellement en cours dans l'équipe).

c) Dépendance vis-à-vis du locuteur.

Le système de reconnaissance de mot centi-seconde présente certaines caractéristiques permettant l'utilisation multilocuteur. En effet, ce système a l'avantage de supprimer la prise en compte des variations temporelles intra et inter-locuteurs, ainsi que les effets de coarticulation inter-phonèmes propres au locuteur.

Le lexique, par une description multiple des mots, permet la prise en compte des variations phonétiques et phonologiques intra et inter-locuteurs.

On peut estimer que les différents paramètres du système, comme les seuils de segmentation (voisement, variation spectrale maximale) ou d'identification (taille minimum d'un phonème) sont quasiment indépendants du locuteur.

En contre partie, la représentation acoustique des phonèmes est naturellement totalement dépendante du locuteur. Et pour un locuteur n'ayant pas effectué l'apprentissage des phonèmes de référence, il est possible de voir chuter les scores de reconnaissance de moitié.

L'adaptation au locuteur consiste donc à fournir automatiquement au système des formes de référence propres au locuteur ou proches de celle du locuteur.

1.3.- Apprentissage manuel des formes de référence des phonèmes.

Chaque locuteur désirant utiliser le système doit posséder un fichier contenant les formes de références des phonèmes. Le fichier prévu peut contenir 39 phonèmes. Un programme intéractif permet au locuteur de réaliser l'apprentissage des phonèmes de la façon suivante :

Pour chaque phonème, le locuteur prononce un mot contenant ce phonème. Le spectrogramme du mot prononcé est affiché sur l'écran de la console alphanumérique. Le locuteur désigne la partie du spectrogramme qui représente le phonème, si possible dans une zone de stabilité. Ceci suppose une certaine habitude de la lecture de spectrogrammes.

Le programme calcule alors le prélèvement moyen en effectuant la moyenne des énergies sur chaque canal pour les prélèvements de la partie désignée :

Soit d le prélèvement de début de la zone du phonème k
Soit f le prélèvement de fin de la zone du phonème k .
La forme de référence du phonème k, E_k est donnée par :

pour c de 1 à 21

$$E_k(c) = \sum_{i=d}^{f} E_i(c) \; / \; (f-d+1) \; .$$

De plus, on associe au phonème son type (voisé, non voisé). Le locuteur peut avoir dans son fichier plusieurs formes de référence pour un phonème et, inversement, n'en avoir aucune. C'est le cas en général des diphtongues /w/, /j/ et /ɥ/.

Il faut noter que cet apprentissage est souvent long, de 15 à 20 minutes, et fastidieux et qu'il nécessite que le locuteur pratique la lecture de spectrogramme ou dans le cas contraire, qu'un opérateur soit présent pendant cette phase.

2.- APPRENTISSAGE AUTOMATIQUE

2.1.- Description et principes de base

Cet apprentissage consiste à extraire, de façon entièrement automatique, les phonèmes de références d'un locuteur, de quelques phrases prédéfinies. L'extraction des phonèmes d'une phrase est réalisée par le cadrage des phonèmes résultats d'un module de décodage phonétique, sur une transcription phonétique standard de la phrase.

Il est important de noter que les résultats du cadrage doivent être corrects à 100 % puisqu'aucune intervention manuelle ne rattrape d'éventuelles erreurs. C'est pourquoi de nombreuses vérifications-validations automatiques sont entreprises durant cette phase.

D'autre part le décodage phonétique doit être valable pour un nombre maximum d'individus, ce qui nous a conduit à rejeter le décodage centi-seconde trop dépendant du locuteur. De plus, le décodage centi-seconde présente des caractéristiques de segmentation trop variables (tantôt bonne segmentation, tantôt sous-segmentation, tantôt sur-segmentation) peu favorables au cadrage.

Nous avons retenu le système de décodage acoustico-phonétique par identification de traits phonétiques, élaboré par J.F. MARI dans le cadre de la reconnaissance de mots isolés pour grands vocabulaires (1000 mots et plus) (MARI-84). Ce système fournit une identification de phonèmes plus grossière (classe de phonèmes) mais l'analyse à partir de traits phonétiques constitue un degré d'abstraction qui permet d'éliminer certaines variations inter-locuteurs. De plus la segmentation est stable et ses paramètres sont facilement adaptables au locuteur.

Toute cette phase d'apprentissage est réalisée de manière conversationnelle, et est inclue dans le système de reconnaissance centi-seconde. Si pour une phrase, le cadrage n'aboutit pas, le locuteur a la possibilité d'énoncer à nouveau cette phrase ou de passer à la phrase suivante.

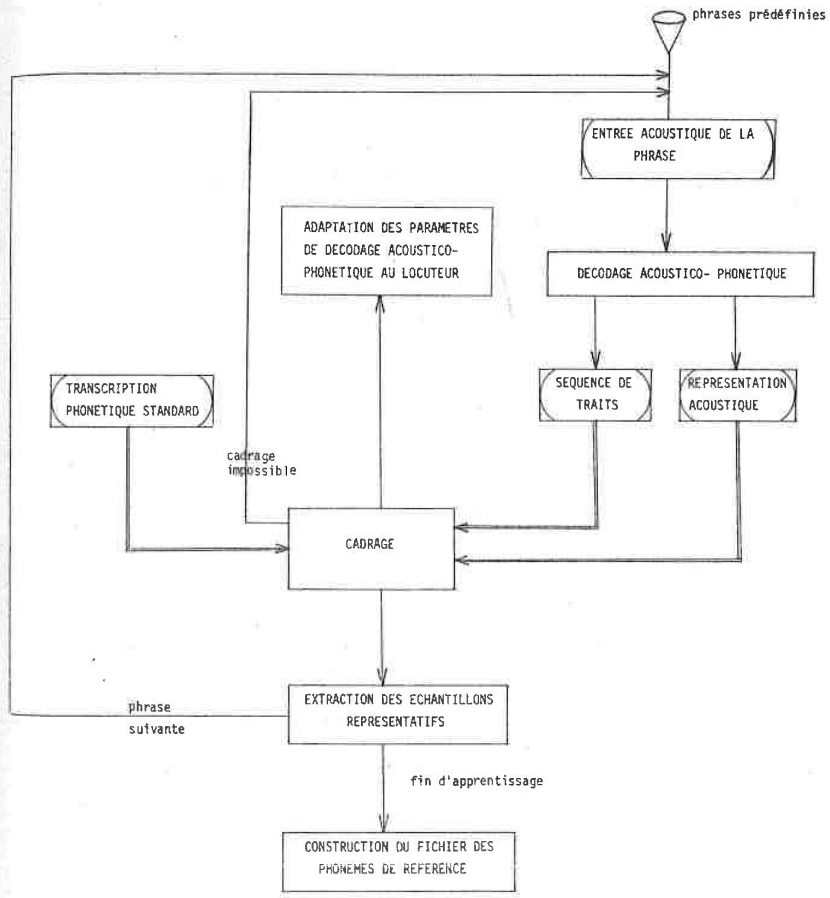


Figure III-3 : Principe de l'apprentissage automatique

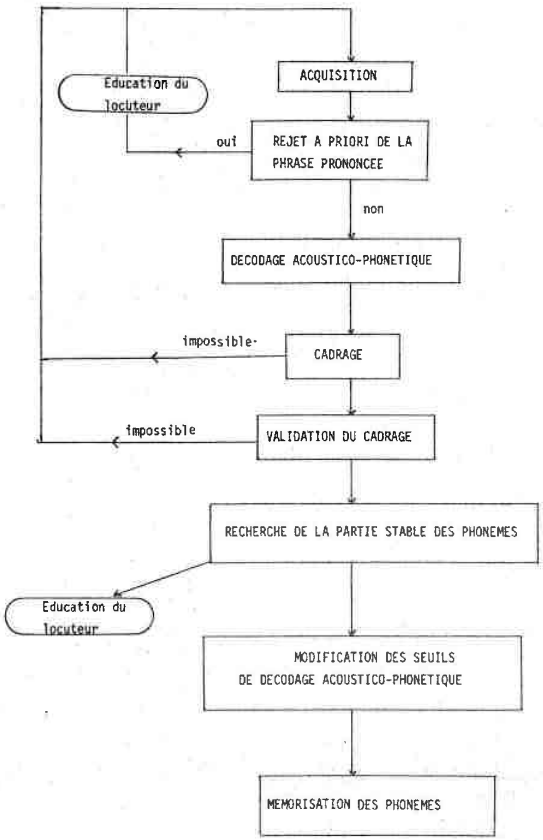


Figure III-4 : Traitements pour une phrase

Une certaine éducation du locuteur est réalisée par l'apport de quelques informations issues des résultats du cadrage. La figure III.3 représente le principe de l'apprentissage automatique. Sur la figure III.4 on présente les différents traitements effectués sur une phrase.

2.2.- Décodage acoustico-phonétique par identification de traits phonétiques.

Ce système de décodage repose sur l'utilisation d'indices acoustico-phonétiques qui permettent la détection de traits phonétiques. Le résultat de ce module n'est pas une suite ou un treillis de phonèmes, mais une séquence de traits phonétiques ; chaque trait représentant une classe de phonèmes. (Dans la suite il y a assimilation entre ces 2 notions).

a) Nature des indices.

Les indices utilisés pour la recherche des traits sont :

- Voisé - Non voisé :

Ces indices sont obtenus en testant la valeur de la fréquence du fondamental en sortie du Mélographe.

Un prélèvement i est voisé si $MELO_i > 0$

- Fricatif - Non fricatif :

Ces indices sont obtenus en testant la valeur du taux de friction TF calculé par le rapport des énergies en basses fréquences (450-1900 Hz) sur les énergies en hautes fréquences (1900-6500 Hz)

$$TF_i \text{ du prélèvement } i = \frac{\sum_{j=2}^{7} E_i(j)}{\sum_{j=0}^{16} E_i(j)}$$

Un prélèvement i est fricatif si $TF_i < FRICA_{(seuil)}$

- Energie du signal vocal :

l'énergie du signal vocal permet de distinguer 4 types de prélèvements :

- Très forte énergie,
- Forte énergie,
- Faible énergie,
- Très faible énergie.

L'énergie d'un prélèvement i est calculée par la somme des énergies dans les 16 canaux.

$$EN_i = \sum_{j=1}^{16} E_i(j)$$

Un prélèvement i est de :

très forte énergie si	EN_i > IEVYMA
forte énergie si	IEVY < EN_i < IEVYMA
faible énergie si	JSFRIC < EN_i < MAXEPV
très faible énergie si	EN_i < JSFRIC

où IEVYMA, IEVY, MAXEPV et JSFRIC représentent des seuils.

b) Nature des traits

les traits reconnus sont :	La classe de phonèmes associée :
PS : Plosive sourde	/p/, /t/, /k/
PV : Plosive voisée	/b/, /d/, /g/
FS : Fricative sourde	/f/, /s/, /ʃ/
FV : Fricative voisée	/v/, /z/, /ʒ/
VY : Voyelle	Les voyelles
VO : Autres voisées	/l/, /m/, /n/, /R/ .

Notons que cette répartition théorique des phonèmes dans les classes ne correspond pas dans certains cas à la réalité expérimentale.

Ainsi /v/ pourrait appartenir à la classe des Plosives et /i/ dans certains contextes à la classe des Autres voisées.

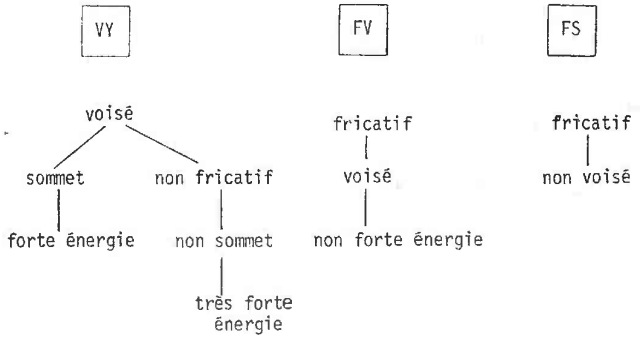
c) Recherche de traits

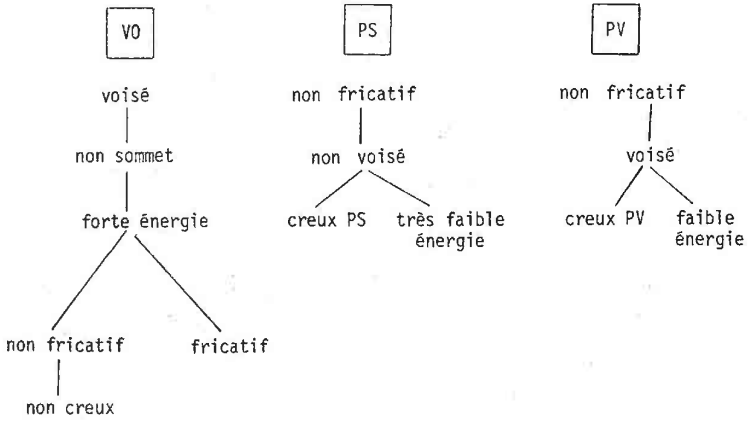
La recherche des segments, identifiés par les traits phonétiques précédemment exposés, est réalisée par l'étiquetage des prélèvements centi-secondes à l'aide de ces mêmes traits, à partir des indices et par l'allure locale de la courbe d'énergie.

On obtient, en fonction des dérivées première et seconde de la courbe d'énergie lissée, quatre types de courbes :

- sommet, creux de PS, creux de PV, autres courbes.

Les différents traits, ou classes de phonèmes, sont décrits par les arbres suivants :





Les segments ne correspondant à aucune de ces descriptions ne sont pas identifiés. La figure III.5 montre l'utilisation des seuils d'énergie et de l'allure locale de la courbe pour les traits VY, VO, PV, PS.

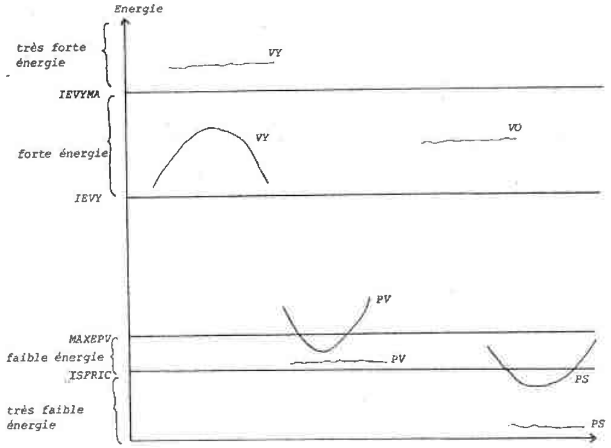


Figure III-5 : Utilisation des seuils d'énergie et de l'allure locale de la courbe d'énergie pour l'identification des traits VY, VO, PV, PS.

d) Exemples et propriétés des séquences de traits résultats.

Sur les exemples qui suivent, la liste des traits reconnus est accompagnée pour chaque trait trouvé des numéros du prélèvement de début et du prélèvement de fin du segment.

Exemples de séquences de traits résultats.

MOT PRONONCE:UN AMAS DE RIZ
PAR CHRISTIANE
LISTE DES TRAITS RECONNUS

N.	DEB	FIN		
VY	6	16	/œ̃/	
VO	17	21	/n/	
VY	28	38	/a/	omission de /m/
VY	53	65	/a/	
PV	66	74	/d/	
VY	76	84	/ə/	
PV	90	100	/R/	confusion
VY	101	109	/i/	

MOT PRONONCE:JE VEUX L ABORDER
PAR FLORENCE
LISTE DES TRAITS RECONNUS

N.	DEB	FIN		
FV	7	12	/ʒ/	
VY	14	21	/ə/	
PV	25	34	/v/	confusion
VY	37	52	/ø/	
VO	53	58	/ɛ/	
VY	59	64	/a/	
PV	66	72	/b/	
VY	75	80	/ɔ/	omission de /R/
PV	89	97	/d/	
VY	100	104	/e/	

MOT PRONONCE:ON AMASSAIT BEAUCOUP DE RATS
PAR JEAN-JULIEN
LISTE DES TRAITS RECONNUS

N.	DEB	FIN		
VY	7	12	/ɔ̃/	omission de /n/
VY	17	24	/a/	omission de /m/
VY	29	37	/a/	
FS	40	47	/s/	
VY	50	62	/ɛ/	
PV	65	68	/b/	omission de /o/
PS	77	81	/k/	omission de /u/
PV	88	91	/d/	
VY	93	101	/ə/	
PV	103	106	/R/	confusion
VY	107	118	/a/	

On remarque que, sur les quelques exemples présentés, ainsi que sur de nombreux autres cas, la séquence de traits résultats ne comporte jamais d'insertion de segment provoquée par le décodage phonétique, c'est-à-dire de sur-segmentation. Cette remarque est d'une grande importance pour l'élaboration de l'algorithme de cadrage.

Les seules erreurs de décodage sont des erreurs de confusion de traits et des omissions, ces dernières correspondant aux phonèmes non identifiés.

2.3.- Le cadrage.

a) Le problème.

Le but du cadrage est de mettre en correspondance les phonèmes de la transcription phonétique standard avec les segments de la séquence de traits résultats entachée d'erreurs, pour identifier avec certitude les segments de la phrase prononcée.

La résolution de ce problème peut s'appuyer sur des méthodes qui ont été mises en oeuvre pour la reconnaissance analytique de mots, bien que les objectifs soient différents.

Dans un cas il s'agit de reconnaître des phonèmes, dans l'autre cas il s'agit d'identifier des mots par l'intermédiaire de phonèmes.

La difficulté du cadrage, ainsi que sa raison d'être, proviennent de la présence d'erreurs dans la séquence de traits de la phrase analysée. Ces erreurs ont des origines différentes dont il faut faire la distinction pendant le cadrage :

- erreurs dues à une mauvaise acquisition

Ces erreurs proviennent des limites du programme d'acquisition ainsi que des conditions d'acquisition que nous exposons plus loin dans le paragraphe 2.5 a).

Ainsi, il faut tenir compte de la non-acquisition de certains phonèmes non voisés en début de mot que nous considérons comme une élision.

- erreurs dues à l'élocution.

Ces erreurs proviennent des altérations phonologiques provoquées par le locuteur lors de la prononciation de la phrase.

Nous considérons les erreurs suivantes :

Elision : correspond à un phonème non prononcé, ou trop faiblement prononcé notamment en fin de phrase,

Insertion : correspond à un phonème ajouté,

Substitution : correspond à une prononciation différente d'un même phonème provoquée par des accents différents.

- erreurs dues au décodage phonétique

Ces erreurs proviennent des faiblesses du système de décodage acoustico-phonétique et de l'analyse du signal.

Nous considérons les erreurs suivantes :

Omission : correspond à un phonème non identifié

Confusion : correspond à une identification erronée.

Notons que ce décodage phonétique ne provoque jamais d'insertions de phonèmes. Ces différentes erreurs peuvent être prises en compte soit dans la transcription standard de la phrase, soit par la matrice de confusion, soit par l'algorithme de cadrage. Dans le cas d'une erreur non prise en compte, la phrase est rejetée. Dans le tableau récapitulatif qui suit, sont répertoriés les différents types d'erreurs en fonction des différentes sources d'erreurs, accompagnés de leurs prises en compte.

Sources d'erreurs	ACQUISITION	ELOCUTION	DECODAGE
types d'erreurs	ELISION en début de mot ③	ELISION en fin de mot ③ autre ④ INSERTION prévisible ① autre ④ SUBSTITUTION ②	OMISSION ③ CONFUSION ②

① : transcription standard de la phrase
② : matrice de confusion
③ : algorithme de cadrage
④ : pas pris en compte.

b) Transcription phonétique standard.

