Université de Nancy I

U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

REALISATION D'UN SYSTEME A BASE DE PROTOTYPES POUR LE CONTROLE DU DECODAGE ACOUSTICO-PHONETIQUE DE LA PAROLE

THESE

présentée et soutenue publiquement le 6 Janvier 1986

A L'UNIVERSITE DE NANCY I

pour l'obtention du titre de

DOCTEUR D'UNIVERSITE EN INFORMATIQUE

par

Jean-Paul DAMESTOY

Composition du Jury

Président:	Jean-Marie PIERREL
Rapporteurs:	Daniel COULON
	Jean-Pierre TUBACH
Examinateurs:	Alain BONNET
	Jean-Paul HATON
Invité	François LONCHAMP

Université de Nancy I — U.E.R. Sciences Mathématiques

Centre de Recherche en Informatique de Nancy

REALISATION D'UN SYSTEME A BASE DE PROTOTYPES POUR LE CONTROLE DU DECODAGE ACOUSTICO-PHONETIQUE DE LA PAROLE

THESE

présentée et soutenue publiquement le 6 Janvier 1986

A L'UNIVERSITE DE NANCY I

pour l'obtention du titre de
DOCTEUR D'UNIVERSITE EN INFORMATIQUE

par
Jean-Paul DAMESTOY

Composition du Jury

Président:	Jean-Marie PIERREL
Rapporteurs:	Daniel COULON
	Jean-Pierre TUBACH
Examinateurs:	Alain BONNET
	Jean-Paul HATON
Invité	François LONCHAMP

Je tiens à exprimer ma profonde gratitude à Mr le Professeur Haton qui a bien voulu m'accueillir dans son laboratoire. Je lui suis infiniment redevable d'avoir su inspirer et diriger ce travail.

Que Mr le professeur Pierrel trouve ici l'expression de ma reconnaissance pour le bienveillant intérêt qu'il a porté à mes recherches et l'appui amical qu'il m'a toujours manifesté.

Je remercie Messieurs les Professeurs Coulon et Tubach qui ont bien voulu juger ce travail malgré l'importance des nombreuses tâches qu'ils assument.

Tous mes remerciement vont également à A. Bonnet, professeur à l'E.N.S.T., et à F. Lonchamp, maitre de conférences à l'université de Nancy 2, pour avoir accepté de participer à ce jury.

Merci enfin à tous mes camarades du CRIN pour l'aide précieuse et amicale qu'ils m'ont tous apportée lors de la réalisation de ce travail.

SOMMAIRE

CHAPITRE 6
L'implantation du système: LFRAP

CHAPITRE 7
Résultats expérimentaux

INTRODUCTION

La parole est un moyen de communication privilégié pour l'homme, et l'invention du téléphone, ainsi que le développement de l'électronique ont donné à ce médium une nouvelle dimension. A l'aide de ces nouvelles techniques, il est devenu possible d'enregistrer la parole, de la transmettre et de la diffuser à distance grâce aux réseaux téléphonique, radiophonique et télévisuel. Mais jusqu'à ces dernières années, le dialogue avec un calculateur est resté rebelle à la parole: il passe par l'utilisation d'un code écrit, et clavier, imprimante ou écran restent encore les seuls moyens de dialogue possibles avec la machine.

Depuis peu, le progrès des techniques de traitement du signal de la parole et l'apport des techniques de l'intelligence artificielle permettent d'envisager dans un futur proche un véritable dialogue oral entre l'homme et la machine.

Le décodage acoustico-phonétique est l'étape du traitement automatique de la parole qui cherche à transformer le signal acoustique en une suite discrète de symboles phonétiques. Le décodage joue un rôle fondamental dans un tel système, car placé en frontal entre le signal vocal et les niveaux linguistiques du traitement, il détermine en bonne partie la performance globale du système de compréhension. La difficulté principale pour les systèmes développés est de gérer la grande variabilité intra et inter locuteurs de la parole qui se traduit par des variations acoustiques, phonétiques et phonologiques du message vocal, et le décodage acoustico-phonétique constitue actuellement une des principales pierres d'achoppement en reconnaissance de la parole continue. Les causes sont nombreuses, et résident pour la plupart dans une mauvaise description de l'univers phonétique.

C'est la raison pour laquelle nous proposons un nouveau formalisme pour représenter et utiliser le savoir-faire des phonéticiens: les frames. Le système décrit ici représente l'actuel niveau phonétique des systèmes de reconnaissance de la parole développés dans notre laboratoire.

Notre système comporte sept chapitres:

Dans les premier et second chapitres sont présentés le système phonatoire ainsi que les caractéristiques phonétiques utiles à la reconnaissance des sons émis par l'homme. Synthèse du cours de phonétique assuré par F. Lonchamp à l'univeristé de Nancy, ils permettent 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 troisième chapitre présente le cadre générale dans lequel s'insère notre recherche, le décodage acoustico-phonétique, et détaille les différents niveaux d'analyse sous-jacents: l'analyse acoustique, la segmentation du continuum vocal, l'identification des segments et l'apprentissage.

Les différents outil mathématiques qui permettent de traiter le problème phonétique font l'objet du quatrième chapitre. Le point sur les différentes techniques actuelles en les divisant en deux grandes classes: les techniques relevant de la reconnaissance des formes et les techniques propres à l'intelligence artificielle.

Le cinquième chapitre introduit le concept de frames pour décrire l'univers phonétique. L'interét essentiel de cette modélisation qui met en oeuvre une méthode ascendante d'analyse, approche classique en décodage acoustico-phonétique, réside dans le fait qu'elle autorise les interactions indispensables entre les processus de segmentation et d'identification phonétique, grâce aux relations qui existent entre les notions de règle, de prototype et de procédure. Ces principes sont mis en application dans les système de décodage acoustico-phonétique LFRAP, dont le processus d'analyse est explicité dans le sixième chapitre.

Le septième chapitre termine sur quelques résultats expérimentaux.

CHAPITRE 1
LA PAROLE ET LE SYSTEME PHONATOIRE

1.1 Introduction

Comprendre le mécanisme de production de la parole est un aspect de l'étude de la parole qui a une grande importance. En effet, différents algorithmes de paramétrisation de l'onde vocale sont obtenus a partir de modèles du conduit vocal plus ou moins fidèles. De plus, c'est l'étude du système phonatoire qui va nous permettre d'identifier et de caractériser les grandes classes de sons élémentaires, et par là-même d'expliquer les variations de ces sons élémentaires dans les contextes intra et inter-locuteurs.

1.2 Le système de production de la parole

La figure 1 est une représentation schématique du système phonatoire. Le conduit vocal se compose essentiellement de la cavité du pharynx, de la cavité buccale et des cavités nasales, et c'est de l'excitation de ce système par l'ensemble poumons-cordes vocales que résulte la production de la parole.

1.2.1 Les sources d'excitation du conduit vocal

L'énergie nécessaire pour produire un son provient de la contraction de la cage thoracique qui augmente la pression dans les poumons. L'air est alors chassé en direction du larynx qui constitue l'entrée du conduit vocal. Sous la pression de l'air, les cordes vocales accolées oscillent un peu à la manière d'une anche double de hautbois :

- si les cordes sont tendues, elles se mettent à vibrer laissant passer l'air dans le larynx par saccades ou par impulsions
- si les cordes sont relachées, l'air passe librement à travers la glotte (orifice crée par les cordes vocales)

La fréquence de ce phénomène de relaxation est essentiellement déterminée par la tension, la longueur, l'épaisseur des cordes vocales ainsi que par la pression