Une phrase possède deux représentations internes :

- une séquence de traits représentant la transcription phonétique standard de la phrase, utilisée par le programme de cadrage, mémorisée dans le tableau ITRA (N).
- une représentation phonémique permettant au programme d'apprentissage automatique d'associer les traits cadrés aux phonèmes cadrés, mémorisée dans le tableau IPHON (N).

exemple : le mot APPATER est représenté par :

ITRA (5) :	VY	PS	VY	PS	VY
IPHON (5) :	a	p	a	t	e

Outre la description simple de la phrase sous forme de traits, cette transcription prend en compte les insertions prévisibles de petites pauses entre les mots ainsi que certaines insertions phonologiques. Ces segments ne sont pas représentés par une classe de phonèmes mais par 0.

exemple : insertion phonologique

	UN		OURS(E)		BLANC				
ITRA (9) :	VY	VO	VY	VO	FS	0	PV	VO	VY
IPHON (9) :	œ̃	n	u	R	s	␣	b	ɑ̃	ɑ̃

exemple : insertion d'une pause

	INCONNU					_	PENDANT			LE		JEU	
ITRA (14) :	VY	PS	VY	VO	VY	0	PS	VY	PV	VY	VO	VY	FV	VY
IPHON (14) :	ɛ̃	k	o	n	y	␣	p	ɑ̃	d	ɑ̃	ɑ̃	ə	ʒ	ø

Un cas particulier a été envisagé pour le phonème /i/ qui peut appartenir soit à la classe VY soit à la classe VO dans le cas où il est entouré de phonèmes voisés de forte énergie, ou en fin de mot.

exemple : phrase contenant /i/

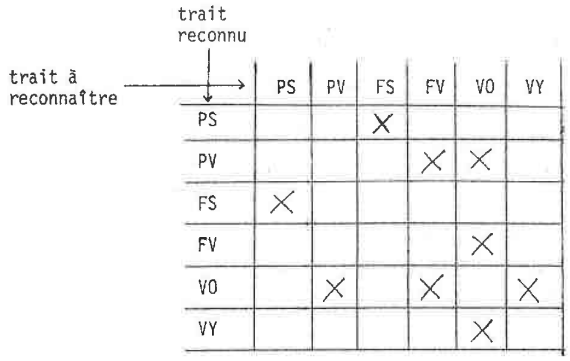
	DIX		ROTIS			
ITRA (6) :	PV	VO	VO	VY	PS	VO
IPHON (6) :	d	i	R	o	t	i

c) Matrice de confusion.

La matrice de confusion renseigne sur les confusions, substitutions admises pendant le cadrage. Elle a été obtenue d'une part par de nombreuses observations du signal vocal et des séquences de traits résultats, et d'autre part par l'application de règles phonétiques et phonologiques. Elle exprime donc les faiblesses de l'analyse et de l'acquisition du signal, les faiblesses du programme de recherche de traits appliqué à de nombreux locuteurs et certaines variations phonétiques.

Ne sont pas admises les substitutions (telle que PV - PS) qui peuvent se produire pour certains locuteurs et qui orientent le programme de cadrage vers des hypothèses fausses qui ne peuvent pas être détectées. Le choix et le nombre de confusions, substitutions admises est un compromis entre l'information apportée au programme de cadrage et le pouvoir d'identification des traits utilisés (si toutes les confusions sont permises les traits n'ont plus aucun pouvoir discriminant).

MATRICE DE CONFUSION



Justifications des confusions - substitutions :

trait à reconnaître		trait reconnu	
FS	-	PS	: dans le cas où la friction n'a pas été détectée dans la FS - due à l'analyse du signal incorrecte en haute fréquence.
PS	-	FS	: dans le cas où une friction a été détectée dans une PS - due à l'influence de l'environnement de la PS
FV	-	PV	: dans le cas où la friction n'a pas été détectée dans FV - due à l'analyse du signal incorrecte.
VO	-	PV	: dans le cas où la VO a très peu d'énergie - due à l'insuffisance des traits utilisés.
VO	-	FV	: dans le cas où la VO a peu d'énergie et possède un indice de friction - due à l'environnement de la VO (et à l'insuffisance des traits utilisés).
PV	-	VO	: dans le cas où la PV a une forte énergie. - due à l'insuffisance des traits utilisés.
VY	-	VO	: dans le cas où la VY a peu d'énergie - due à l'influence de l'environnement de la VY
VO	-	VY	: dans le cas où la VO a une très forte énergie - due à l'influence de l'environnement de la VO

Les améliorations envisageables de ces confusions sont d'une part d'utiliser des traits acoustico-phonétiques indépendants du locuteur plus nombreux pouvant porter sur la répartition de l'énergie dans le spectre (FONSALE-84) et d'autre part de tenir compte de l'environnement des phonèmes pendant l'identification des traits. Dans tous les cas cela nécessite une meilleure définition du signal dans les basses et hautes fréquences.

d) Algorithme de cadrage.

L'algorithme de cadrage consiste à associer à tous les traits de la transcription standard de la phrase les prélèvements centi-secondes correspondants dans l'énoncé avec l'aide de la séquence de traits reconnus et des segments de prélèvements associés.

Dans un premier temps on effectue un parcours de gauche à droite des deux séquences de traits en ne tenant pas compte des insertions prévisibles et en émettant des hypothèses d'omission. Ces hypothèses d'omission sont émises en fonction d'un paramètre temporel (nombre de prélèvements centi-secondes). Quand le cadrage n'aboutit pas, des retours-arrière sont possibles en émettant alors des hypothèses d'insertion dans le cas des insertions prévisibles. En dernier lieu, l'émission des hypothèses d'omission est remise en question en modifiant pas à pas le seuil de décision du paramètre temporel dans une limite raisonnable jusqu'à ce que le cadrage aboutisse. Dans le cas contraire le cadrage est déclaré impossible.

Les hypothèses d'insertions servent à ne pas faire d'éventuelles hypothèses d'omissions fausses.

En vue de valider le cadrage et d'améliorer l'apprentissage automatique par une modification des paramètres de décodage phonétique, chaque trait cadré est assorti d'un indice d'exactitude pouvant représenter la fiabilité du segment cadré.

Algorithme général :

Tentative de cadrage

- Initialisations
- Traitement de début de phrase

Tant que tous les traits de la transcription strandard ne sont pas cadrés et qu'il reste des traits de la séquence résultat non encore exploités répéter :

- Traitement général du cadrage du trait courant

Si cadrage du trait non possible alors

Si il y a des hypothèses modifiables alors

soit changement d'hypothèse et retour-arrière

soit on recommence une *Tentative de cadrage* avec un paramètre modifié

sinon

cadrage impossible

fsi

sinon

passer au trait suivant

fsi

fin tant que

- Traitement de fin de phrase

Les données :

- la séquence de traits standard : ITRA (LT)
- la séquence de traits reconnus : NATTR (NTR)
- les segments associés aux traits de NATTR et qui sont représentés par :

. le numéro du prélèvement de début des segments : DEBTR(NTR)
. le numéro du prélèvement de fin des segments : FINTR(NTR)

avec

LT : nombre de traits de la transcription standard
NTR : nombre de traits de la séquence résultat
LINF : taille minimum d'un phonème omis.
(seuil de décision pour les hypothèses d'omission).

Les résultats :

Si le cadrage est impossible : message d'erreur
sinon

- les segments associés aux traits de ITRA qui sont représentés par :

. le numéro du prélèvement de début des segments associés : DEB(LT)
. le numéro du prélèvement de fin des segments associés : FIN(LT)

- les indices d'exactitude des segments associés aux traits : POIDS(LT) .

Ces indices sont : 3 si égalité entre les traits
2 si confusion entre les traits
1 si omission de trait
0 si insertion prévisible ou élision en début ou fin de phrase.

Algorithmes détaillés :

Nous présentons en détail les différents modules de l'algorithme général :

- Initialisations
- Traitement de début de phrase
- Traitement général du cadrage du trait courant
- Traitement du cadrage de trait non possible
- Traitement de fin de phrase.

Notations : TCS : représente le trait courant de la transcription standard
TCR : représente le trait courant de la séquence de traits résultat.
INSER : représente le nombre d'hypothèses de non-insertion faites.

Initialisations : TCS = ITRA (1)
TCR = NATTR (1)
LINF = 4
INSER = 0

Traitement de début de phrase

‡ Test de début mal acquis
Elision de PS ou FS ‡

Si ITRA(1) = PS ou ITRA(1) = FS et
NATTR(1) ≠ FS et NATTR(1) ≠ PS alors
on associe à TCS : - aucun prélèvement
- l'indice d'exactitude 0
TCS = trait suivant
sinon rien.

Traitement général du cadrage du trait courant TCS

‡ insertion prévisible ‡

Si TCS = 0 alors on fait l'hypothèse qu'il n'y a pas d'insertion :
- on associe à TCS - aucun prélèvement
- l'indice d'exactitude 0
- mémorisation d'informations relatives à TCS et TCR en cas de retour-arrière
- INSER = INSER + 1

sinon

‡ Test égalité ‡

Si TCS = TCR alors
- on associe à TCS - le segment de TCR
- l'indice d'exactitude 3

sinon

‡ Test omission ‡

Si le nombre de prélèvements entre le segment courant et le segment précédent > LINF alors
- on fait l'hypothèse d'une omission :
- on associe à TCS - les prélèvements qui se trouvent entre le segment courant et le segment précédent
- l'indice d'exactitude 1

sinon

‡ Test confusion ‡

Si (TCS, TCR) appartient à la matrice de confusion
alors
- on associe à TCS - le segment de TCR
- l'indice d'exactitude 2

sinon
Le cadrage du trait n'est pas possible.

Traitement du cadrage de trait non possible

‡ changement d'hypothèse ‡

Si il existe des hypothèses de non-insertion (INSER > 0) alors
- retour-arrière : TCS = trait de la dernière hypothèse de non insertion faite
- on fait l'hypothèse d'insertion :
on associe à TCS - les prélèvements qui se trouvent entre la fin du trait cadré avant TCS et le début du segment suivant
- l'indice d'exactitude 0
- INSER = INSER - 1

sinon

Si il existe des hypothèses d'omission alors

- modification de LINF : LINF = LINF + 1
- si LINF ≤ limite supérieure que peut prendre LINF alors
  nouvelle *Tentative de cadrage*
  sinon
  le cadrage est impossible

sinon

le cadrage est impossible.

Traitement de fin de phrase

Le traitement de la fin d'une phrase est un problème délicat car c'est à ce moment qu'on décide si le cadrage est correct, alors que le dernier phonème d'une phrase est souvent mal prononcé ou mal acquis donc mal reconnu ou même élidé.

Si il ne reste plus de traits non encore cadrés alors
fin correct
sinon
si trop de prélèvements restant en fin de mot alors
cadrage impossible
sinon

‡ *Test trop de phonèmes manquants* ‡

si il reste plus d'un trait (différent d'une insertion prévisible)
non encore cadré alors
cadrage impossible
sinon

‡ *Test dernier trait élidé* ‡

si le nombre de prélèvement restant est suffisant alors
- on associe à TCS - ces prélèvements
  - l'indice d'exactitude 1

sinon

le dernier phonème a été élidé :

- on associe à TCS - aucun prélèvement
  - l'indice d'exactitude 0

e) Exemples de cadrage et cas de rejet de la phrase

exemples :

Dans les exemples de cadrage qui suivent une codification numérique des traits est utilisée dont voici la table :

insertion de prélèvements : 0
PS : 1
PV : 2
FS : 3
FV : 4
VO : 5
VY : 6

Les deux premiers exemples sont des exemples théoriques.
L'exemple 1 montre la succession de toutes les hypothèses faites ainsi que les modifications de LINF dans le cas où la transcription standard contient deux insertions prévisibles et où il y a une hypothèse d'omission. Le cadrage est impossible car la confusion (PV, FV) n'est pas admise.

L'exemple 2 illustre le traitement des élisions de phonèmes en début et fin de mot.

Les trois exemples qui suivent sont des exemples réels.

L'exemple 3 montre une succession de confusions et d'omissions, les exemples 4 et 5 illustrent la récupération d'hypothèses d'omission fausses par modification de LINF ou par émission d'hypothèse d'insertion.

Séquence de traits reconnus :

```
N.  DEB  FIN
PV   1   10
VY  11   20
PS  21   30
FV  31   40
VY  41   50
FV  61   70
VY  71   80

NOMBRE DE SEGMENTS DONNES  =    9
NOMBRE DE SEGMENTS TROUVES =    7

CHAINE STANDARD :   2  6  0  3  5  6  0  2  6

CHAINE TROUVEE  :   2  6  1  4  6  4  6
       LINF=   4
EMPILER 1
EMPILER 2
RETOUR 1
RETOUR 6
EMPILER 1
RETOUR 0
       LINF=   5
EMPILER 1
EMPILER 2
RETOUR 1
RETOUR 0
EMPILER 1
RETOUR 0
       LINF=   6
EMPILER 1
EMPILER 2
RETOUR 1
RETOUR 0
EMPILER 1
RETOUR 0
       LINF=   7
EMPILER 1
EMPILER 2
RETOUR 1
RETOUR 0
EMPILER 1
RETOUR 0
       LINF=   8
EMPILER 1
EMPILER 2
RETOUR 1
RETOUR 0
EMPILER 1
RETOUR 0
CADRAGE IMPOSSIBLE
```

Confusion non-admise

Séquence de traits standards :

PV
VY
O
FS
VO
VY
O
PV
VY

non-insertion 1, non-insertion 2
non-insertion 1, insertion 2
insertion 1, non-insertion 2
insertion 1, insertion 2

EXEMPLE THEORIQUE 1

Les hypothèses d'insertions

EXEMPLE THEORIQUE 2

Les élisions en début et fin de mot.

Séquence de traits reconnus

```
N.  DEB  FIN
VY   3   10
FV  11   20
VY  30   40
FS  41   50
VY  51   60
PS  61   70

NOMBRE DE SEGMENTS DONNES  =   10
NOMBRE DE SEGMENTS TROUVES =    6

CHAINE STANDARD :   1  6  4  0  2  6  3  6  1

CHAINE TROUVEE  :   6  4  6  3  6  1
       LINF=   4
EMPILER 1
```

Séquence de traits standards

PS
VY
FV
O
PV
VY
FS
VY
PS
VO

RESULTAT DU CADRAGE

NAT	DEB	FIN	EXAC	
PS	0	0	0	← *élision de PS*
VY	3	10	3	
FV	11	20	3	
∂∂	21	21	0	
PV	22	29	1	← *élision de PV*
VY	30	40	3	
FS	41	50	3	
VY	51	60	3	
PS	61	70	3	
VO	0	0	0	← *élision du dernier phonème*


```
MOT PRONONCE DIX NOUVEAUX ARRETS
PAR CHRISTINE
LISTE DES TRAITS RECONNUS
N. DEB FIN
PV   2  17
VY  21  29
PV  31  41
PV  50  60
VY  63  72
VY  84  94
PV  95 105
VY 106 113
VO 114 121

NOMBRE DE SEGMENTS DONNES  =   11
NOMBRE DE SEGMENTS TROUVES =    9

CHAINE STANDARD :   2  5  0  5  6  4  6  4  6  5  6

CHAINE TROUVEE  :   2  6  2  2  6  6  2  6  5
       LINF=   4
EMPILER 1
```

RESULTAT DU CADRAGE

NAT	DEB	FIN	EXAC	
PV	2	17	3	*égalité*
VO	21	29	2	*confusion*
∂∂	30	30	0	
VO	31	41	2	*confusion*
VY	42	49	1	*omission*
FV	50	60	2	*confusion*
VY	63	72	3	*égalité*
FV	73	83	1	*omission*
VY	84	94	3	*égalité*
VO	95	105	2	*confusion*
VY	106	113	3	*égalité*

Exemple 3 : Les confusions et omissions.

```
MOT PRONONCE ISSUE FICHEE
PAR NOELLE
LISTE DES TRAITS RECONNUS
N. DEB FIN
FS  21  37
VY  40  52
PS  63  67
FS  77  95
VY  99 110

NOMBRE DE SEGMENTS DONNES  =    8
NOMBRE DE SEGMENTS TROUVES =    5

CHAINE STANDARD :  6  3  6  0  3  6  3  6
CHAINE TROUVEE  :  3  6  1  3  6
       LINF=   4
EMPILER 1
RETOUR 0
```

Emission d'une hypothèse d'omission fausse, récupérée par une hypothèse d'insertion

RESULTAT DU CADRAGE

NAT	DEB	FIN	EXAC
VY	2	20	1
FS	21	37	3
VY	40	52	3
@@	53	62	0
FS	63	67	2
VY	68	76	1
FS	77	95	3
VY	99	110	3

Exemple 4 : Récupération d'une hypothèse d'omission fausse par une hypothèse d'insertion


```
MOT PRONONCE DIX NOUVEAUX ARRETS
PAR YOLANDE
LISTE DES TRAITS RECONNUS
N. DEB FIN
PV   3  16
VY  19  37
VO  38  47
VY  48  53
PV  59  64
VY  74  80
PV  83  90
VY  93 106
VY 115 129

NOMBRE DE SEGMENTS DONNES  =   11
NOMBRE DE SEGMENTS TROUVES =    9

CHAINE STANDARD :  2  5  0  5  6  4  6  4  6  5  6
CHAINE TROUVEE  :  2  6  5  6  2  6  2  6  6
       LINF=   4
EMPILER 1
RETOUR 0
       LINF=   5
EMPILER 1
RETOUR 0
       LINF=   6
EMPILER 1
```

Emission d'une hypothèse d'omission fausse récupérée par modification de LINF

RESULTAT DU CADRAGE

NAT	DEB	FIN	EXAC
PV	3	16	3
VO	19	37	2
@@	38	37	0
VO	38	47	3
VY	48	53	3
PV	59	64	2
VY	74	80	3
PV	83	90	2
VY	93	106	3
VO	107	114	1
VY	115	129	3

Exemple 5 : Récupération d'une hypothèse d'omission fausse par modification du seuil LINF

```
MOT PRONONCE DIX NOUVEAUX ARRETS
PAR JEAN-JULIEN
LISTE DES TRAITS RECONNUS
N. DEB FIN
PS  10  14
VO  17  24
VY  37  41
VY  53  72
VY  83  91
VY  97 102

NOMBRE DE SEGMENTS DONNES  =   11
NOMBRE DE SEGMENTS TROUVES =    6

CHAINE STANDARD :  2  5  0  5  6  4  6  4  6  5  6

CHAINE TROUVEE  :  1  5  6  6  6  6
        LINF=   4
        LINF=   5
        LINF=   6
        LINF=   7
        LINF=   8
CADRAGE IMPOSSIBLE
```

Substitution non admise

Exemple 6 : Cadrage impossible car substitution non admise

```
MOT PRONONCE INCONNU PENDANT LE JEU
PAR JEAN-JULIEN
LISTE DES TRAITS RECONNUS
N. DEB FIN
VY   4  12
PS  16  21
VY  26  30
PV  32  38
VY  40  52
PS  57  67
PV  81  85
VY  87  96
VO  97 106

NOMBRE DE SEGMENTS DONNES  =   14
NOMBRE DE SEGMENTS TROUVES =    9

CHAINE STANDARD :  6  1  6  5  6  0  1  6  2  6  5  6  4  6

CHAINE TROUVEE  :  6  1  6  2  6  1  2  6  5
        LINF=   4
EMPILER 1
TROP DE PHONEMES MANQUANT EN FIN DE MOT
```

Exemple 7 : Cadrage impossible car plusieurs omissions adjacentes en fin de phrase.

Cas de rejet de la phrase :

Le cadrage est impossible lorsque la phrase prononcée contient une insertion non prévisible, une élision (sauf en début et fin de mot), une confusion non plausible ou encore plusieurs omissions adjacentes (notamment en fin de mot). Voir les exemples 6 et 7 de cadrages impossibles par suite d'une substitution de PV par PS non admise, et par suite d'un manque trop important de phonèmes en fin de mot.

f) Validation du cadrage

Le but de la validation du cadrage est de vérifier, à partir des résultats du programme de cadrage, que les prélèvements qui ont été associés à un trait à la suite d'une hypothèse d'omission peuvent effectivement l'être. On procède à l'élimination des prélèvements incohérents de ces phonèmes à l'aide de critères simples. Si la taille du phonème restant est insuffisante (inférieure à la taille minimum du trait correspondant) la validation est impossible et le cadrage est refusé.