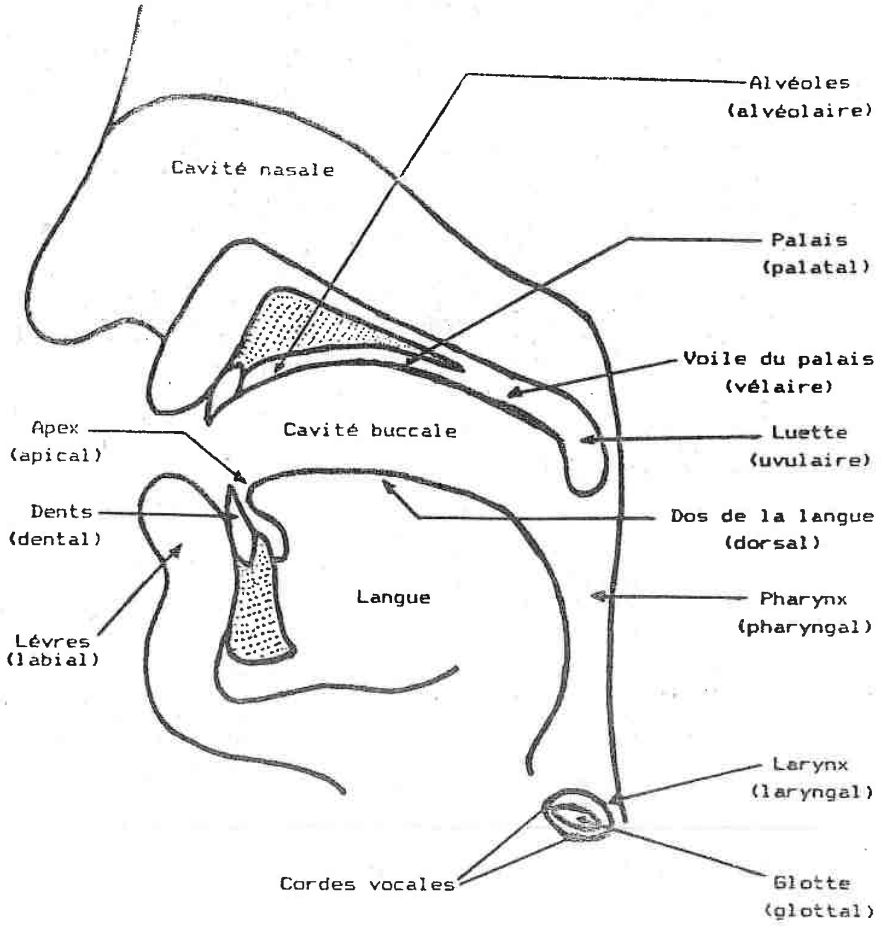


Figure 1
Schéma de l'appareil phonatoire
d'après [MELO-82]

sub-glottique. A l'entrée du conduit vocal, le débit d'air est donc modulé périodiquement. La fréquence de ces vibrations désignée sous les noms de fréquence fondamentale ou de pitch, caractérise la hauteur de la voix et ses variations contribuent à la perception de la mélodie de la phrase. Cette fréquence fondamentale est en moyenne de 125 Hz pour un homme et de 250 Hz pour une femme.

La présence ou l'absence de vibrations des cordes vocales permet de différencier les sons en deux grandes classes :
- les sons dits sonores ou voisés, qui résultent d'une vibration périodique des cordes vocales
- les sons dits sourds ou non voisés pour lesquels l'air passe librement à travers la glotte

Dans les deux cas, l'onde de pression acoustique créee est modifiée par les phénomènes de résonance dans les cavités du pharynx, de la bouche et du nez avant d'être émise à l'air libre. Ce sont les formes et la position des différents organes de la cavité buccale qui vont nous permettre de distinguer deux autres catégories de sons : les sons fricatifs et les sons occlusifs. Un son est fricatif lorsque l'air s'échappe du conduit vocal au travers d'une constriction de ce dernier créee par les dents, les lèvres ou la langue et le palais. Par contre, on a une occlusive ou plosive lorsque l'air sort du système phonatoire de façon brutale après avoir vu sa pression croitre derrière un obstacle : lèvres, dents, etc...

On peut donc considérer qu'il existe trois modes d'excitation du conduit vocal, la source sonore ou voisement, la friction et l'occlusion, qui induisent les catégories de sons suivants :
- les sons voisés, non fricatifs et non plosifs
- les fricatives sourdes
- les fricatives sonores
- les plosives sourdes
- les plosives sonores
- les silences qui correspondent aux sons sourds, non fricatifs et non plosifs

1.2.2 L'articulation des sons - les formants

L'articulation, qui définit la forme du conduit vocal que l'on peut considérer comme un conduit de section et de longueur variables, limité à son origine par les cordes vocales et à son extremité par les lèvres (figure 1), joue un role important dans la production de la parole. Le débit d'air issu de la "source vocale" est modifié lors de la traversée du conduit vocal, constitué pour l'essentiel de deux cavités couplées, le pharynx et la bouche. Les positions de la machoire et de la langue déterminent des cavités jouant le role de caisses de résonance qui vont renforcer certaines régions du spectre acoustique. Les maxima de la courbe de réponse en fréquence du conduit vocal sont appelés "formants". La fréquence du premier formant peut varier de 200 Hz à 750 Hz, celle du second formant de 900 Hz à 2500 Hz. Il existe des formants d'ordre supérieur pouvant aller jusqu'à 5000 Hz; l'ensemble des formants contribue en particulier à caractériser le timbre de la voix.

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

Les phonéticiens distinguent l'endroit où le conduit vocal est le plus resserré qui est appelé "lieu d'articulation". Il peut être au niveau des lèvres, des dents, des alvéoles, du palais ou du velum. Il en est de même pour l'organe qui conditionne ce resserrement et qui est appelé "organe d'articulation". Ce dernier peut être soit la lèvre inférieure, soit la la langue dans laquelle on distingue la pointe, le dos et le dessous. Ces points d'articulation seront étudiés en détail dans le chapitre suivant.

1.3 Modélisation numérique du système de production de la parole

A partir du modèle fonctionnel du système phonatoire que nous venons de décrire, il est possible de définir un modèle numérique. La figure 2 visualise un tel modèle.

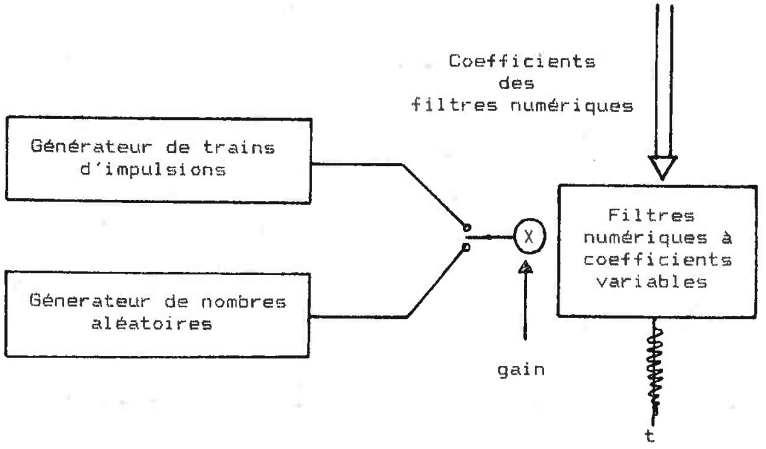


Figure 2
Modèle du système phonatoire d'après SCHAFFER
[SCHA-72]

L'idée de base qui a été émise pour établir celui-ci est l'indépendance des sources d'excitation et du conduit vocal. C'est cette hypothèse qui permet de définir la notion de transfert du conduit vocal.

Celle-ci est valable dans la majorité des cas, mais malheureusement certains sons transitoires comme les plosives sourdes rendent le modèle incomplet, du fait de la nature non périodique de ces sons.

Les différentes sources du conduit vocal ont été modélisées par des générateurs de trains d'impulsions :
- impulsions périodiques pour les sons voisés, la fréquence de récurrence de ces impulsions est commandée par la fréquence fondamentale, ce qui permet de reproduire la hauteur de la voix.
- impulsions aléatoires pour les sons non voisés.

Ces différentes impulsions sont envoyées à l'entrée d'un banc de filtres numériques à coefficients variant dans le temps pour simuler la déformation spatiale du conduit vocal lors de l'élocution. Un dernier paramètre agissant sur l'amplitude des signaux d'excitation permet de controler le gain du système.

Le modèle du système de phonation que nous venons de décrire permet de comprendre et de simuler les mécanismes de production de la parole. C'est à partir de ce modèle qu'ont été élaborés les principaux systèmes d'analyse de la parole, les vocodeurs à canaux, à prédiction linéaire ou à formants, et les algorithmes de paramétrisation de l'onde vocale associés [CARR-79],[DIMA-84].