Les critères d'élimination des prélèvements sont :

pour les PV : prélèvements non voisés ou prélèvements de forte énergie,

pour les PS : prélèvements voisés ou prélèvements d'énergie pas très faible.

pour les FV,VO,VY : prélèvements non voisés.

L'exemple 8 illustre le cas d'une validation impossible se justifiant par le dévoisement des prélèvements d'un segment cadré FV. Sur l'exemple 9 on voit le résultat de l'élimination de certains prélèvements d'une PS et d'une PV.

```
MOT PRONONCE JE VEUX L ABORDER
PAR ODILE
LISTE DES TRAITS RECONNUS
N. DEB FIN
VY  11  16
PV  19  29
VY  31  41
VY  47  52
PV  53  60
VY  62  68
PV  73  85
VY  87  93

NOMBRE DE SEGMENTS DONNES  =   12
NOMBRE DE SEGMENTS TROUVES =    8

CHAINE STANDARD :   0  4  6  4  6  5  6  2  6  5  2  6

CHAINE TROUVEE  :   6  2  6  6  2  6  2  6
      LINF=   4
EMPILER 1
```

RESULTAT DU CADRAGE

NAT	DEB	FIN	EXAC
@@	1	1	0
FV	2	10	1
VY	11	16	3
FV	19	29	2
VY	31	41	3
VO	42	46	1
VY	47	52	3
PV	53	60	3
VY	62	68	3
VO	69	72	1
PV	73	85	3
VY	87	93	3

VALIDATION IMPOSSIBLE

```
PREL                    ENERGIE          MELODIE          HZ
 2:  3          54      )*                *              0
 3:  :         4::      )*                *              0
 4:  4         43:      )*                *              0
 5:  :         3:.       *                *              0
 6:  5         .:43     )*                *              0
 7:  5.      35463:     ) *               *              0
 8:  75.   :3465653     )   *             )            * 226
 9:  874.. 5456454.     )     *           )          *   240
10:  995534655645.      )       *         )          *   248
11:  9A66467666564      )         *       )          *   248
12:  9876468666566      )         *       )          *   248
13:  98765686664475     )          *      )          *   248
14:  98765786664365     )          *      )        *     256
15:  AB765786664465     )         *       )        *     256
16:  AB765766663455     )         *       )          *   252
17:  A875576545 :34     )       *         )          *   252
18:  AA55573444  .:     )     *           )          *   240
19:  984445::::   .     )   *             )        *     233
20:  864.:3.            ) *               )      *       217
21:  85: .              )*                )      *       220
22:  64                  *                )      *       217
23:  63                  *                )      *       214
24:  6:                  *                )          *   252
25:  5.                  *                )          *   252
26:  63                  *                )          *   240
27:  85                 )*                )            * 264
28:  863 33             ) *               )            + 279
29:  9853453  :         )   *             )            + 289
30:  A966575345   .     )     *           )            + 300
31:  8977696555 :44     )       *         )            + 312
32:  89777965651444     )       *         )            + 306
33:  8988796566:455     )         *       )            + 312
34:  8988797576:455     )         *       )            + 318
35:  8998797576:455     )         *       )            + 318
36:  8998797566.455     )         *       )            + 325
37:  8998788576:454     )         *       )            + 339
38:  8997788676.454     )         *       )            + 347
```

Exemple 8 : Validation impossible car prélèvements non-voisés dans une fricative voisée.


```
MOT PRONONCE ON AMASSAIT BEAUCOUP DE RATS
PAR GERALD
NO. D ANALYSEUR    0      MINI DEBUT          5      MINI FIN          20
MAJORATION BRUIT  450     BRUIT DE FOND     450     NBR DE PRELEV.   143
SEUIL DE PITCH=    75     SEUIL DE FRIC.=     4     ENERGIE MINI  D 1 FRICAT.=  200
VITESSES D UNE PS: 2  PV: 2  FS: 4  FV: 4  VO: 3  VY: 3
ENERGIE MINI  VOYELLE= 612     ENERGIE MAXI PV = 300     SEUIL MAXI FRICAT. =  10
NC1= 2   NC2= 7   NC3= 8   NC4=16
LISTE DES TRAITS RECONNUS
N. DEB FIN
VY   4  12
VO  13  18
VY  19  25
VY  33  39
FS  44  51
VY  54  65
PV  67  75
VY  78  83
VY  98 102
VY 109 120
VY 126 137

NOMBRE DE SEGMENTS DONNES  =   15
NOMBRE DE SEGMENTS TROUVES =   11

CHAINE STANDARD :   5  5  6  5  6  3  6  2  6  1  6  2  6  5  6

CHAINE TROUVEE  :   6  5  6  6  3  6  2  6  6  6  6
      LINF=   4
```

RESULTAT DU CADRAGE

NAT	DEB	FIN	EXAC
VO	4	12	2
VO	13	18	3
VY	19	25	3
VO	26	32	1
VY	33	39	3
FS	44	51	3
VY	54	65	3
PV	67	75	3
VY	78	83	3
PS	84	97	1
VY	98	102	3
PV	103	108	1
VY	109	120	3
VO	121	125	1
VY	126	137	3

RESULTAT VALIDATION

NAT	DEB	FIN	EXAC
VO	4	12	2
VO	13	18	3
VY	19	25	3
VO	26	32	1
VY	33	39	3
FS	44	51	3
VY	54	65	3
PV	67	75	3
VY	78	83	3
PS	88	91	1
VY	98	102	3
PV	104	107	1
VY	109	120	3
VO	121	125	1
VY	126	137	3

Exemple 9 : Elimination des prélèvements incohérents d'une plosive sourde et d'une plosive voisée pendant la validation.

2.4.- Améliorations.

Le mode interactif de la session d'apprentissage nous permet d'effectuer certaines améliorations, comme le rejet a priori d'une phrase mal acquise, ou l'affichage de conseils de prononciation en fonction des informations extraites des phrases prononcées, parvenant ainsi à une certaine éducation du locuteur.

La recherche de la partie stable des phonèmes conduit à une meilleure représentation, en vue de la reconnaissance, des formes de références moyennes, et fournit les données nécessaires à l'adaptation automatique des seuils d'énergie du système au fur et à mesure de l'apprentissage.

a) Rejet a priori de la phrase énoncée.

Une fois la phrase énoncée, le locuteur a la possibilité, comme dans le système centi-seconde, de la visualiser sur l'écran sous la forme d'un spectrogramme pour la rejeter dans le cas d'une mauvaise acquisition. Toujours dans l'optique de rendre le système accessible à des locuteurs non habitués à la lecture de spectrogrammes, nous avons réalisé un module de rejet automatique a priori de la phrase énoncée qui évite de poursuivre inutilement les traitements quand l'acquisition ou la prononciation ont été incorrectes. Les vérifications se font sur la représentation acoustique de la phrase et se décomposent en :

Vérification de durée :

Cas de rejet : Si la durée de la phrase acquise est inférieure à la durée minimum de la phrase . (La durée est représentée par le nombre total de prélèvements centi-secondes, la durée minimum est estimée par le nombre de phonèmes composant la phrase).

Vérification d'intensité :

Cas de rejet : si la phrase acquise a trop ou pas assez d'intensité

- pas assez d'intensité : si moins de quatre prélèvements sur l'ensemble des prélèvements de la phrase ont chacun une énergie > ENMIN
- trop d'intensité : si plus de quatre prélèvements ont une énergie > ENMAX

A l'issue de ces tests on fournit au locuteur des informations lui permettant de modifier son comportement ou les conditions d'acquisition. Toutefois il faut savoir que souvent, quand on demande à un locuteur de parler plus fort, il parle plus fort, mais aussi plus vite.

b) Recherche de la partie stable des segments

La recherche de la partie stable des segments cadrés a pour but d'obtenir des phonèmes de référence présentant les mêmes caractéristiques de stabilité que les phonèmes de référence de la phase d'apprentissage manuel décrite précédemment.

L'extraction automatique de la partie stable d'un segment consiste à éliminer les prélèvements des extrémités du segment, qui introduisent une certaine variabilité. Le critère de stabilité retenu est fondé sur l'énergie des prélèvements. Pour avoir une idée de la variabilité de cette énergie dans le segment nous calculons le coefficient de variation V des énergies :

$$V = \frac{\sigma}{\overline{EN}} \times 100$$

$\overline{EN}$: moyenne des énergies $\quad \overline{EN} = \frac{1}{N} \sum_{i=iD}^{iF} EN_i$

σ : écart-type des énergies $\quad \sigma = \sqrt{\sum_{i=iD}^{iF} \frac{(EN_i - \overline{EN})^2}{N - 1}}$

avec ID et IF : numéro des prélèvements de début et de fin de segment

N : nombre de prélèvements du segment (N = IF - ID + 1)

Un segment est considéré comme stable si V est inférieur au seuil limite de variation S.

S = 30 pour PS, PV, PS 1er groupe

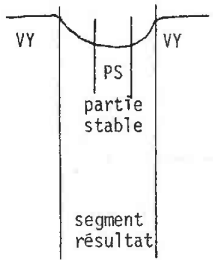
S = 15 pour FV, VO, VY 2ème groupe

Il est normal que le seuil S du 2ème groupe soit inférieur à celui du 1er groupe car, relativement à la moyenne des énergies assez forte des phonèmes du 2ème groupe, l'écart-type est plus petit que pour les phonèmes du 1er groupe (qui ont une énergie moyenne faible).

- Extraction de la partie stable des plosives sourdes :

Les segments PS résultats du décodage acoustico-phonétique possèdent aux extrémités, des prélèvements qui ont en général une énergie nettement supérieure à la moyenne des prélèvements du segment (entre autre le burst) et que nous éliminons :

exemples de courbes d'énergie



Algorithme :

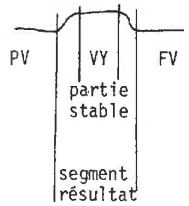
partie 1 {
- Tant que V > 30 et taille du segment > 3 prélèvements
- supprimer parmi les 2 prélèvements extrêmes celui de plus forte énergie
- Calcul de V
- fttque

partie 2 {
- Si taille du segment restant < 3 alors le segment est non stable
  sinon mémorisation des bornes du segment

- Extraction de la partie stable des voyelles :

Les segments VY résultats possèdent aux extrémités des prélèvements qui ont en général une énergie inférieure à la moyenne des prélèvements du segment, et qu'il faut éliminer.

exemple de courbe d'énergie



Algorithme :

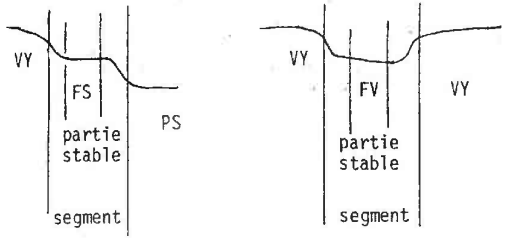
partie 1 {
- Tant que V > 15 et taille du segment > 3 prélèvements
- Supprimer parmi les deux prélèvements extrêmes celui de plus faible énergie
- Calcul de V
- fttque

partie 2 même chose que pour PS

- Extraction de la partie stable des consonnes voisées et fricatives sourdes.

Les segments résultats de ces classes de phonèmes possèdent aux extrémités des prélèvements dont l'énergie est plus basse ou plus élevée que la moyenne des prélèvements du segment.

exemple de courbe d'énergie



Algorithme :

partie ① { Tant que V > S et taille du segment ⩾ 3 prélèvements
Supprimer parmi les deux prélèvements extrêmes celui dont l'énergie est la plus éloignée de l'énergie moyenne
Calcul de V
fttque

partie 2 même chose que pour les PS

L'exemple 10 illustre d'une part le rejet a priori d'une première énonciation de la phrase, puis les différents traitements d'une phrase, jusqu'à la recherche de la partie stable des segments cadrés et validés. Sur la figure III.6 on voit le découpage final de la représentation acoustique de la phrase en phonèmes stables.


```
EST-CE UNE BONNE LOCUTION ? ←— rejet a priori de la phrase énoncée
PARLEZ + FORT

EST-CE UNE BONNE LOCUTION ?
 OUI

******

ESSAI  1
******

0000000011111111033333333311111111111111333333333311111333333333331111111111111111111100


MOT PRONONCE DE BONS GOUTERS
PAR GERALD
NO. D ANALYSEUR    0      MINI DEBUT        5      MINI FIN          20
MAJORATION BRUIT  450     BRUIT DE FOND    481     NBR DE PRELEV.    90
SEUIL DE PITCH=   75      SEUIL DE FRIC.=    4     ENERGIE MINI  D 1 FRICAT.=  200
VITESSES D UNE PS: 2  PV: 2  FS: 4  FV: 4  VO: 3  VY: 3
ENERGIE MINI  VOYELLE= 700     ENERGIE MAXI PV = 450     SEUIL MAXI FRICAT. =  10
NC1= 2   NC2= 7   NC3= 8   NC4=16
LISTE DES TRAITS RECONNUS
N. DEB FIN
PV   2   7
VY   9  15
PV  17  25
PV  40  49
PS  59  67
VY  71  83

NOMBRE DE SEGMENTS DONNES  =    8
NOMBRE DE SEGMENTS TROUVES =    6

CHAINE STANDARD :   2  6  2  6  2  6  1  6

CHAINE TROUVEE  :   2  6  2  2  1  6
       LINF=   5
```

RESULTAT DU CADRAGE

NAT	DEB	FIN	EXAC
PV	2	7	3
VY	9	15	3
PV	17	25	3
VY	26	39	1
PV	40	49	3
VY	50	58	1
PS	59	67	3
VY	71	83	3

RESULTAT VALIDATION

NAT	DEB	FIN	EXAC
PV	2	7	3
VY	9	15	3
PV	17	25	3
VY	26	39	1
PV	40	49	3
VY	50	58	1
PS	59	67	3
VY	71	83	3

PARTIES STABLES

NAT	DEB	FIN	ENM	ECA	EMA	EMI	STA
PV	2	6	58	4	64	53	F
VY	9	14	853	37	900	793	F
PV	19	24	191	13	207	165	F
VY	28	34	606	50	671	520	F
PV	42	45	160	15	182	144	F
VY	51	54	729	46	789	658	F
PS	59	66	28	3	34	23	F
VY	71	83	778	23	807	720	F

Exemple 10 : - Rejet a priori de la phrase énoncée

- Cadrage, validation et recherche de la partie stable des segments cadrés.

MOT PRONONCE DE BONS GOUTERS
PAR GERALD

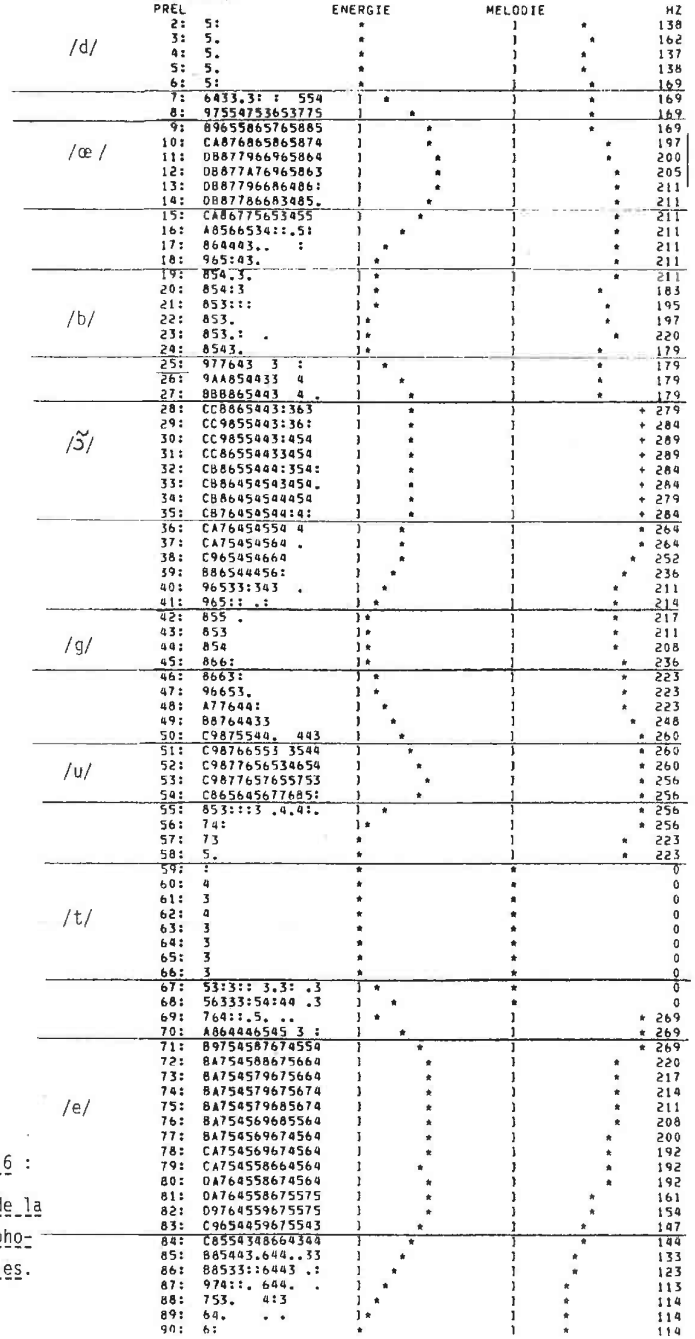


Figure III.6 : Découpage de la phrase en phonèmes stables.

c) Adaptation du système de reconnaissance de traits au locuteur.

Cette phase consiste à améliorer le système de reconnaissance de traits pendant la session d'apprentissage en modifiant automatiquement les paramètres et seuils du système en fonction d'informations relatives au locuteur extraites des phrases déjà cadrées.

- Modification du calcul du taux de friction.

Après avoir étudié les fricatives de nombreux locuteurs nous n'avons pas pu tirer de règles constantes concernant le calcul du taux de friction le mieux adapté au locuteur, la réalisation des fricatives étant trop variable et très dépendante du contexte (ceci étant en partie dû à l'analyse incorrecte du signal dans les hautes fréquences. voir chapitre II). On se contente donc d'un seul calcul de taux de friction donnant de bons résultats pour la plupart des locuteurs pour des fricatives en contexte favorable /i/, /y/ ...

- Ajustements des seuils d'énergie.

Les valeurs des énergies des différentes classes de phonèmes varient légèrement, de façon non linéaire d'un locuteur à l'autre. L'adaptation au locuteur consiste à ajuster les trois seuils MAXEPV, IEVY, IEVYMA (figure III-5) relatifs au paramètre d'énergie dans l'algorithme de classification de traits, à partir de la valeur de l'énergie moyenne de la partie stable de segments cadrés qui ont été mal ou pas identifiés par le système. Chaque seuil possède un domaine de variation possible pour éviter les modifications abusives ou incohérentes.

α) La première phrase :

La première phrase de la session d'apprentissage a un rôle particulier qui permet d'une part des vérifications sur les conditions d'acquisition et d'autre part de fournir quelques conseils au locuteur qui utilise

peut-être pour la première fois le système. A cet effet la phrase est composée de plosives sourdes et de voyelles de forte énergie.

Vérifications :

Des vérifications sont effectuées sur :

- les conditions d'acquisition : si il existe une PS qui contient dans sa partie stable un prélèvement dont l'énergie est supérieure à un seuil maximum alors on fait l'hypothèse qu'il y a trop de bruit de fond ou un mauvais réglage d'acquisition.
- la vitesse d'élocution : si il existe un phonème non stable alors on demande au locuteur de parler plus lentement.
- l'intensité du message : si il existe une voyelle (exceptée une voyelle de fin de phrase) qui ne contient pas de prélèvements dont l'énergie est supérieure à un seuil minimum alors on demande au locuteur de parler plus fort ou plus près du micro.