CHAPITRE 2
ETUDE PHONETIQUE DU FRANCAIS

2.1 Introduction de la notion de phonème

Des travaux menés sur la manière dont est organisée et structurée une langue ont mis en évidence que derrière les concepts de production et de perception de la parole, il existe une suite de segments élémentaires et discrets, qui sont concaténés dans l'espace temporel. Ces segments, appelés "phonèmes", sont supposés posséder des caractéristiques articulatoires et acoustiques uniques. Jakobson, Fant et Halle [JAKO-63a], ont montré que les phonèmes pouvaient etre caractérisés par un ensemble d'attributs invariant, que l'on appelle "fonction distinctive" ou "traits distinctifs". Il existe une relation étroite entre la fonction distinctive et le geste articulatoire qui a produit le son associé au phonème; et au niveau phonémique, la structure linguistique d'une phrase pourrait être représentée par une matrice à deux dimensions dont les colonnes représenteraient les phonèmes, les lignes les traits distinctifs et chaque élément de la matrice indiquerait alors la présence ou l'absence d'un trait pour un phonème donné.

Cette représentation, bien que séduisante, n'est pas réaliste. Pendant la production de la parole, le contenu linguistique de la matrice des traits est transformé en commandes neuro-musculaires qui sont exécutées par les articulateurs (lèvres, langue, etc...). Une exécution qui est peu fidèle, car bien que ces commandes soient dicrétisées dans le temps, le travail des articulateurs et le signal acoustique qui en résulte sont quant à eux continus dans le temps. Cette continuité provient de la structure musculaire, donc élastique, des articulateurs, qui introduit des "transitions" entre les différents lieux d'articulation des phonèmes émis. Il s'ensuit une altération des traits distinctifs de chaque phonème par l'apport d'information phonémique des transitions avec les phonèmes contigus. En d'autres termes, un phonème verra sa réalisation varier énormément en fonction du contexte dans le discours continu.

Ce n'est pas le seul facteur de variabilité de la réalisation d'un phonème; on peut citer entre autres l'état général du locuteur et le locuteur lui-même, mais nous étudierons de façon plus détaillée ces problèmes de variabilité dans les chapitres suivants.

Il se dégage de cette introduction que les phonèmes constituent en fait une abstraction de la réalité, car un phonème n'a pas de représentation unique. Et c'est cette pluralité de formes possibles que peut prendre un phonème qui rend la tache de décodage acoustico-phonétique extrémement ardue.

2.2 Les phonèmes du français

On détermine couramment les phonèmes d'une langue suivant la méthode des paires minimales, qui consiste à rechercher la plus petite variation dans un segment de niveau supérieur, le mot en général, qui le change de catégorie. Ainsi les mots "pire", "pur", "père", "peur", "port", "pour" et "par" qui diffèrent par le second phonème, constituent une paire minimale. En opposant les différentes paires minimales ainsi constituées, on établit de proche en proche l'inventaire complet des phonèmes d'une langue.

Cette liste ne doit cependant pas prendre en compte les variantes de certains phonèmes ou allophones, qui sont imputables à des phénomènes contextuels ou à des particularismes régionaux ou individuels (prononciation différente du phonème /R/ par les Parisiens et les bourguignons par exemple). Par définition, le nombre de phonèmes est limité, mais les variantes ou réalisations concrètes sont théoriquement illimitées.

Le système phonétique du français comporte environ 36 phonèmes qui sont présentés dans le tableau 3 que nous avons emprunté à [CART-74]. Pour chaque phonème y figurant, sont donnés le mot-clef dans lequel le phonème apparait et la classe phonétique à laquelle il appartient.

Phonème	Mot-clef	Classe	Phonème	Mot-clef	Classe
ɑ	patte	Voyelles orales	p	passe	Plosives
ɑ	pâte		t	toux	
i	riz		k	cou	
y	rue		b	basse	
ɔ	port		d	doux	
o	pot		g	goût	
ə	le				
ɛ	raie		m	masse	Nasales
e	ré		n	nous	
ø	peu		ɲ	signe	
œ	peur				
u	roue		f	fer	Fricatives
			s	assis	
ɑ̃	blanc	Voyelles nasales	ʃ	chou	
ɔ̃	bon		v	verre	
ɛ̃	vin		z	Asie	
œ̃	un		ʒ	joue	
j	hier	Semi-consonnes	l	la	Liquides
ɥ	huit		R	rat	
w	oui				

Tableau 3
Les phonèmes du français d'après CARTON [CART-74]

2.3 Les voyelles

Du point de vue articulatoire, les voyelles sont étudiées à partir de l'examen de films radiocinématographiques de l'appareil phonatoire. Elles se caractérisent par le maintien d'une même position de la partie de la langue qui produit l'aperture vocalique. Aucun changement n'intervient durant ce laps de temps, entre la langue et la voute palatale au lieu d'articulation. Par contre, les autres organes continuent leurs mouvements. La figure 4 présente les positions stables de la langue, des machoires et des lèvres pendant la phase d'articulation des voyelles /i/, /e/, /ɛ/, /a/, /ɑ/, /o/, /ɔ/ et /u/, ainsi que la représentation schématique que l'on peut déduire en disposant sur un axe le degré d'ouverture et sur l'autre le caractère antérieur ou postérieur de l'articulation.

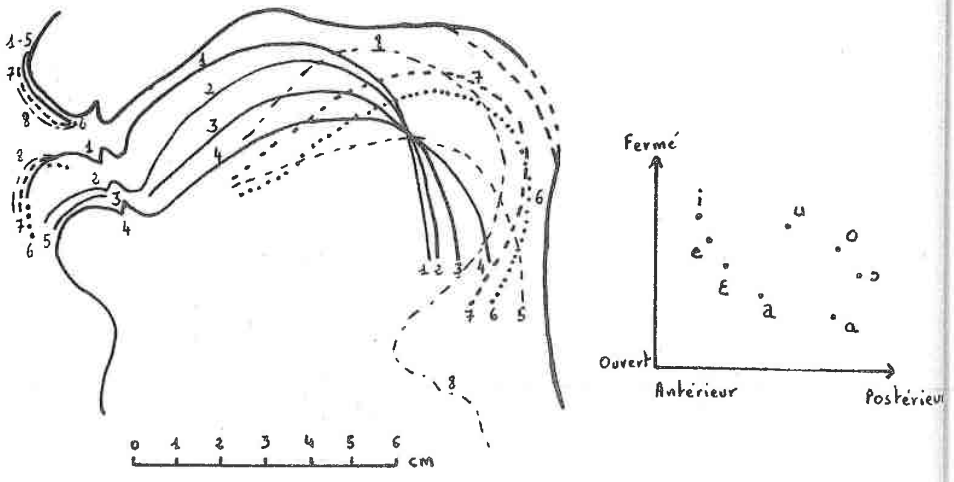


Figure 4
Classification articulatoire des voyelles orales
i(1),e(2),ɛ(3),a(4),ɑ(5),ɔ(6),o(7) et u(8)

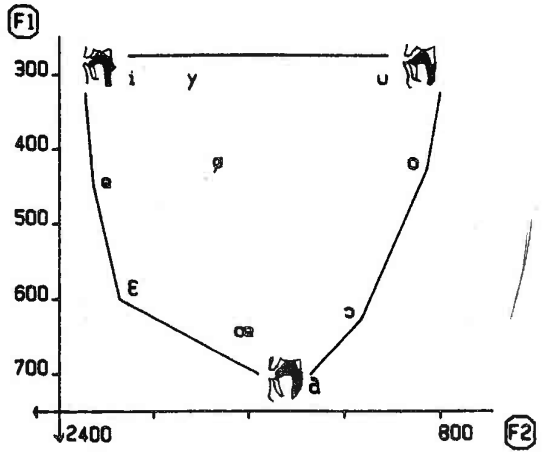
Cette caractérisation articulatoire permet d'établir une première classification des voyelles.

Mais c'est sur le plan acoustique que sont vraiment caractérisées les voyelles. Une analyse spectrale de la parole, que l'on peut établir à l'aide d'un spectrographe, distingue les voyelles des autres phonèmes par la présence d'énergie substancielle dans les régions de basse et moyenne fréquence, qui caractérisent principalement les valeurs des trois premières fréquences formantiques appelées F1, F2 et F3. Lors de la production d'une voyelle le conduit vocal se déforme très peu, donc les fréquences de formants sont stables et spécifiques de la voyelle émise (figure 5). Ce critère distinctif a permis d'introduire les classifications acoustiques des voyelles dites "Triangle vocalique du français" et du "Chapeau chinois" présentées sur la figure 6, qui sont très utilisées en décodage acoustico-phonétique de la parole continue.