Ajustement de IEVYMA :

- le phonème /a/ étant considéré comme le phonème contenant le plus d'énergie, IEVYMA est systématiquement ajusté à la valeur moyenne des énergies moyennes des différents /a/ de la phrase.

β) Les autres phrases :

Pour chaque phrase on établit des hypothèses de mauvaise adaptation des seuils d'énergie en fonction des erreurs de confusion ou d'omission (repérées par l'indice d'exactitude du programme de cadrage) effectuées sur les différentes classes de phonèmes. Ces hypothèses sont alors vérifiées directement sur la valeur des énergies des phonèmes incorrectement reconnus et s'il y a lieu les seuils sont modifiés en fonction de leur valeur précédente et de la valeur de l'énergie de ces phonèmes.

Hypothèses émises en fonction des erreurs produites :

confusion de FV confusion de PV	⟶	IEVY trop petit
omission de VY	⟶	IEVYMA trop grand
omission de PV	⟶	MAXEPV trop petit
omission de VO	⟶	IEVY trop grand
confusion de VO	⟶	IEVY trop grand IEVYMA trop petit.

Exemple de modification :

Dans le cas d'une confusion de VO, si IEVY est trop grand alors IEVY prend la valeur moyenne entre son ancienne valeur et la valeur de l'énergie moyenne de la VO.

d) Construction du fichier des formes de référence des phonèmes.

Les formes de référence sont stockées sous la forme d'un fichier ayant la même structure que celui de l'apprentissage manuel mais qui peut contenir jusqu'à 70 phonèmes.

- Construction des formes de référence.

Pour chaque phonème cadré stable ne résultant pas d'une hypothèse d'élision prévisible, on mémorise la forme de référence moyenne correspondante, sous la forme :

E(1) à E(16) : moyennes des énergies des canaux des prélèvements constituant le phonème.

E(17) à E(21) : moyennes des répartitions d'énergie des prélèvements constituant le phonème.

- Construction du fichier

A la suite de la construction des formes de référence, chaque phonème possède une ou plusieurs formes de référence provenant des phonèmes extraits des différentes phrases. Afin de minimiser le nombre de formes à stocker dans le fichier, chaque nouvelle forme de référence d'un phonème est comparée aux formes correspondantes déjà créées dans le fichier. Si la distance entre les deux formes est faible, on remplace la forme de référence du fichier par la moyenne des deux formes, sinon on créé dans le fichier une nouvelle référence. Cette étape permet de diminuer de 25 % le nombre de formes de références à mémoriser sans altérer les performances du système de reconnaissance.

2.5.- Conditions expérimentales.

a) Les conditions d'enregistrement.

Un des problèmes majeurs que nous avons rencontré a été de dissocier les effets de la variabilité de la parole inter-locuteur de la variation du signal vocal due aux conditions d'acquisition. Comme on l'a vu précédemment la qualité du décodage phonétique à partir des traits est en grande partie fondée sur l'énergie des prélèvements centi-secondes. Les paramètres utilisés sont assez généraux pour présenter des qualités d'indépendance vis-à-vis des locuteurs, lorsque ceux-ci sont placés dans des conditions d'acquisition semblables.

Or, il faut noter que l'enregistrement a été effectué dans une salle machine dont le bruit ambiant varie en fonction des différentes activités s'y déroulant, et introduisant un bruit de fond plus ou moins important. Notons également que les appareils d'acquisition et d'analyse du signal servant à plusieurs chercheurs qui travaillent sur des applications différentes, les réglages des gains étaient fréquemment modifiés et difficilement reproductibles de façon identique d'une session à l'autre. Un autre

facteur de variabilité qu'il faut prendre en compte et qui conditionne la qualité du rapport signal sur bruit est la distance entre le micro et la bouche du locuteur qui fluctue considérablement.

b) Les phrases d'apprentissage.

Nous présentons les différents critères justifiant le choix des phrases à faire prononcer pendant la session d'apprentissage. Ces phrases doivent être limitées en nombre (< 10) pour que la session d'apprentissage reste courte, mais elles doivent contenir le maximum d'informations relatives aux phonèmes (différents contextes ...).

Ces phrases sont présentées par ordre de difficulté de décodage phonétique croissant, permettant ainsi au système d'adapter ses seuils au fur et à mesure de l'apprentissage, accompagné par une familiarisation progressive du locuteur au système d'acquisition.

Notons que pour restreindre les difficultés à surmonter par le système encore en phase d'apprentissage, ces phrases comportent le moins d'altérations phonologiques possibles telles qu'assimilations complètes de phonèmes à l'intérieur des mots ou à la jonction des mots.

De plus, connaissant certaines faiblesses de l'analyse du signal pour certains phonèmes, notamment les fricatives sourdes, ceux-ci sont placés dans des contextes favorables pour améliorer le décodage phonétique en traits.

Voici les sept phrases retenues, dans l'ordre de leur présentation au locuteur, avec leurs descriptions internes. Ces phrases permettent d'extraire en moyenne 65 phonèmes dans un temps relativement court, inférieur en moyenne à 5 minutes.

1 APPATER

VY PS VY PS VY

/a/ /p/ /a/ /t/ /e/

2 UN BON GOUTER

VY PV VY PV VY PS VY

/œ̃/ /b/ /ɔ̃/ /g/ /u/ /t/ /e/

3 ISSUE FICHEE

VY FS VY 0 FS VY FS VY

/i/ /s/ /y/ ␣ /f/ /i/ /ʃ/ /e/

4 JE VEUX L'ABORDER

0 FV VY FV VY VO VY PV VY VO PV VY

␣ /ʒ/ /ə/ /v/ /ø/ /l/ /a/ /b/ /ɔ/ /R/ /d/ /e/

5 DIX NOUVEAUX ARRETS

PV VO 0 VO VY FV VY FV VY VO VY

/d/ /i/ ␣ /n/ /u/ /v/ /o/ /z/ /a/ /R/ /ɛ/

6 INCONNU PENDANT LE JEU

VY PS VY VO VY 0 PS VY PV VY VO VY FV VY

/œ̃/ /k/ /o/ /n/ /y/ ␣ /p/ /ɑ̃/ /d/ /ɑ̃/ /l/ /ə/ /ʒ/ /ø/

7 ON AMASSAIT BEAUCOUP DE RATS

VY VO VY VO VY FS VY PV VY PS VY PV VY VO VY

/ɔ̃/ /n/ /a/ /m/ /a/ /s/ /ɛ/ /b/ /o/ /k/ /u/ /d/ /ə/ /R/ /a/

␣ insertion prévisible

Nous avons évalué à 1,5 le nombre moyen d'énonciations par phrase pour l'ensemble des locuteurs, dû à des impossibilités de cadrage ; les phrases nécessitant le plus de répétitions étant différentes d'un locuteur à l'autre, mais étant souvent les phrases :

4 à cause des élisions de phonèmes dues à certains locuteurs,

5 par manque d'intensité du dernier mot de la phrase ou par assourdissement du /d/ en début de phrase,

7 par manque d'intensité du dernier mot de la phrase.

c) Les locuteurs

Le corpus de locuteurs ayant testé l'apprentissage automatique, se compose de 20 individus dont 10 hommes et 10 femmes parmi lesquels on trouve différents accents : vosgien, alsacien, mosellan, parisien...

Sept personnes du corpus étaient habituées au système d'acquisition de la parole pour l'avoir déjà utilisé dans d'autres applications. Notons que pour un homme et une femme l'apprentissage n'a pas été possible, vu qu'ils produisaient systématiquement des phonèmes dévoisés à la place de phonèmes normalement voisés ; cette substitution n'étant pas admise dans la matrice de confusion. Il faut également remarquer que certains locuteurs non habitués au système ont eu tendance à parler trop lentement en détachant les syllabes ce qui faisait introduire au programme de cadrage des hypothèses d'omissions fausses. D'autres locuteurs ont diminué de façon trop excessive (pour le système) l'intensité des phonèmes de fin de phrase.

Une dernière remarque pour constater qu'en général la personne qui met au point un système de reconnaissance automatique de la parole a fortement tendance à s'adapter à son système, puisque, contrairement à un grand nombre de locuteurs du corpus, aucune répétition n'a été nécessaire au cours de son apprentissage.

3.- LA RECONNAISSANCE DE PHONEMES.

Afin de tester la validité des formes de références obtenues par apprentissage automatique nous avons évalué les performances du décodage acoustico-phonétique centi-seconde du système de reconnaissance de mots, en fonction de différentes formes de références (obtenues par apprentissage manuel, automatique et normalisées). Il ne nous a pas semblé nécessaire de revenir sur les performances du système de reconnaissance de mots (exposées dans le paragraphe 1.2.b) qui varient naturellement dans le même sens que la qualité du décodage acoustico-phonétique.

3.1.- Les différentes formes de références testées

a) Formes de références obtenues par apprentissage manuel.

L'ensemble de ces formes est limité entre 35 et 40 formes moyennes de phonèmes. Quelques phonèmes, /a/, /R/, .. ont deux représentants. Les voyelles sont en général acquises sans phonèmes adjacents, et les consonnes dans un environnement simple VCV.

b) Formes de références obtenues par apprentissage automatique.

L'ensemble peut aller jusqu'à 55 formes de références moyennes avec plusieurs représentants (jusqu'à quatre) pour un phonème. Ces différents représentants d'un phonème sont issus des divers contextes possibles présents dans les phrases de l'apprentissage.

c) Formes de références normalisées.

Nous avons construit, à partir du fichier des formes de référence obtenues par apprentissage automatique, un fichier de formes de référence normalisées par rapport à l'énergie totale de la forme afin de minimiser l'influence de la variation de l'intensité des phonèmes pendant leur identification.

Ainsi la valeur dans le canal k de la $i^{ème}$ forme de référence normalisée est donnée par :

$$ENORM_i(k) = (E_i(k) * 1000) / \left(\left(\sum_{j=1}^{NC} E_i(j) \right) + 1 \right)$$

avec $E_i(k)$: énergie dans le $k^{ème}$ canal de la forme de référence de l'apprentissage automatique,

NC : nombre de canaux.

3.2.- Résultats expérimentaux.

a) Les conditions de test.

Notre préoccupation principale étant de tester les formes de références de l'apprentissage automatique, c'est sur ces formes qu'ont porté le plus grand nombre d'essais. Ainsi huit locuteurs ont participé aux tests, prononçant en moyenne 80 voyelles et 75 consonnes dans des mots ou des phrases de difficultés diverses ; ces phrases étant approximativement les mêmes pour tous les locuteurs. (Voir le détail de ces phrases en annexe 3).

Certains tests se sont déroulés le même jour que l'apprentissage pour conserver des conditions d'acquisition semblables. Ceci présente l'inconvénient de supprimer quelques effets de variabilité de la parole intra-locuteurs.

Les tests sur les références de l'apprentissage manuel et les références normalisées n'ont été effectués que par cinq locuteurs appartenant à l'ensemble des huit locuteurs précédemment cités.

b) Les performances

Le tableau III.1 donne :

- pour les voyelles : le pourcentage moyen, pour l'ensemble des locuteurs, des voyelles correctement reconnues, c'est-à-dire apparaissant dans les trois phonèmes réponses du décodage centi-seconde.
- pour les consonnes : le pourcentage moyen, pour l'ensemble des locuteurs, des consonnes globalement correctement reconnues à l'intérieur des classes (/p/, /t/, /k/), (/b/, /d/, /g/), (/v/, /z/, /ʒ/), (/f/, /s/, /ʃ/), (/m/, /n/), (/ɲ/), (/R/). Ceci signifie que le module de décodage propose dans les trois phonèmes réponses au moins un phonème appartenant à la classe du phonème à reconnaître.

Tableau III.1.

Scores moyens inter-locuteurs de reconnaissance de phonèmes en fonction des formes de références utilisées.

	références de l'apprentissage automatique	références de l'apprentissage manuel	références de l'apprentissage automatique normalisées
Voyelles	0,80	0,65	0,83
Consonnes	0,70	0,75	0,61

Ce tableau qui représente des moyennes inter-locuteurs conserve les tendances observées pour chacun des locuteurs :

- l'amélioration de l'identification des voyelles à l'aide des références de l'apprentissage automatique, par rapport à l'apprentissage manuel, et qui s'explique par le nombre plus important des formes représentant un phonème.

- l'amélioration de l'identification des voyelles obtenue par les formes de références normalisées et qui est surtout due à la meilleure reconnaissance des voyelles /a/, /ø/, /œ/.

- la détérioration des scores pour la reconnaissance des consonnes à l'aide des références de l'apprentissage automatique qui est due en grande partie à la confusion fréquente des fricatives sourdes par des plosives sourdes ; ceci provenant de l'environnement particulier des fricatives sourdes extraites de la phrase d'apprentissage [isy fiʃe] qui favorise la présence

d'énergie dans les hautes fréquences que l'on ne retrouve pas obligatoirement dans des environnements différents. Notons que cette détérioration est beaucoup moins importante pour les fricatives normalisées.

- le score extrêmement faible pour la reconnaissance des consonnes normalisées qui est dû à la mauvaise identification des consonnes /l/, /m/, /n/ et /R/, lesquelles sont confondues avec des voyelles de faible énergie /i/, /u/, /5̃/, /o/.

CONCLUSION

Les formes normalisées demandent à être étudiées plus en détail et pour un nombre de locuteurs plus important. Pour avoir de bons scores de reconnaissance à partir des références de l'apprentissage automatique il faudrait ajouter une phrase pendant l'apprentissage pour représenter les fricatives sourdes dans des contextes différents de ceux déjà existants.

Toutefois, on peut estimer que l'apprentissage automatique, qui ne demande aucune intervention humaine et qui est très rapide, fournit avec sûreté des phonèmes de bonne qualité pour un nombre important d'individus.

Les travaux en cours dans notre laboratoire nous permettent d'envisager des améliorations quant à la reconnaissance des phonèmes, notamment en tenant compte de l'influence de l'environnement des phonèmes pendant leur identification. Au niveau de la reconnaissance de mots, il faut préciser que nous n'utilisons qu'un seul lexique pour tous les locuteurs potentiels. Actuellement la description d'un mot ne peut contenir qu'un seul phonème de substitution et d'insertion pour chaque phonème du mot. On peut envisager une amélioration en constituant un lexique plus complet contenant toutes les substitutions et insertions possibles effectuées par les différents utilisateurs du système.

CHAPITRE IV

ANALYSE DES VOYELLES MULTILOCUTEURS

INTRODUCTION.

Les nombreuses données obtenues par apprentissage automatique et provenant de 15 locuteurs ont été réunies afin de pouvoir mener une étude multilocuteur des voyelles représentées par les valeurs des canaux d'un vocoder. Les premières études entreprises relèvent de la statistique descriptive et ont pour but de mieux représenter et observer les formes de références multilocuteurs. Ainsi une technique de classification automatique est utilisée pour définir d'éventuels groupes homogènes de locuteurs. Ceci peut permettre de réduire les données à mémoriser dans un système de reconnaissance automatique de la parole. Ensuite une interprétation statistique est entreprise sous la forme d'une analyse en régression multiple pour rechercher les relations qui peuvent exister entre les formes de références des différentes voyelles des différents locuteurs. Ceci devrait nous permettre de générer automatiquement les références d'un locuteur en fonction de seulement quelques voyelles propres à celui-ci.

1.- LES VOYELLES TRAITEES

L'ensemble des voyelles étudiées est constitué de 10 voyelles orales /a/, /i/, /u/, /e/, /ɛ/, /y/, /o/, /ɔ/, /ə/, /ø/, et de 3 voyelles nasales /$\tilde{ɔ}$/, /$\tilde{a}$/, /$\tilde{ɛ}$/. Celles-ci ont été obtenues à partir des voyelles de l'apprentissage automatique effectué par 15 locuteurs dont 9 hommes et 6 femmes. Afin de faciliter les analyses statistiques, on ne conserve pour chaque locuteur qu'un seul représentant (c'est-à-dire un spectre à court terme) par voyelle. Ce nouveau fichier réunissant les formes de références uniques est obtenu par l'algorithme suivant :

Pour chaque locuteur :

Pour chaque voyelle :

Si une seule référence initiale alors
la référence unique = la référence initiale
sinon
* *plusieurs références initiales* *
pour chaque référence initiale :
on recherche les deux premiers canaux contenant un pic d'énergie
fin pour
on moyenne les références qui ont les pics d'énergie sur les mêmes canaux
la référence unique = la référence moyenne pour laquelle le plus grand nombre de références initiales interviennent dans le calcul de la moyenne.

fin pour

fin pour

Cette sélection des références à partir de l'emplacement des pics d'énergie permet de moyenner des spectres variables en amplitude mais proches sur le plan spectral, et permet également de s'abstraire des variations pas toujours significatives en haute fréquence. Notons en dernier lieu, que l'énergie dans le 16ème canal étant généralement nulle pour les voyelles nous avons supprimé cette donnée. Les voyelles traitées se résument donc à la donnée des énergies dans 15 bandes de fréquence pour 13 voyelles de 15 locuteurs. Sur le plan statistique, cela représente pour une voyelle un petit échantillon et limite donc les interprétations que l'on peut faire.

2.- ANALYSES STATISTIQUES

2.1.- Les voyelles moyennes multi-locuteurs.

Les formes de références de ces voyelles sont obtenues par :

$$E_{VM_j}(i) = \sum_{\ell=1}^{NLOC} E_{V_j}^{\ell}(i) \;/\; NLOC$$

où $E_{VM_j}(i)$ est l'énergie calculée du ième canal pour la voyelle moyenne j

NLOC est le nombre de locuteurs

$E_{V_j}^{\ell}(i)$ est l'énergie du ième canal pour la voyelle j du locuteur ℓ .

Comme les voyelles moyennes peuvent servir directement ou indirectement dans les systèmes multilocuteurs, nous avons calculé certaines matrices de confusion afin d'établir le degré de ressemblance entre ces références moyennes et les voyelles propres aux locuteurs.

a) Matrices de confusion

Nous présentons les résultats de deux groupes de matrices de confusion.

Les matrices ①, ② et ③ qui font partie du premier groupe donnent pour chaque voyelle, la répartition des voyelles moyennes les plus proches des voyelles des 15 locuteurs.

Dans les matrices de confusion ④ , ⑤ et ⑥ on comptabilise le nombre de fois où la voyelle moyenne correspondant à la voyelle étudiée fait partie des deux voyelles moyennes les plus proches.

A l'intersection de la ligne i et de la colonne j d'une matrice on trouve le nombre d'individus dont la voyelle, correspondant à la ligne i, a été affectée à la classe de la voyelle moyenne j.

Ce qui différencie les matrices ①, ② , ③ ou ④, ⑤ ,⑥ est l'évaluation différente de la distance entre deux formes. En effet, à la vue des spectres vocoder des voyelles des 15 locuteurs (Annexe 1) on se rend compte d'une grande variabilité, pour certaines voyelles, dans les hautes fréquences. C'est pourquoi l'évaluation des différentes distances fait intervenir les énergies dans les 15 canaux ou dans les 12 ou 13 premiers canaux.

La distance générale utilisée est :

$$d_{V,VM} = \sum_{i=1}^{NC} \left| E_V(i) - E_{VM}(i) \right|$$

avec V la voyelle à comparer
VM la voyelle moyenne
$E_V(i)$ l'énergie dans le canal i de la voyelle V
NC le nombre de canaux intervenant dans la somme :

pour les matrices ① et ④ NC = 15
② et ⑤ NC = 13
③ et ⑥ NC = 12

b) Analyse des résultats

Sur la matrice de confusion ① on remarque notamment que :

- 4 locuteurs possèdent un /a/ plus proche du /ɛ̃/ moyen que du /a/ moyen. Inversement 3 locuteurs possèdent un /ɛ̃/ plus proche du /a/ moyen que du /ɛ̃/ moyen.
- Relativement peu de locuteurs ont un /e/ proche du /e/ moyen, /i/ et /y/ moyens étant souvent plus proches.
- La voyelle /ə/ est plus souvent plus proche du /ø/ moyen que de /ə/ moyen.

MATRICES DU 1er GROUPE : CHOIX DE LA PLUS PROCHE VOYELLE MOYENNE.

MATRICE DE CONFUSION ①

A	I	OU	EI	AI	U	AU	E	O	EU	ON	IN	AN	
11	0	0	0	0	0	0	0	0	0	0	4	0	A
0	12	0	0	3	0	0	0	0	0	0	0	0	I
0	0	11	0	0	0	1	0	0	3	0	0	0	OU
0	0	2	9	1	0	0	1	0	0	0	2	0	EI
0	2	1	0	7	4	0	1	0	0	0	0	0	AI
0	0	1	0	3	10	0	0	0	1	0	0	0	U
0	0	1	0	0	0	8	0	2	0	2	1	1	AU
0	0	0	3	0	0	0	4	0	7	0	1	0	E
0	0	0	0	0	0	0	0	12	0	0	0	3	O
0	0	1	2	1	0	0	2	0	8	0	1	0	EU
0	0	2	0	0	0	0	0	2	0	10	0	1	ON
3	0	0	0	0	0	0	1	1	1	0	9	0	IN
0	0	0	0	0	0	0	0	3	0	2	0	10	AN

distances calculées sur les 15 canaux

MATRICE DE CONFUSION ②

A	I	OU	EI	AI	U	AU	E	O	EU	ON	IN	AN	
12	0	0	0	0	0	0	0	0	0	0	3	0	A
0	10	0	0	5	0	0	0	0	0	0	0	0	I
0	0	11	0	0	0	1	0	0	3	0	0	0	OU
0	0	2	10	1	0	0	1	0	0	0	1	0	EI
0	2	1	0	8	3	0	1	0	0	0	0	0	AI
0	0	0	0	3	11	0	0	0	1	0	0	0	U
0	0	2	0	0	0	8	0	2	0	1	1	1	AU
0	0	0	2	0	0	0	9	0	4	0	0	0	E
0	0	0	0	0	0	1	0	12	0	0	0	2	O
0	0	1	1	1	0	0	4	0	7	0	1	0	EU
0	0	2	0	0	0	1	0	2	0	10	0	0	ON
3	0	0	0	0	0	0	1	0	0	0	11	0	IN
0	0	0	0	0	0	0	0	3	0	2	0	10	AN

distances calculées sur les 13 premiers canaux

MATRICE DE CONFUSION ③

A	I	OU	EI	AI	U	AU	E	O	EU	ON	IN	AN	
12	0	0	0	0	0	0	0	0	0	0	3	0	A
0	9	0	0	6	0	0	0	0	0	0	0	0	I
0	1	10	0	0	0	1	0	0	3	0	0	0	OU
0	0	1	9	1	0	0	1	0	1	0	2	0	EI
0	2	0	1	9	0	0	0	0	0	0	0	0	AI
0	0	0	0	2	12	0	0	0	1	0	0	0	U
1	0	1	0	0	0	8	0	2	0	2	0	1	AU
0	0	0	2	0	0	0	9	0	4	0	0	0	E
0	0	0	0	0	0	0	0	12	0	0	0	3	O
0	0	1	1	1	0	0	4	0	7	0	1	0	EU
0	0	2	0	0	0	1	0	1	0	10	0	1	ON
5	0	0	1	0	0	0	1	0	0	0	8	0	IN
0	0	0	0	0	0	0	0	3	0	2	0	10	AN

distances calculées sur les 12 premiers canaux

MATRICES DU 2ème GROUPE : CHOIX PARMI LES DEUX PLUS PROCHES VOYELLES MOYENNES.

MATRICE DE CONFUSION ④

A	I	OU	EI	AI	U	AU	E	O	EU	ON	IN	AN	
15	0	0	0	0	0	0	0	0	0	0	0	0	A
0	14	0	0	1	0	0	0	0	0	0	0	0	I
0	0	12	0	0	0	0	0	0	3	0	0	0	OU
0	0	2	10	1	0	0	1	0	0	0	1	0	EI
0	0	0	0	14	0	0	1	0	0	0	0	0	AI
0	0	1	0	0	13	0	0	0	1	0	0	0	U
0	0	1	0	0	0	11	0	1	0	1	1	0	AU
0	0	0	0	0	0	0	13	0	1	0	1	0	E
0	0	0	0	0	0	0	0	15	0	0	0	0	O
0	0	1	1	0	0	0	0	0	12	0	1	0	EU
0	0	0	0	0	0	0	0	2	0	12	0	1	ON
1	0	0	0	0	0	0	1	1	1	0	11	0	IN
0	0	0	0	0	0	0	0	0	0	1	0	14	AN

distances calculées sur les 15 canaux

MATRICE DE CONFUSION ⑤

A	I	OU	EI	AI	U	AU	E	O	EU	ON	IN	AN	
15	0	0	0	0	0	0	0	0	0	0	0	0	A
0	12	0	0	3	0	0	0	0	0	0	0	0	I
0	0	11	0	0	0	1	0	0	3	0	0	0	OU
0	0	2	11	1	0	0	0	0	0	0	1	0	EI
0	0	0	0	14	0	0	1	0	0	0	0	0	AI
0	0	0	0	0	14	0	0	0	1	0	0	0	U
0	0	2	0	0	0	11	0	1	0	0	1	0	AU
0	0	0	1	0	0	0	14	0	0	0	0	0	E
0	0	0	0	0	0	1	0	14	0	0	0	0	O
0	0	1	1	0	0	0	0	0	12	0	1	0	EU
0	0	0	0	0	0	0	0	2	0	13	0	0	ON
1	0	0	0	0	0	0	0	0	0	0	14	0	IN
0	0	0	0	0	0	0	0	1	0	1	0	13	AN

distances calculées sur les 13 premiers canaux

MATRICE DE CONFUSION ⑥

A	I	OU	EI	AI	U	AU	E	O	EU	ON	IN	AN	
15	0	0	0	0	0	0	0	0	0	0	0	0	A
0	12	0	0	3	0	0	0	0	0	0	0	0	I
0	0	11	0	0	0	1	0	0	3	0	0	0	OU
0	0	1	11	1	0	0	0	0	1	0	1	0	EI
0	0	0	1	14	0	0	0	0	0	0	0	0	AI
0	0	0	0	0	14	0	0	0	1	0	0	0	U
1	0	1	0	0	0	12	0	0	0	1	0	0	AU
0	0	0	0	0	0	0	15	0	0	0	0	0	E
0	0	0	0	0	0	0	0	14	0	0	0	1	O
0	0	1	1	1	0	0	0	0	11	0	1	0	EU
0	0	2	0	0	0	0	0	1	0	11	0	1	ON
0	0	0	0	0	0	0	1	0	0	0	14	0	IN
0	0	0	0	0	0	0	0	1	0	1	0	13	AN

distances calculées sur les 12 premiers canaux

- En général les confusions se font entre voyelles qui sont voisines sur le triangle acoustique : exemple : /i/ - /e/, /u/ - /ø/

Sur la matrice ② obtenue en calculant les distances sans tenir compte des 3 derniers canaux on remarque que, comparativement à la matrice ① :

- le /i/ moyen est moins souvent le plus proche phonème du /i/ des locuteurs et ceci à l'avantage du /e/ moyen.
- par contre le /ə/ moyen gagne 5 locuteurs au détriment du /ø/ et du /i/ moyens.

En résumé, par rapport à la matrice ① , dans la matrice ② les voyelles moyennes
* /a/, /e/, /ɛ/, /y/, /ə/, /$\tilde{\varepsilon}$/ : sont plus représentatives pour un plus grand nombre de locuteurs
* /i/, /ø/ : sont moins représentatives
* /u/, /o/, /$\tilde{ɔ}$/, /$\tilde{ɔ}$/, /$\tilde{a}$/ : restent aussi satisfaisantes.

Par rapport à la matrice ① , dans la matrice ③ les voyelles moyennes
* /a/, /e/, /ə/ : sont plus représentatives
* /i/, /u/, /y/, /ø/, /$\tilde{\varepsilon}$/, /ɛ/: sont moins représentatives
* /o/, /ɔ/, /$\tilde{ɔ}$/, /$\tilde{a}$/ : restent aussi satisfaisantes.

Dans la matrice ④ il reste des voyelles moyennes, /o/, /$\tilde{\varepsilon}$/ et /ɛ/ qui ne font pas partie des deux voyelles moyennes les plus proches pour au moins 4 locuteurs. Ceci signifie que ces voyelles moyennes sont trop éloignées des voyelles d'un trop grand nombre de locuteurs pour pouvoir être de bonnes formes de références pour un système multilocuteur.

On voit sur les matrices ①, ② et ③ ou ④, ⑤ et ⑥ que les différentes distances favorisent ou défavorisent certaines voyelles. Pour se rendre compte de l'influence globale des distances sur l'ensemble des voyelles nous avons calculé le nombre moyen de locuteurs bien classés. Ces nombres figurent sur le tableau IV.1 sous l'intitulé : moyenne inter-voyelle. Les pourcentages figurant sur ce tableau représentent pour chaque matrice le pourcentage moyen interlocuteur des voyelles proches de la voyelle moyenne correspondante.

Tableau IV.1

Matrice	①	②	③	④	⑤	⑥
moyenne inter-voyelle	9,3	9,9	9,6	12,76	12,92	12,8
pourcentage	62 %	66 %	64 %	85 %	86 %	85 %

On remarque donc que la suppression des 3 derniers canaux dans le calcul des distances favorise la reconnaissance des voyelles multilocuteurs à partir des voyelles moyennes. Par contre la suppression d'un canal supplémentaire commence à faire diminuer les performances. Les pourcentages, inférieurs à 70 % des matrices ① , ② et ③ indiquent que les voyelles moyennes ne peuvent pas être utilisées telles quelles dans un système multilocuteur.

2.2.- Les voyelles centrées réduites

Nous proposons ici une normalisation des spectres des voyelles en calculant les énergies centrées réduites des canaux à partir de la moyenne et de l'écart type des énergies des 15 canaux de la voyelle. L'énergie normalisée dans le canal i d'une voyelle est :

$$ECR(i) = \frac{E(i) - EM}{\sigma}$$

avec $E(i)$: énergie dans le canal i

EM : énergie moyenne du spectre de la voyelle

$$EM = \sum_{k=1}^{NC} E(k)/NC$$

σ : écart-type de l'énergie dans le spectre

$$\sigma = \sqrt{\sum_{k=1}^{NC} (E(k) - EM)^2 /NC}$$

avec NC : nombre de canaux.

Nous n'avons pas pu expérimenter ces nouvelles références du fait de problèmes matériels. Toutefois le rapprochement important des courbes des voyelles multilocuteurs ainsi normalisées autour de la courbe moyenne montre l'efficacité de cette normalisation.

Les figures IV.1.a et IV.1.b, portant sur les spectres des locuteurs masculins et féminins pris séparément pour la voyelle /i/, illustrent bien la réduction de la variabilité entre les courbes obtenues à partir des valeurs prises en sortie du Vocoder (courbes brutes) et les courbes obtenues à partir des valeurs centrées réduites. (La courbe moyenne est représentée en trait plein)

On peut faire la même remarque pour la voyelle /ɛ/ (figures IV.2.a et IV.2.b) bien que pour les locuteurs féminins il apparaisse que cette normalisation ne soit pas suffisante et qu'une normalisation fréquentielle soit nécessaire.

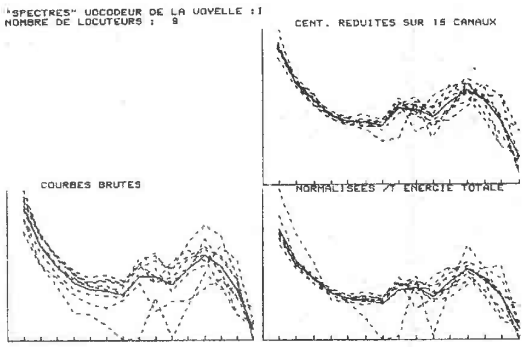


Figure IV.1.a : Voyelle /i/ - locuteurs masculins

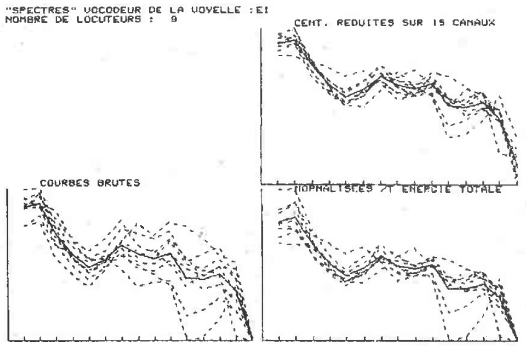


Figure IV.2.a : Voyelle /ɛ/ - locuteurs masculins

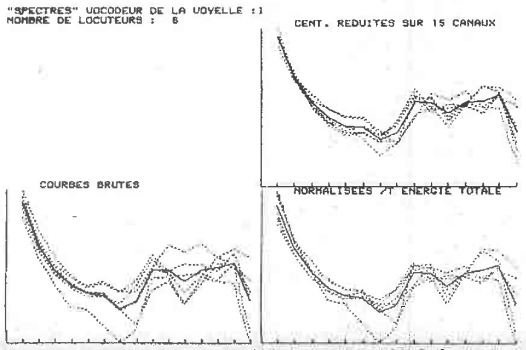


Figure IV.1.b : Voyelle /i/ locuteurs féminins

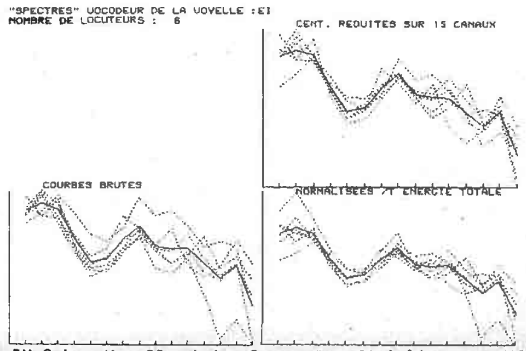


Figure IV.2.b : Voyelle /ɛ/ - locuteurs féminins

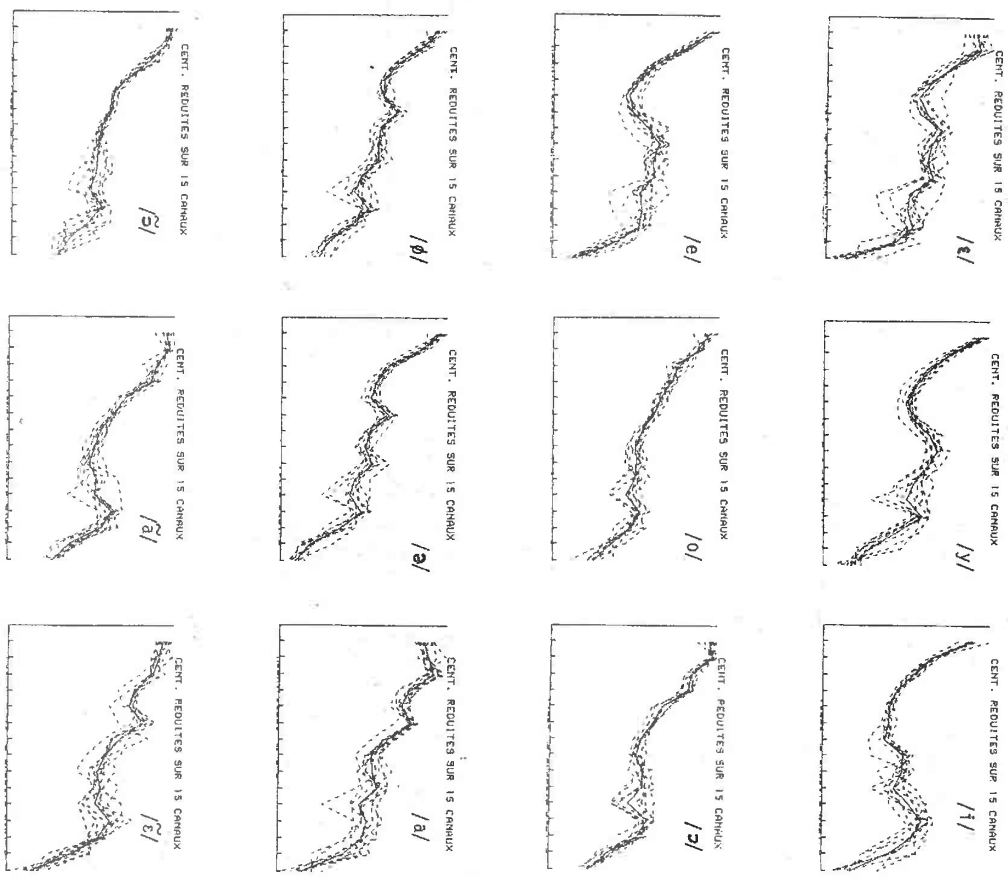
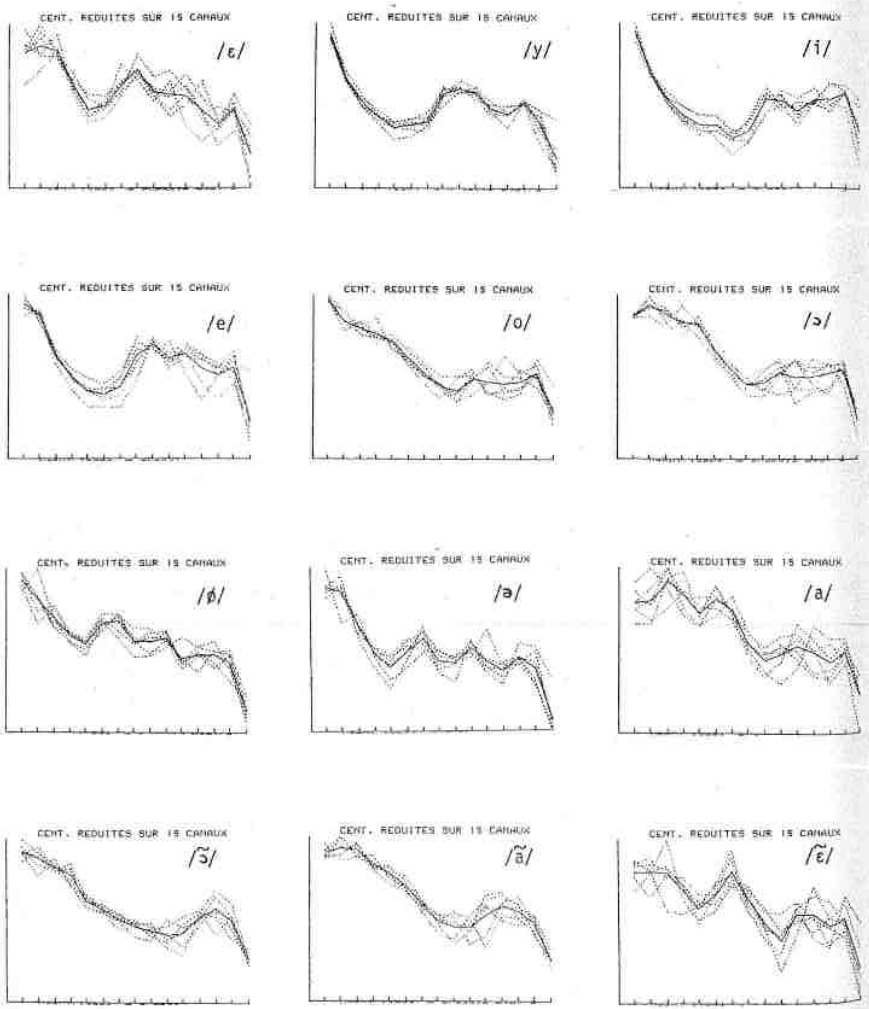


Figure IV.3 : Spectres centrés réduits des voyelles de neuf locuteurs masculins

Figure IV.4 : Spectres centrés réduits des voyelles de six locuteurs féminins



Les résultats de cette normalisation pour toutes les voyelles sont présentés sur la figure IV.3 pour les locuteurs masculins et IV.4 pour les locuteurs féminins. On y remarque que pour certaines voyelles, /ø/, /y/... il suffit de considérer séparément les sexes pour que les formes de références ainsi construites soient très proches les unes des autres pour les différents locuteurs. D'autre part la répartition de l'énergie dans les différentes bandes de fréquences est beaucoup plus variable chez les locuteurs féminins que chez les locuteurs masculins notamment pour les voyelles /a/, /ɛ/, /ɛ̃/. Il est intéressant de noter que pour les locuteurs masculins les voyelles /y/, /i/, /ɔ/, /o/, /e/, /ø/, /ə/, /œ̃/ présentent une très grande stabilité de la répartition de l'énergie dans les sept premiers canaux (250 Hz ⟶ 1900 Hz).