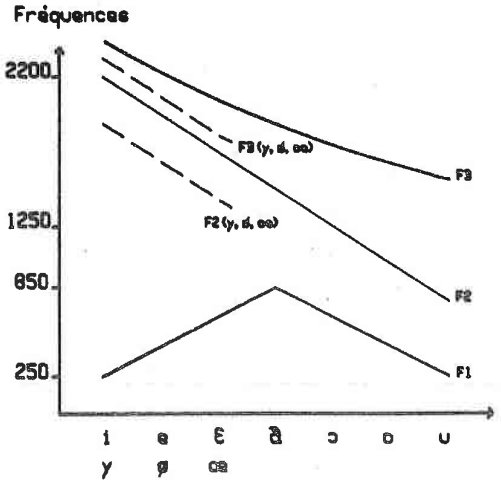
On remarquera que l'étude des formants des voyelles nécessite non pas l'observation de fréquences précises mais de plages fréquentielles susceptibles de se recouvrir partiellement. Ces variations sont imputables à plusieurs facteurs :

- la corrélation qui existe entre la fréquence d'un formant et le fondamental
- l'instabilité de l'articulation qui émet la voyelle : pour un même individu et à une même fréquence fondamentale la dispersion des formants est déjà importante du fait que l'articulation n'est jamais répétée de façon identique.
- la fonction de transfert qui dépend des dimensions du conduit vocal et varie donc d'un individu à l'autre

Ces considérations acoustiques parfaitement adaptées à l'étude de voyelles orales semblent plus difficiles à caractériser les voyelles nasales. Pour la production des voyelles nasales, le voile du palais s'abaisse et permet à l'air de passer simultanément dans les conduits buccal et nasal. Ce couplage acoustique a pour effet d'affaiblir les formants à fréquence élevée, et l'ouverture du vélum du fait de sa lenteur par rapport aux autres mouvements, introduit une certaine instabilité dans le son produit par rapport à la production des voyelles orales. Il s'ensuit que les quatre voyelles