En conclusion la normalisation de l'énergie à partir de la moyenne et de l'écart-type des énergies dans les canaux permet d'obtenir pour quelques voyelles des formes de références semblables pour des locuteurs de même sexe. Pour les autres voyelles cette normalisation devra être accompagnée d'une normalisation fréquentielle.

2.3.- Recherche d'une typologie de locuteur par classification hiérarchique

Nous avons appliqué la méthode de la classification hiérarchique sur l'ensemble des locuteurs pour une voyelle donnée. Ceci a été motivé par la recherche d'éventuelles classes de locuteurs exploitables dans les systèmes multilocuteurs.

a) La méthode :

Soit P : le nombre d'individus à classer, c'est-à-dire les 15 locuteurs.

Soit NC : le nombre de caractères, c'est-à-dire les énergies dans les 15 canaux d'une forme de référence,

Soit X(P, NC) : la matrice des données, dont l'élément x_{ij} représente l'énergie dans le canal j de la voyelle donnée pour le locuteur i

Soit $\bar{X}(NC)$: le vecteur représentant la voyelle moyenne, dont l'élément

$$\bar{x}_j = \frac{1}{p} \sum_{i=1}^{p} x_{ij}$$

$1 \leqslant j \leqslant NC$

On calcule la matrice $D_0(P,P)$ des distances entre les formes de références des différents locuteurs pour la voyelle donnée. L'élément de la matrice d_{i_1,i_2} qui représente la distance normalisée par la variance entre les formes de références des locuteurs i_1 et i_2 est donnée par :

$$d_{i_1,i_2} = \sum_{j=1}^{NC} (x_{i_1 j} - x_{i_2 j})^2 / \sigma_j^2$$

$\{1 \leqslant i_1 \leqslant p\}$

$\{1 \leqslant i_2 \leqslant p\}$

avec $\sigma_j^2 = \frac{1}{p} \sum_{i=1}^{p} (x_{ij} - \bar{x}_j)^2$

$\{1 \leqslant j \leqslant NC\}$

On cherche à regrouper les différents locuteurs en classes en fonction des distances entre leurs formes de références :

Algorithme de classification hiérarchique :

- On considère initialement l'ensemble constitué par les 15 classes formées d'un seul élément : $C_1 = \{i_1\}$,... $C_p = \{i_p\}$ et la matrice des distances $D = D_0$

Pour i de 1 à p-1 :

- On cherche parmi les p-i+1 classes formées précédemment : $C_1, \ldots, C_{p-i+1}$ les 2 classes les plus proches, soit C_a et C_b.
- On agrège C_a et C_b en une nouvelle classe $C_{aUb} = \{ i \in C_a \text{ ou } i \in C_b \}$

- On cherche la nouvelle matrice des distances D en calculant les p-i-1 distances des classes, autres que C_a et C_b, à la nouvelle classe formée C_{aUb}, à l'aide d'une des trois méthodes de calcul de distances entre parties d'ensembles : INF, SUP et MOY décrites ci-après.

fin pour

- On agrège les deux dernières classes.

Remarque : La distance entre C_a et C_b représente l'indice de stratification de la classification hiérarchique, qui est porté en ordonnée sur la représentation graphique des arbres.

Le critère INF correspond au critère du plus proche voisin et le critère SUP au critère du voisin le plus éloigné. Ils font tous les deux partie des méthodes hiérarchiques ordinales. Le critère MOY fait partie des méthodes hiérarchiques non ordinales.

INF : $d(C, C_{aUb}) = \text{Inf}\{d(C,C_a), d(C,C_b)\}$

SUP : $d(C, C_{aUb}) = \text{Sup}\{d(C,C_a), d(C,C_b)\}$

MOY : $d(C, C_{aUb}) = \frac{1}{2}\left(d(C,C_a) + d(C,C_b)\right)$.

b) Résultats.

Nous présentons sur les figures IV.5 et IV.6 les résultats de la classification hiérarchique pour les voyelles /a/ et /ɔ/. Les trois arbres de chaque page illustrent les trois méthodes de calcul des distances.

Rappelons que les locuteurs masculins sont numérotés de 1 à 9 et les locuteurs féminins de 10 à 15.

Nous définissons un groupe de locuteurs comme étant une classe de locuteurs pour un indice de stratification donnée, ce qui nous permet de trouver des groupes homogènes de locuteurs.

On trouvera dans (CHANDON-81) les effets des différents critères sur les résultats de la classification. On retrouve notamment la propriété de chaînage du critère du plus proche voisin sur les figures IV.5.a et IV.6.a, qui donne des résultats peu satisfaisants : pas de groupes séparés de locuteurs.

Par contre les deux autres critères permettent de trouver des groupes de locuteurs bien séparés. Ainsi pour /ɔ/ les deux critères donnent pour un même niveau de stratification, des groupes séparés de locuteurs presque identiques (figures IV.6.b et IV.6.c)

- du côté masculin on a deux ensembles de locuteurs qui font toujours partie du même groupe : {2, 3, 4, 6} et {1, 7, 8, 9}
- de même du côté féminin on a l'ensemble {10, 11, 13, 14} .

En dernier lieu, nous pouvons noter qu'en utilisant le même critère (du voisin le plus éloigné) les locuteurs 1, 5, 7 et 9 appartiennent à un même groupe pour les deux voyelles /ɔ/ et /a/.

Toutes voyelles confondues on a fréquemment les locuteurs 7, 1, 9 dans le même groupe ainsi que les locuteurs 11, 13, 14. (Ceci respecte la discrimination de sexe bien connue !). Le locuteur 15 peut être considéré comme un locuteur particulier dont les références des voyelles ne sont jamais semblables aux références des voyelles des autres locuteurs. En effet il est souvent agrégé en fin de classification, ou se trouve dans des groupes réduits de locuteurs agrégés également en fin de classification.

En conclusion, la classification hiérarchique des locuteurs peut sans doute permettre, de minimiser le nombre des références de phonèmes à mémoriser dans un système multilocuteur.

Figures VI.5 : Voyelle /a/

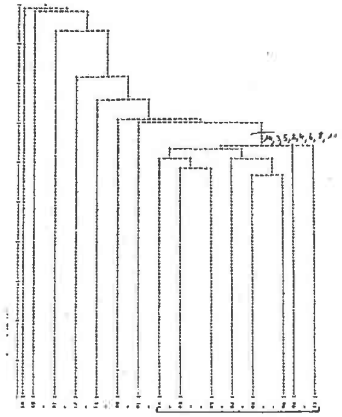


Figure IV.5.a

Plus proche voisin

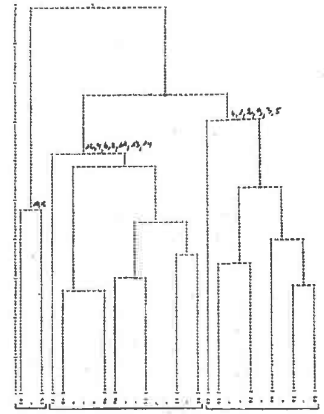


Figure IV.5.b

Voisin le plus éloigné

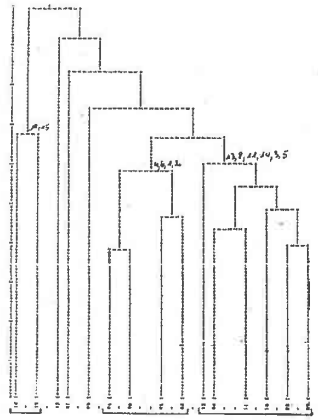


Figure IV.5.c

Moyenne des distances

Figures IV.6. : Voyelle /ɔ/

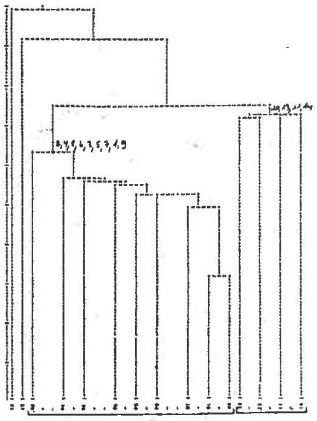


Figure IV.6.a

Critère : Plus proche voisin

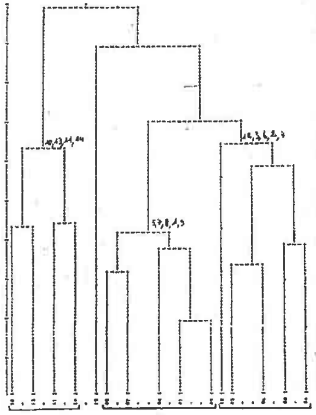


Figure IV.6.b

Critère : Voisin le plus éloigné

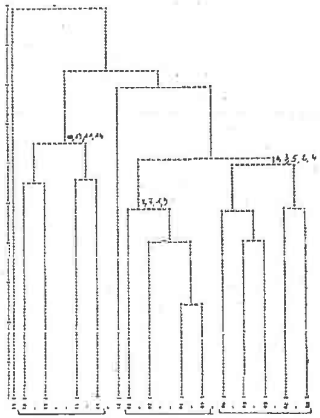


Figure IV.6.c

Critère : moyenne des distances

En outre, il serait intéressant d'effectuer la classification hiérarchique sur les références de toutes les voyelles confondues des différents locuteurs pour se rendre compte de la proximité des voyelles multilocuteurs entre elles.

2.4.- Génération automatique des références des voyelles

a) Objectifs

En vue de réduire la phase d'apprentissage des phonèmes pour un nouveau locuteur, une analyse statistique a été menée, qui doit permettre de générer automatiquement toutes les formes de références des voyelles à partir d'un sous-ensemble restreint de voyelles propres au locuteur. Ainsi nous faisons l'hypothèse que chaque nouveau locuteur n'a besoin d'énoncer que quelques voyelles et que toutes les voyelles de références peuvent alors être estimées à l'aide d'un ensemble de règles de transformation sur les quelques voyelles d'apprentissage. Nous pensons a priori que les trois voyelles /a/, /i/ et /u/,sommets du triangle acoustique des voyelles, peuvent faire partie des voyelles d'apprentissage.

Les règles de transformation sont obtenues à partir des voyelles de références des quinze locuteurs précédemment étudiées. Une analyse en régression multiple permet d'établir les relations linéaires, supposées indépendantes du locuteur, entre les voyelles.

Cette étude a été inspirée des travaux de FURUI (FURUI-80) mais porte sur des données et des hypothèses de travail différentes. En effet dans son système les phonèmes sont représentées par dix coefficients de prédiction linéaire, et les phonèmes d'apprentissage sont extraits des mots du vocabulaire qui présentent le plus de difficulté à être bien reconnus dans un environnement multilocuteur.

Nous n'avons pas pu mener à terme cette étude car l'ordinateur sur lequel les travaux ont été commencés n'a plus été disponible. Nous présentons donc essentiellement l'analyse en régression multiple accompagnée des premiers résultats obtenus. Une proposition pour l'estimation de toutes les voyelles à partir des quelques voyelles d'apprentissage est présentée sans avoir été testée.

b) Analyse en régression multiple

On trouvera dans (RAO-73) une présentation complète de l'analyse en régression multiple. Nous appliquons cette méthode au cas particulier des formes de référence des voyelles.

- Soit EV l'ensemble de toutes les voyelles $\{Y_1,\ldots, Y_i,\ldots, Y_{NV}\}$ avec NV = 13
- Soit EVE l'ensemble des voyelles d'apprentissage $\{X_1,\ldots, X_j,..X_{NVE}\}$ que l'on nomme voyelles explicatives. Dans le cas présent NVE = 3.

On fait l'hypothèse qu'il existe une relation linéaire entre les formes de références de chaque voyelle Y_i de EV et celles de chaque voyelle X_j de EVE de la forme :

$$Y_i = \alpha_{ij}\, X_j + E_{ij} \qquad (1)$$

où E_{ij} représente un vecteur aléatoire résiduel,

α_{ij} représente la matrice des coefficients de régression entre les deux voyelles Y_i et X_j .

Dans la suite, pour faciliter l'écriture nous omettrons les indices i et j des voyelles.

Y et X étant des vecteurs à NC composantes qui représentent les énergies dans les NC canaux (NC = 15) on pose :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{NC} \end{pmatrix} \qquad X = \begin{pmatrix} X_1 \\ \vdots \\ X_{NC} \end{pmatrix}$$

$$\alpha = \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,NC} \\ \vdots & & \vdots \\ \alpha_{NC,1} & \cdots & \alpha_{NC,NC} \end{pmatrix}$$

l'équation (1) devient :

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_{NC} \end{pmatrix} = \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,NC} \\ \vdots & & \vdots \\ \alpha_{NC,1} & \cdots & \alpha_{NC,NC} \end{pmatrix} X \begin{pmatrix} X_1 \\ \vdots \\ X_{NC} \end{pmatrix} + \begin{pmatrix} E_1 \\ \vdots \\ E_{NC} \end{pmatrix}$$

Elle peut s'écrire sous la forme de NC relations linéaires à NC variables explicatives :

$$Y_k = \alpha_{k,1}\, X_1 + \cdots \alpha_{k,NC}\, X_{NC} + E_k \qquad (2)$$

$1 \leq k \leq NC$

En clair, ceci signifie que le canal k de la voyelle Y peut être expliqué par une combinaison linéaire des valeurs dans les NC canaux de la voyelle X .

Estimation de α :

A partir des observations provenant de la prononciation par L locuteurs (L = 15) des deux voyelles considérées X et Y on construit les deux matrices x(NC,L) et y(NC,L) (appelées matrices de structures) des énergies mesurées sur les NC canaux. Ces données permettent d'estimer au sens des moindres carrés la matrice des coefficients de régression.

On note :

$$y = \begin{pmatrix} y_{1,1} & \cdots & y_{1,L} \\ \vdots & & \vdots \\ y_{NC,1} & \cdots & y_{NC,L} \end{pmatrix} \qquad x = \begin{pmatrix} x_{1,1} & \cdots & x_{1,L} \\ \vdots & & \vdots \\ x_{NC,1} & \cdots & x_{NC,L} \end{pmatrix}$$

↑ forme de référence de la voyelle y du locuteur 1

↑ forme de référence de la voyelle X du locuteur L

- Soit y_k le vecteur $(y_{k,1}, \ldots, y_{k,L})$ des valeurs prises par le $k^{ème}$ canal pour les différents locuteurs (c.à.d. la $k^{ème}$ ligne de y) et
- α_k le vecteur $(\alpha_{k,1}, \ldots, \alpha_{k,NC})$ (c.à.d la $k^{ème}$ ligne de α)
- Soit $A = x\, x^t$

α_k est alors estimé par : $\hat{\alpha}_k = A^{-1}\, xy_k$

d'où la matrice estimée de α : $\boxed{\hat{\alpha} = A^{-1}\, xy}$

On note $e_{k\ell}$ le résidu de la régression pour le $k^{ème}$ canal du $\ell^{ème}$ locuteur :

$$e_{k\ell} = y_{k,\ell} - \sum_{n=1}^{NC} \hat{\alpha}_{k,n}\, x_{n,\ell}$$

Les résidus $e_{k\ell}$ $(1 \leq k \leq NC\ ,\ 1 \leq \ell \leq L)$ représentent la différence entre les valeurs y_k réellement observées et celles qui sont fournies en fonction des $x_1, \ldots, x_{NC}$ par l'équation de regression et ceci pour chaque locuteur. Il permet donc d'apprécier l'adéquation du modèle au cas étudié.

- Les coefficients de corrélation multiple

Le coefficient de corrélation multiple donne la valeur globale de la régression sur toutes les observations. C'est-à-dire qu'il renseigne sur la valeur explicative de l'équation de régression obtenue.

Par définition le carré du coefficient de corrélation multiple est le rapport de la variance de la voyelle expliquée par la régression à la variance totale de la voyelle observée. L'estimation est d'autant meilleure que ce carré est proche de 1.

Le calcul du carré du coefficient de corrélation R^2_{ijk} pour le $k^{ième}$ canal de la voyelle Y_i expliquée par la voyelle X_j est :

$$R^2_{ijk} = \frac{\hat{\alpha}_k \times y_k^t}{\text{variance de } y_k}$$

Exemple des valeurs des carrés des coefficients de corrélation multiple pour les régressions de 15 canaux de la voyelle nasale /ɛ̃/ par les canaux de la voyelle /a/ :

canaux	R^2
1	.78833E+00
2	.97493E+00
3	.95849E+00
4	.79492E+00
5	.53086E+00
6	.99819E+00
7	.97090E+00
8	.99661E+00
9	.77241E+00
10	.87624E+00
11	.70441E+00
12	.77554E+00
13	.87643E+00
14	.81764E+00
15	.83414E+00

Il apparaît sur cet exemple que pour le canal 5 dont $R^2 = 0,53$, la valeur explicative de l'équation de régression obtenue concernant ce canal est faible par rapport aux équations concernant les canaux 6 et 8 pour lesquels R^2 est proche de 1.

c) Résultats.

Nous avons effectué quelques régressions en choisissant pour chaque régression deux voyelles proches sur le triangle acoustique : régressions de /ø/ par /y/, /e/ par /y/, /i/ par /y/, /ɔ/ par /a/, /ɛ̃/ par /a/ et /o/ par /u/ .

Les interprétations que l'on peut faire des résultats des régressions obtenues portent sur les coefficients de corrélation multiple et sont essentiellement comparatives. Toutefois on pourra considérer ici qu'un coefficient de corrélation est bon si son carré est compris entre 0,9 et 1 et médiocre s'il est inférieur à 0,9.

Ainsi on s'aperçoit sur les exemples de résultats qui suivent que les régressions des canaux de /ø/ par /y/ sont meilleures que celles de /e/ par /y/ et ceci pour la presque totalité des canaux.

Carré des coefficients de corrélation multiple des régressions des 15 canaux de :

/e/ par /y/

canaux ↓	R^2 ↓
1	.87120E+00
2	.46979E+00
3	.53750E+00
4	.67599E+00
5	.71496E+00
6	.72028E+00
7	.92477E+00
8	.88901E+00
9	.97024E+00
10	.75575E+00
11	.97314E+00
12	.91381E+00
13	.81893E+00
14	.81959E+00
15	.87290E+00

/ø/ par /y/

canaux ↓	R^2 ↓
1	.96155E+00
2	.81817E+00
3	.98325E+00
4	.93136E+00
5	.96605E+00
6	.96262E+00
7	.87000E+00
8	.94063E+00
9	.98971E+00
10	.94176E+00
11	.99553E+00
12	.90723E+00
13	.95584E+00
14	.96692E+00
15	.84491E+00

On s'aperçoit que sur l'ensemble des régressions effectuées, les résultats obtenus sont peu satisfaisants, c'est-à-dire que beaucoup de R^2 sont inférieurs à 0,9. Cela s'explique par l'importante variabilité des références étudiées.

C'est pourquoi nous avons tenté d'appliquer cette analyse aux formes centrées réduites présentées au §.2.2 de ce chapitre. On remarque alors une très nette amélioration des coefficients de corrélation multiple comme le montre les résultats de la régression de la voyelle /i/ par la voyelle /y/, toutes deux centrées réduites, comparativement aux résultats de la même régression sur les voyelles non centrées réduites.

Carré des coefficients de corrélation multiple des régressions des 15 canaux de /i/ par /y/ :

voyelles non centrées réduites

canaux ↓	R^2 ↓
1	.86354E+00
2	.82102E+00
3	.77027E+00
4	.68671E+00
5	.68474E+00
6	.64007E+00
7	.65633E+00
8	.82663E+00
9	.95677E+00
10	.73109E+00
11	.81389E+00
12	.90058E+00
13	.84755E+00
14	.98241E+00
15	.91800E+00

voyelles centrées réduites

canaux ↓	R^2 ↓
1	.88014E+00
2	.73860E+00
3	.72858E+00
4	.85012E+00
5	.97993E+00
6	.99539E+00
7	.93409E+00
8	.86803E+00
9	.97044E+00
10	.97348E+00
11	.97835E+00
12	.99037E+00
13	.94900E+00
14	.96946E+00
15	.81964E+00

On trouve en moyenne pour l'ensemble des régressions étudiées sur les références centrées réduites 10 canaux sur 15 correctement expliqués par voyelle.

d) Propositions pour l'estimation des références des voyelles.

L'ordinateur n'étant plus disponible nous n'avons pas pu déterminer les voyelles qui sont les plus explicatives pour l'ensemble de toutes les voyelles. Ceci nécessite d'étudier les résultats de toutes les régressions des voyelles entre elles sous la forme de 156 tableaux de 15 coefficients de corrélation multiple, pour déterminer les voyelles qui fournissent les meilleurs résultats.

C'est pourquoi nous pensons faire intervenir dans la formule de l'estimation de la forme de référence d'une voyelle Y_i :

- Les trois voyelles /a/ /i/ /u/ sommets du triangle acoustique
- Les coefficients de régression et les coefficients de corrélation multiple des régressions de la voyelle Y_i par ces trois voyelles
- La voyelle moyenne $\overline{Y}_i$ (étudiée au §.2.1).

Estimation de la voyelle Y_i :

$$\hat{Y}_i = \frac{r}{\sum_{j=1}^{NVE} w_{ij}} \sum_{j=1}^{NVE} w_{ij}\ \hat{\alpha}_{ij}\ X_j + (1-r)\ \overline{Y}_i$$

avec NVE = 3 , les voyelles X étant /a/ /i/ /u/ et

$$w_{ij} = \frac{1}{NC} \sum_{k=1}^{NC} R^2_{ijk}$$

w_{ij} est le coefficient de pondération d'une voyelle explicative qui donne relativement plus d'importance aux voyelles qui ont un bon pouvoir explicatif (au sens de la moyenne de la somme des carrés des coefficients de corrélation multiple pour les différents canaux).

Selon la valeur du coefficient r on peut donner plus ou moins d'importance à la voyelle moyenne interlocuteur (r proche de o) ou à l'estimation

pondérée de cette voyelle par les voyelles explicatives (r proche de 1), dans l'estimation de la voyelle Y_i d'un locuteur.

Nous n'avons malheureusement par pu tester cette formule et trouver la valeur optimum de r qui fournit les résultats de l'estimation les meilleurs. Dans un premier temps r = 0,5 semble être une solution intéressante (FURUI-80) Il est sans doute possible de diminuer le nombre des données à mémoriser pour prévoir les références des voyelles, notamment en utilisant des variables qui résument au mieux les différents canaux de chaque voyelle explicative, par exemple par des composantes principales.

Toutefois les régressions multiples sur les formes centrées réduites des voyelles ont fourni des résultats satisfaisants que nous comptons exploiter dans la suite de nos travaux.

Conclusion.

Toutes les études abordées dans ce chapitre gagneraient à porter sur un plus grand nombre d'observations donc de locuteurs, ainsi que sur un plus grand nombre de représentants par voyelle en prenant en compte différents contextes. Elles nous ont tout de même permis de tirer des conclusions intéressantes quant à la connaissance des références des voyelles des quinze individus. Certaines propositions fondées sur l'interprétation des résultats statistiques ont été fournies en vue de l'élaboration ou de l'amélioration de systèmes de reconnaissance de la parole multilocuteurs. Toutefois, il reste à les tester réellement dans un système de reconnaissance pour obtenir les résultats expérimentaux qui permettraient de les confirmer.

En outre nous pensons qu'une meilleure représentation du signal de la parole, par des références obtenues par prédiction linéaire ou par analyse cepstrale ou par une analyse spectrale faite à partir d'un plus grand nombre de filtres améliorerait la qualité de l'information traitée, donc l'interprétation des résultats des analyses statistiques.

CONCLUSION

L'apprentissage automatique que nous avons réalisé a deux buts essentiels :

- l'adaptation au locuteur du système de reconnaissance de mots centiseconde, par extraction automatique des formes de références représentées par les sorties d'un analyseur spectral.

- la composition d'un corpus de formes de référence de voyelles multilocuteurs, propice à des études statistiques pour la recherche de certaines propriétés des phonèmes ainsi représentés.

Cet apprentissage possède les caractéristiques suivantes :

. il est opérationnel pour un nombre élevé de locuteurs très divers, grâce à l'utilisation de traits acoustico-phonétiques peu dépendants du locuteur et grâce à l'adaptation des procédures de décodage et de cadrage de phonèmes par l'ajustement automatique de certains paramètres pendant la session.

. il est rapide, une session dure de 2 à 5 minutes.

. il est entièrement automatique donc agréable d'utilisation, car il ne requiert pas de la part du locuteur des connaissances particulières quant à la représentation du signal vocal.

. il réalise par l'apport de certains conseils l'éducation du locuteur et permet une familiarisation de ce dernier au système d'acquisition.

Certaines analyses statistiques ont été entreprises sur les références des voyelles de 15 locuteurs et des propositions ont été émises en vue de l'élaboration ou l'amélioration de systèmes de reconnaissance de la parole multilocuteur. Des problèmes matériels ne nous ont pas permis de mettre en oeuvre ces différentes propositions qui portent sur la recherche de procédures de normalisation des références, la recherche de groupes homogènes de locuteurs et la recherche d'une méthode de génération automatique des références des voyelles.

L'ensemble des indices acoustico-phonétiques indépendants du locuteur utilisés dans notre système, a permis de réaliser une segmentation et une reconnaissance assez grossière en classes de phonèmes, lesquelles ont été effectivement expérimentées sur vingt locuteurs dans des conditions d'acquisition variables.

La recherche d'invariants plus nombreux et plus précis, fondés sur la répartition de l'énergie dans les différentes bandes de fréquence des formes de références centrées réduites semble être prometteuse compte-tenu de la grande stabilité observée de l'énergie dans les sept premiers canaux (250-1900 Hz) lorsque l'on considère séparément les sexes.

Nous comptons poursuivre nos travaux sur le décodage acoustico-phonétique multilocuteur par l'utilisation simultanée des formes de références de phonèmes et un ensemble de règles actuellement à l'étude dans notre laboratoire permettant notamment la prise en compte de l'influence de l'environnement sur les phonèmes.

BIBLIOGRAPHIE

(AIMARD-74) AIMARD P.
"L'enfant et son language"
SIMEP EDITIONS 1974

(BAUDRY-81) BAUDRY M.
"Recherche de traits acoustiques dans le domaine temporel. Formalisation à l'aide de techniques syntaxiques".
Actes du séminaire "Processus d'encodage et de décodage phonétiques" (p. 157-167) Toulouse sept. 1981.

(BAUDRY-80) BAUDRY M. - DUPEYRAT B.
"Speech Analysis Using syntactic methods and a pitch synchronous Formant detector on the direct signal. Isolated word recognition system".
Spoken Language Generation and Understanding
Edited by JC. Simon.

(BOE-80.a) BOE L.J. - ABRY C.
"Caractéristiques individuelles et phonétiques"
Bulletin de l'Institut de Phonétiques de Grenoble
Vol IX (p. 1-16).

(BOE-80.b) BOE L.J. - ABRY C. - CORSI P.
"Les problèmes de normalisation interlocuteurs. Méthodes d'ajustement aux limites"
XIème JEP du Groupe Communication parlée
Mai 1980 - Strasbourg.

(BRIANT-83) BRIANT N. - FLOCON B.
"SYRIL : Système temps réel de reconnaissance de mots isolés multilocuteurs".
Colloque EUSIPCO - Sept. 1983 Erlangen.

(BRIDLE-83) BRIDLE JS. - CHAMBERLAIN RM.
"Automatic labelling of speech using synthesis-by-rule and non linear time alignement"
Elsevier Science Publishers B.V. North Holland
(p. 187-189) 1983.

(BROAD - 74) BROAD DJ. - SHOUP JE.
"Concepts for Acoustic Phonetic Recognition"
Speech Recognition IEEE Symposium 1974
Edited by D. RAJ. REDDY Academi Press 1975.

(CAELEN-81) CAELEN J. - CAELEN G.
"Indices et propriétés dans le projet ARIAL II".
Actes du séminaire "Processus d'encodage et de décodage phonétiques" (p. 128-143) Toulouse sept 1981.

(CAELEN-79) CAELEN J.
"Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique".
Thèse d'Etat - Toulouse 1979.

(CARBONELL-84) CARBONELL N. - FOHR D. - HATON JP. - LONCHAMP F. - PIERREL JM
"An Expert System For the Automatic Reading of French Spectrograms".
IEEE 1984 - ICASSP San Diego.

(CARTON-79) CARTON F.
"Introduction à la phonétique du Français"
BORDAS Etudes 303 2ème édition 1979.

(CHANDON-81) CHANDON J.L. - PINSON S.
"Analyse typologique - Théories et applications"
Edition MASSON 1981.

(CORSI-79) CORSI P.
"Reconnaissance automatique du locuteur"
INPG Thèse de doctorat d'Ingénieur - Grenoble 1979.

(DI MARTINO-84) DI MARTINO J.
"Contribution à la reconnaissance globale de la parole : mots isolés, mots enchaînés".
Thèse de doctorat d'Ingénieur - Univ. de NANCY I - 1984.

(DOURS-81) DOURS D. - FACCA R. - PERENNOU G.
"Une méthode dynamique de segmentation phonémique"
Actes du séminaire "Processus d'encodage et de décodage phonétiques" (p. 169-178) Toulouse sept. 1981.

(FAIRBANKS-40) FAIRBANKS G.
"Recent Experimental Investigations of Vocal Pitch in Speech".
JASA, 11, (p. 457-466).

(FANT-62) FANT C.G.M.
"Descriptive Analysis of the Acoustic Aspects of Speech".
LOGOS volume 5 n° 1 (p. 3-17) 1962.

(FONSALE-84) FONSALE P.
"Simulation informatique d'un système multilocuteur de reconnaissance de parole (mots isolés) sans apprentissage oral - Analyse par traits phonétiques".
INPG Thèse de doctorat d'Ingénieur - Grenoble Janv. 84.

(FURUI-80) FURUI S.
"A Training Procedure For Isolated Word Recognition Systems".
IEEE Trans. on ASSP - vol. 28, n° 2, April 1980.

(GERARD-81) GERARD M. - MERCIER G.
"L'apprentissage des paramètres de reconnaissance phonétique dans un système de reconnaissance de la parole continue"
AFCET Actes 3ème congrès RFIA (p. 641-652) Nancy 81

(GERSTMAN-86) GERSTMAN LJ.
"Classification of Self Normalized Vowels"
IEEE Trans. on audio and Electroacoustics
Vol AU 16, n° 1, (p. 78-80), March 1986.

(GRENIER-80) GRENIER Y.
"Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur"
11ème journées d'Etude sur la Parole - Strasbourg mai 80.

(GROCHOLEWSKI-84) GROCHOLEWSKI S.
"Reconnaissance de mots isolés à l'aide de "spectres temporels"".
AFCET Janvier 1984 Paris.

(GUEGEN-76) GUEGEN C.
"Introduction à l'analyse de la parole"
7ème J.E.P. 1976 Nancy.

(GUPTA-78) GUPTA VN. - BRYAN J.K. - GOWDY JN.
"Speaker independent vowel identification in continuous speech"
IEEE ICASSP (p. 546-548), Tulsa - USA 1978.

(HATON-82) HATON J.P.
"Recherches et développements actuels en reconnaissance automatique de la parole"
4ème journées Francophones d'Informatique
Janvier 1982 - Genève.

(HATON-81) HATON J.P. - SANCHEZ C.
"Méthode synchrone et asynchrone en reconnaissance phonétique de la parole"
Actes du séminaire "Processus d'encodage et de décodage phonétiques" (p. 144-155) Toulouse sept. 81.

(HATON-74) HATON J.P.
"Contribution à l'analyse, la paramétrisation et la reconnaissance automatique de la parole"
Thèse d'Etat, Université de Nancy I, 1974.

(JASCHUL-79) JASCHUL J.
"An Approach to Speaker Normalization for Automatis Speech Recognition"
IEEE ICASSP p. 235-238 - Washington 1979.

(LAZREK-83) LAZREK M.
"Décodage acoustico-phonétique en compréhension automatique de la parole continue"
CRIN Thèse de 3ème cycle Nancy 1983.

(LECORRE-79) LECORRE C. - VIVES R.
"Un programme de cadrage pour l'adaptation au locuteur en reconnaissance automatique de la parole"
3ème congrès AFCET - IRIA
Reconnaissance des Formes et intelligence Artificielle 1979.

(LELIEVRE-81) LELIEVRE A.
"Codage et traitement de certains types de données phonétiques pour une reconnaissance automatique de la parole par la classification"
Thèse 3ème cycle - Rennes 1981.

(LENNIG-83) LENNIG M.
"Automatic alignment of natural speech with a corresponding transcription"
Elsevier Science Publishers B.V. North Holland
(p. 190-192) 1983.

(LE NY-80) LE NY J.F. - DENIS M.
"Identification et compréhension du language naturel : perspectives cognitives"
GALF AFCET 1980 - Extrait de "Syntaxe et sémantique en compréhension de la parole - (édité par HATON J.P. PIERREL J.M. - QUINTON P.).

(LIENARD-77) LIENARD JS.
"Les processus de la communication parlée"
Editions MASSON 1977.

(LIENARD-72) LIENARD JS.
"Analyse synthèse et reconnaissance automatique de la parole"
Thèse d'Etat - Université de Paris VI - 1972.

(LOWERRE-77) LOWERRE BT.
"Dynamic Speaker Adaptation in the HARPY Speech Recognition System"
IEEE ICASSP Hartford - 1977.

(LOBANOV-71) LOBANOV B.M.
"Classification of Russian vowels spoken by different speakers"
JASA 49, (p. 606-608)

(MALMBERG-79) MALMBERG B.
"La phonétique
Presses Universitaires de France 12ème édition 1979.

(MALMBERG-72) MALMBERG B.
"Phonétique Française"
Hermods Malmo Suède
End Edition 1972.

(MARCHAL-80) MARCHAL A.
"Les sons et la parole"
Collection langue et société GUERIN 1980.

(MARI-84) MARI J.F. - HATON J.P.
"Some Experiments in Automatic Recognition of a thousand Word Vocabulary"
IEEE ICASSP - March 1984 - San Diego.

(MARLSEN-WILSON-80) MARLSEN-WILSON - TYLER
"The temporal structure of spoken language understanding"
Cognition 1980, 8, 1-71.

(MARTIN-80) MARTIN P.
"Variations prosodiques inter et intra locuteurs"
XIème JEP du Groupe Communication parlée
28-30 mai 1980 - Strasbourg

(MATSUMOTO-79) MATSUMOTO H. - WAKITA H.
"Frequency Warping For Non uniform talker normalization"
IEEE ICASSP (p. 566-569) Washington 1979.

(MERCIER-78) MERCIER G.
"Evaluation des indices acoustiques utilisés dans l'analyseur phonétique du système KEAL"
9ème journée d'Etudes sur la parole
31 mai- 2 juin 1978 - Lannion.

(MRAYATI-76) MRAYATI M.
"Contribution aux études sur la production de la parole. Modèle électrique du conduit vocal avec perte, conduit nasal et de la source vocale. Etude de leurs intéractions relations entre disposition articulatoire et caractéristiques acoustiques"
Thèse d'Etat, Grenoble, 1976.

(MRAYATI-75) MRAYATI M. - CARRE R.
"Acoustic aspects of French nasal vowels"
JASA 57, 549 (A) 1975.

(NEAREY-77) NEAREY T.
"Phonetic feature systems for vowels"
Doct. Diss. University of Connecticut

(NEEL-83) NEEL F. - ESKEWAZI M. - MARIANI J.J.
"Cadrage automatique pour la constitution de dictionnaires d'entités phonétiques"
Elsevier Science Publishers BV (North Holland)
Speech communication 2 (p. 193-195) 1983.

(PERRIN-81) PERRIN P.
"PERPHO : Programme Experimental de Reconnaissance PHOnétique"
Rapport interne CRIN, juillet 1981.

(PETERSON-52) PETERSON GE - BARNEY HL.
"Control methods used in a study of the vowels"
JASA vol. 24, n° 2 (p. 175-184) March 1952.

(PIERREL-81) PIERREL J.M.
"Etude et mise en oeuvre de contraintes linguistiques en compréhension automatique du discours continu. Application aux languages artificiels : le système Myrtille I - Application aux languages pseudo-naturels : le système Myrtille II"
Thèse de doctorat es-sciences mathématiques,
Nancy, 1981.

(RABINER-80) RABINER L.R. - WILPON J.G.
"A simplified, robust training procedure for speaker trained, isolated words, recognition systems"
JASA 68 (p. 1271-1276) nov. 1980.

(RABINER-79) RABINER L.R. - WILPON J.G.
"Considerations in applying clustering techniques to speaker independant word recognition"
IEEE ICASSP (p. 578-581) Washington 1979.

(RAO-73) RAO C.R.
"Linear statistical inference and its applications"
Edited by WILEY 1973.

(ROSENBERG-82) ROSENBERG AE - RABINER LR - WILPON JG
"Speaker trained recognition of large vocabularies of isolated words"
IEEE ICASSP (p. 2018-2021) Paris 1982

(ROSSI-83) ROSSI M. - NISHINUMA Y. - MERCIER G.
"Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnaissance automatique de la parole"
Elsevier Science Publishers BV (North Holland)
Speech communication 2 (p. 215-217) 1983.

(ROSSI-81) ROSSI M. - NISHINUMA Y. - TREVARAIN O. - MERCIER G.
"Reconnaissance des voyelles par les indices et les traits"
Symposium Franco-soviétique 20-22 - Grenoble oct. 1981.

(ROSSI-77) ROSSI M. - DI CRISTO A.
"Proposition pour un modèle d'analyse de l'intonation"
8ème JEP Aix-en-Provence - 1977.

(SAMBUR-75) SAMBUR MR. - RABINER LR.
"A Speaker independant digit recognition System"
The Bell System Technical Journal (p. 81-102) janv. 75.

(SCHAFER-74) SCHAFER RW. - RABINER LR.
"Parametric representation of speech"
Speech recognition from IEEE Symposium 1974
Edited by D. RAD. REDDY Academic Press 1975.

(SHIRAI-82) SHIRAI K. - KOBAYASHI T.
"Recognition of semi-vowels and consonnants in continuous speech using articulatory parameters"
IEEE ICASSP (p. 2004-2007) Paris 1982.

(SHIRAI-81) SHIRAI K.
"Vowel identification in continuous speech using articulatory parameters"
IEEE ICASSP (p. 1172-1175) Boston 1981.

(SOLI-81) SOLI S.D.
"Second Formants in Fricatives : Acoustic consequences of fricative vowel coarticulation"
JASA (p. 976-984) October 1981.

(STEVENS-71) STEVENS KN.
"Sources of inter and intra speaker variability in the acoustic properties of speech sounds"
7th Int. Cong. Phon. Sciences (p. 206-227) 1971.

(SUGAMURA-83) SUGAMURA N. - SHIKANO K. - FURUI S.
"Isolated word recognition using phoneme - like templates"
IEEE ICASSP 1983 Boston.

(SUNDBERG-80) SUNDBERG J.
"Sons et musiques"
Bibliothèque pour la Science 1980.

(WAGNER-81) WAGNER M.
"Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms".
IEEE ICASSP (p. 1156-1159) Boston 1981

(WAKITA-77) WAKITA H.
"Normalization of vowels by vocal-tract lenght and its application to vowel identification"
IEEE Trans. on A.S.S.P. - vol. 25 - April 1977.

(WEINSTEIN-75) WEINSTEIN CJ. - Cc. CANDLESS SS. - MONDSHEIN LF. - ZUE VW.
"A system for acoustic-phonetic analysis of continuous speech"
IEEE Trans. on A.S.S.P. - vol. 23 (p. 54-67) Feb. 75

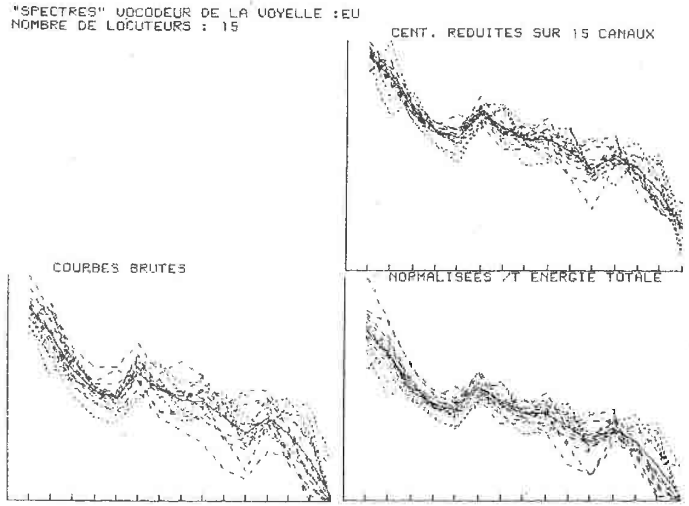
(WILLIAMS-70) WILLIAMS CE. - STEVENS KN. - HECKER MHL.
"Acoustical manifestations of emotional speech"
JASA 47 : 66(A) - 1970.

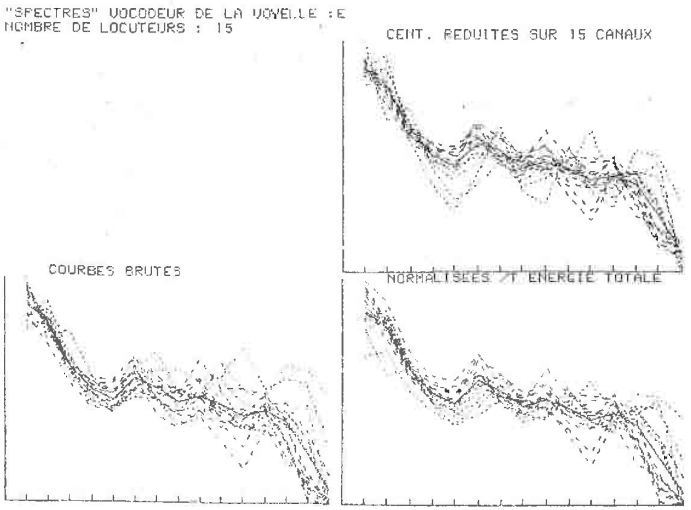
(WILPON-82) WILPON JG. - RABINER LR. - BERGH A.
"Speaker independant isolated word recognition using a 129 word airline vocabulary"
JASA 72(2) (p. 390-396) August 1982.

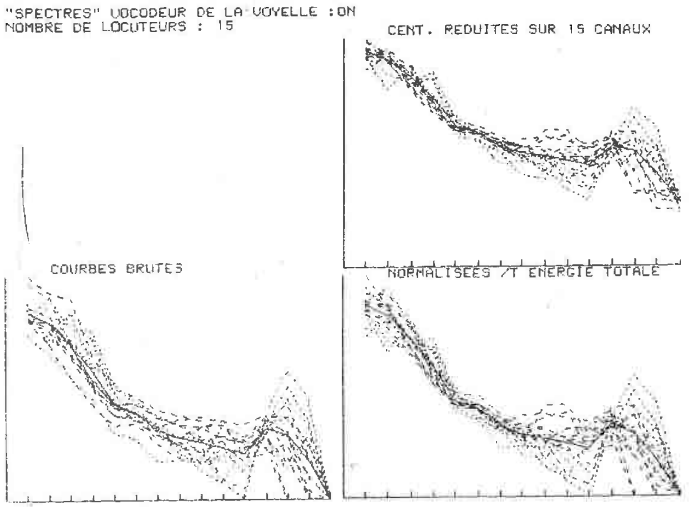
(ZERLING-80) ZERLING JP
"Correlations entre variabilité articulatoire et variabilité acoustique chez deux locuteurs"
XIème journées d'étude sur la parole du groupe communication parlée 28-30 mai 1980 - Strasbourg

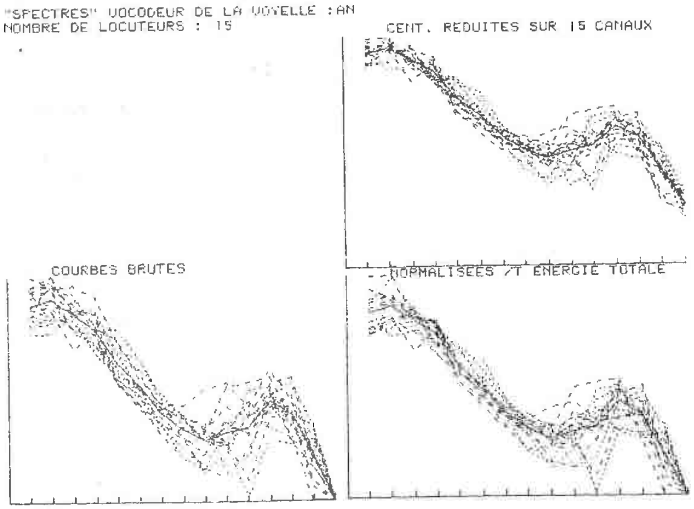
(ZUE-81) ZUE VW. - SCHWARTZ RM.
"Acoustic processing and Phonetic analysis"
From "Trends in Recognition" - Wayne LEA 1981.

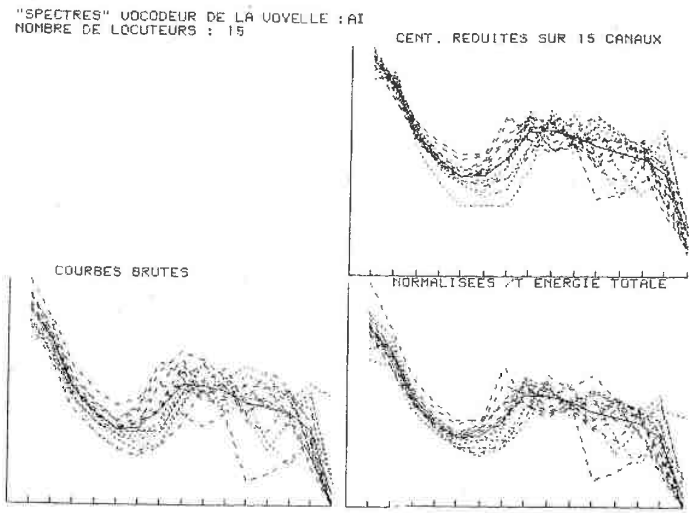
ANNEXE 1 : Graphigues permettant de visualiser les mesures obtenues à l'aide d'un vocoder en portant à l'abscisse i et l'ordonnée y_i l'énergie mesurée sur le canal i .

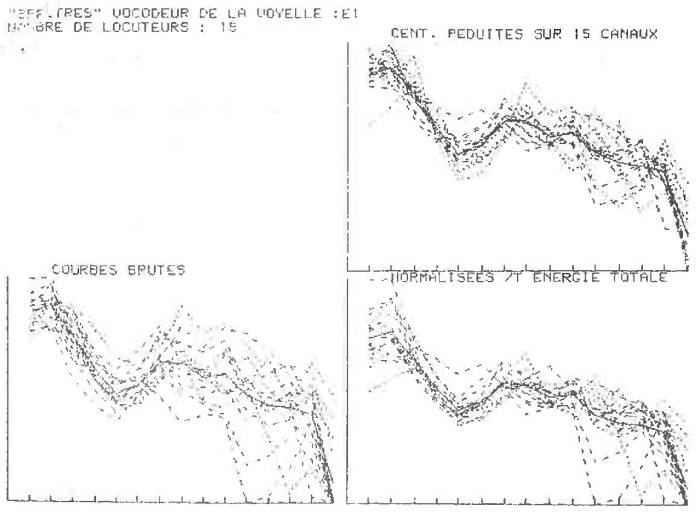

"SPECTRES" VOCODEUR DE LA VOYELLE :EU
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /I ENERGIE TOTALE

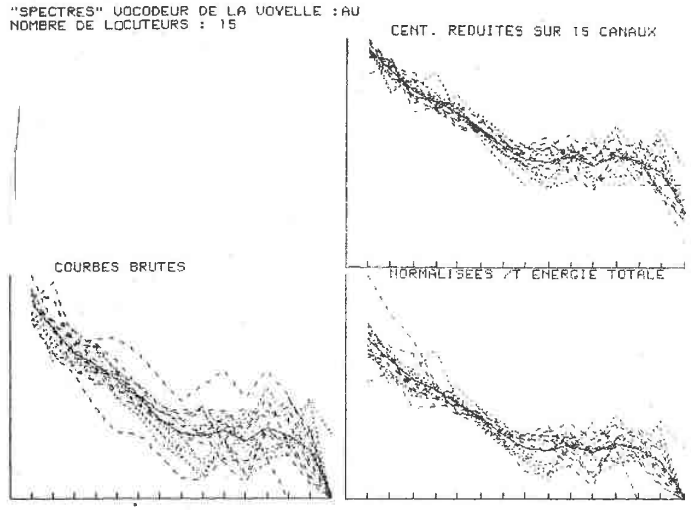

"SPECTRES" VOCODEUR DE LA VOYELLE :E
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /I ENERGIE TOTALE

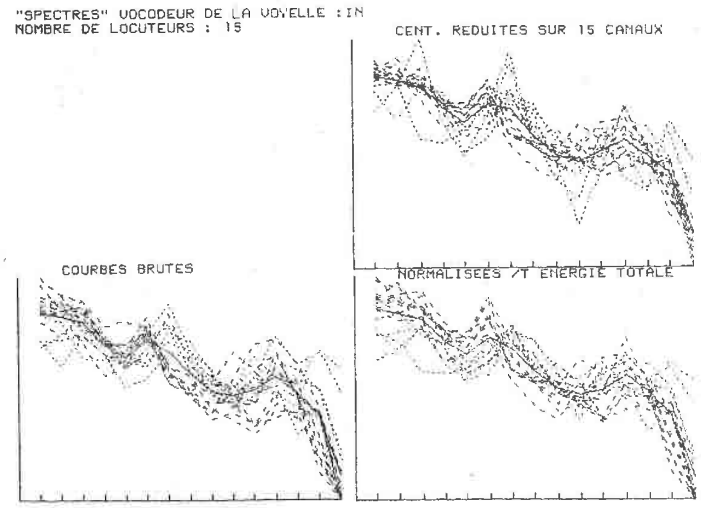

"SPECTRES" VOCODEUR DE LA VOYELLE :ON
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /I ENERGIE TOTALE

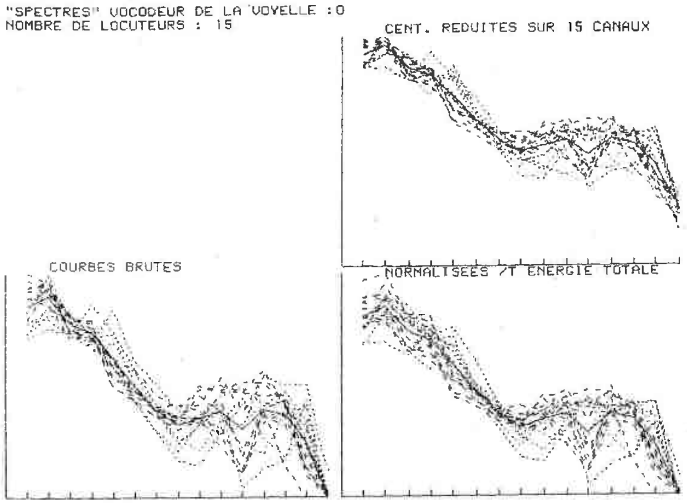

"SPECTRES" VOCODEUR DE LA VOYELLE :AN
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /I ENERGIE TOTALE

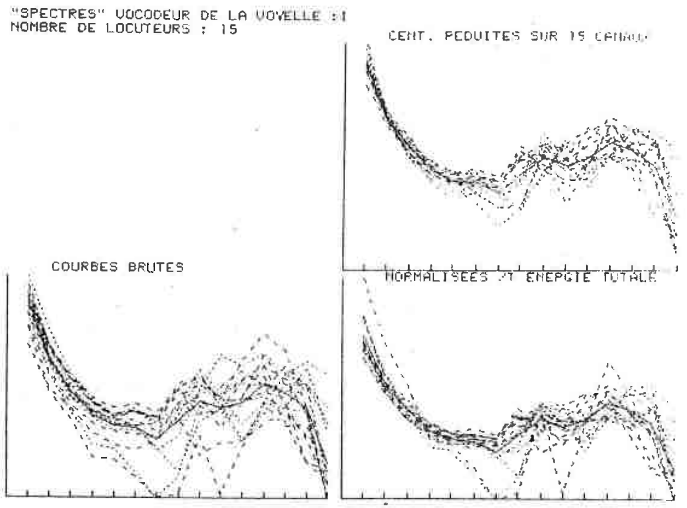

"SPECTRES" VOCODEUR DE LA VOYELLE :AI
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /T ENERGIE TOTALE

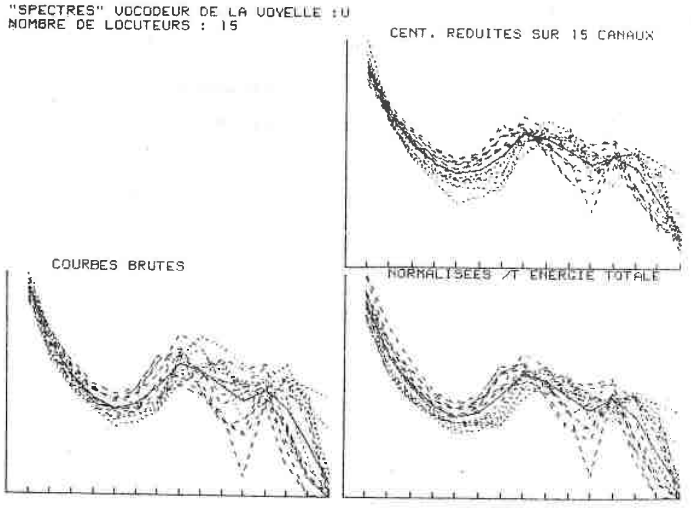

"SPECTRES" VOCODEUR DE LA VOYELLE :E1
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /T ENERGIE TOTALE

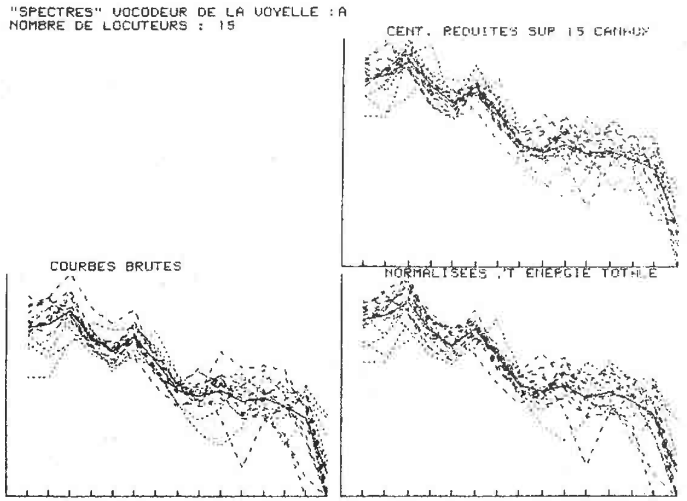

"SPECTRES" VOCODEUR DE LA VOYELLE :AU
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /T ENERGIE TOTALE

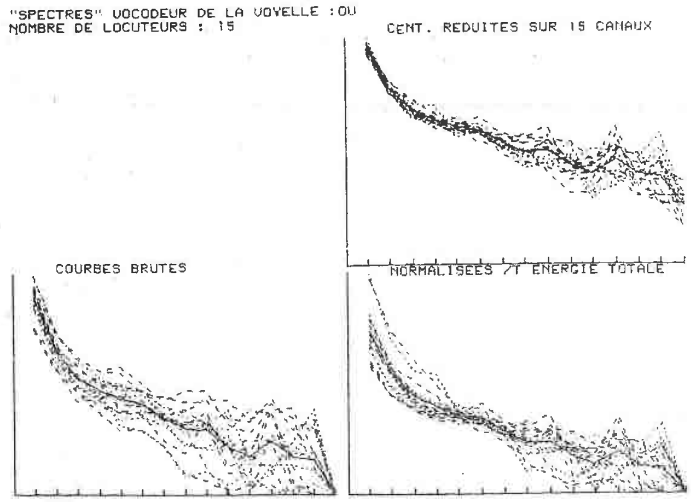

"SPECTRES" VOCODEUR DE LA VOYELLE :IN
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /T ENERGIE TOTALE


"SPECTRES" VOCODEUR DE LA VOYELLE :O
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /T ENERGIE TOTALE


"SPECTRES" VOCODEUR DE LA VOYELLE :I
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /T ENERGIE TOTALE


"SPECTRES" VOCODEUR DE LA VOYELLE :U
NOMBRE DE LOCUTEURS : 15
CENT. REDUITES SUR 15 CANAUX
COURBES BRUTES
NORMALISEES /T ENERGIE TOTALE





ANNEXE 2

Phrases de tests pour la reconnaissance de phonèmes.

Certaines phrases de tests ont été volontairement coupées pour que leur représentation acoustique tienne dans le tableau réservé à cet effet en mémoire centrale.

- BIEN SUR JE CONNAIS
- IL SE GARANTIRA DU FROID
- AVEC CE BON CAPUCHON
- ANNIE S'ENNUIE
- DE MES PARENTS
- LES DEUX CAMIONS
- DEUX FACES
- MAMAN PREND SON VERRE
- DESORMAIS
- QUAND IL PARTIRA
- LES AVIONS TOURNENT
- AU-DESSUS DE LA PLACE
- JE SUIS RESTE SOURD
- A SES CRIS
- LE CHAMEAU EST LOIN
- DE SON ABRI
- CE BONBON CONTENAIT
- TROP DE SUCRE
- LE RENARD
- JE M'ACCOUDAIS AU MURET
- CE PETIT CANARD.

NOM DE L'ETUDIANT : PISTER Christine

NATURE DE LA THESE : Doctorat 3ème cycle en Informatique

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le 22 MAI 1984 n° 916

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

R. MAINARD

RESUME

Dans le cadre d'un système de reconnaissance automatique de mots isolés par approche analytique pour un vocabulaire de grande taille (400 mots et plus) nous avons envisagé une adaptation automatique du système au locuteur. L'adaptation se fait par l'apprentissage automatique des formes de références du locuteur, ainsi que par l'ajustement automatique des paramètres du système.

Un algorithme de cadrage de segments phonétiques décodés à l'aide de traits acoustico-phonétique peu dépendants du locuteur est à la base de cet apprentissage.

Une analyse des voyelles de 15 locuteurs relevant de la statistique descriptive et de l'interprétation statistique a été entreprise en vue d'élaborer des procédures de normalisation et de génération automatique des formes de références des voyelles d'un locuteur.

MOTS-CLES

- Variabilité de la parole
- Système multilocuteur
- Adaptation au locuteur
- Apprentissage automatique.