TRIANGLE VOCALIQUE DU FRANCAIS



CHAPEAU CHINOIS

Figure 6 : Classifications acoustiques des voyelles

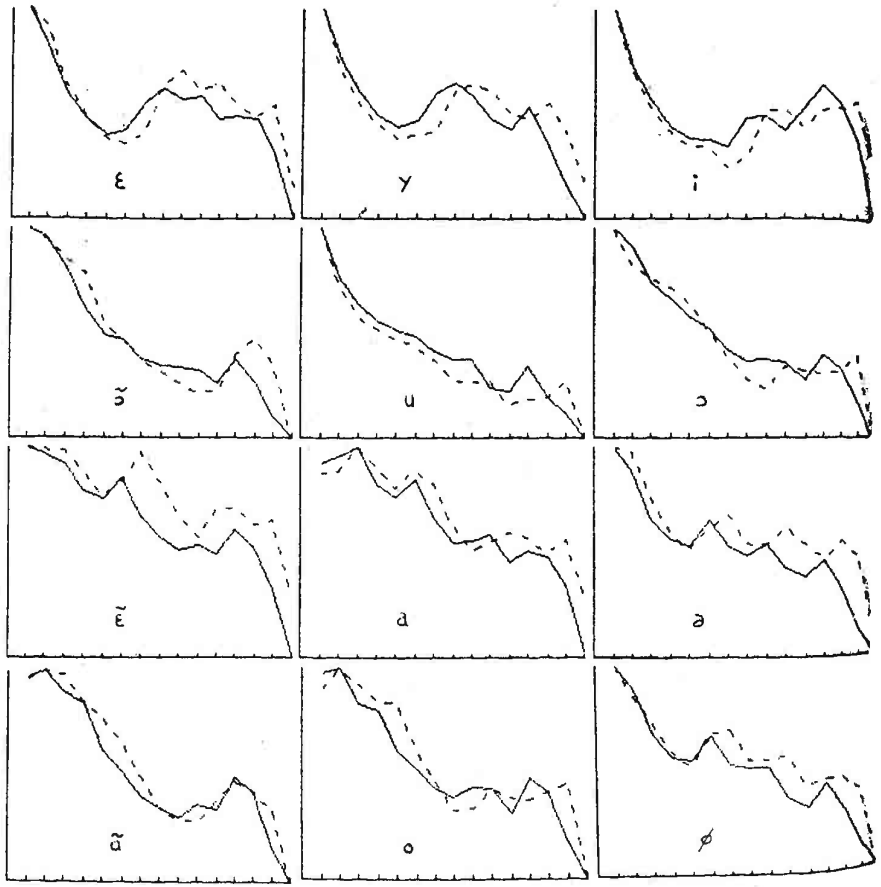


Figure 5

Fréquences formantiques de certaines voyelles

[PIST-84]

—— Hommes --- Femmes

nasales du français /$\tilde{\varepsilon}$/, /$\tilde{\text{œ}}$/, /$\tilde{\text{ɑ}}$/, et /$\tilde{\text{ɔ}}$/, dont les positions articulatoires sont proches des voyelles orales /ɛ/, /œ/,/ɑ/ et /ɔ/ respectivement, sont difficiles à caractériser, l'apparition de pôles nasals dans le spectre compliquant de façon sensible la recherche et l'étude des formants (figure 7).

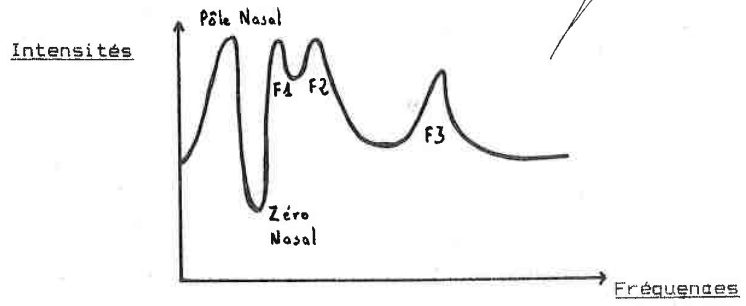


Figure 7
Spectre de la voyelle nasale $\tilde{\text{ɑ}}$

2.4 Les consonnes

2.4.1 Les fricatives ou constrites

Lorsque les cordes vocales sont ouvertes, le son est produit au niveau d'un resserrement du conduit vocal: /f/,/s/,/ʃ/. Ces sons ne sont pas périodiques et sont riches en hautes fréquences. La présence de bruit et de turbulences est caractéristique du spectre des fricatives qui est variable suivant le contexte phonétique. De façon générale, l'amplitude maximale du signal est toujours plus faible que dans la voyelle et décroissante dans l'ordre /ʃ,s,f/. De ce fait, la détection de la fricative /f/ en discours continu, qui est caractérisée par une faible énergie, est fortement liée à la valeur du seuil de détection parole - non parole, qui s'il est élevé assimile souvent cette fricative au silence.

Lorsque les cordes vocales vibrent et que la position de la constriction est la même que celle correspondant à la production de /f/, /s/, /ʃ/, il y a création simultanée de bruit et de voisement pour émettre les sons /v/, /z/ et /ʒ/, que l'on appelle fricatives voisées par opposition aux phonèmes /f/, /s/ et /ʃ/ classés sous le terme fricatives sourdes.

2.4.2 Les plosives ou occlusives

Les plosives se caractérisent acoustiquement par une période plus ou moins importante de silence, aussitôt suivie par une brusque explosion de bruit de friction ou burst. Ces phénomènes correspondent articulatoirement à une occlusion du conduit vocal (silence) et au relachement brusque de ce dernier qui produit l'explosion de la plosive. Selon que les cordes vocales vibrent ou pas, la plosive sera voisée (/b/, /d/, /g/) ou sourde (/p/, /t/, /k/).

Les phénomènes contextuels ont une grande influence sur la forme du spectre engendré, et l'étude des transitions des formants des voyelles adjacentes est un critère tout aussi important que les caractéristiques fréquentielles du burst dans l'analyse spectrale d'une plosive [ZUE-76]. Ainsi l'étude des spectres dans les transitions nous montre que:

- /p/ est caractérisé par un spectre ayant des basses fréquences importantes bien plus rapidement que /t/ et /k/.
- /t/ est caractérisé par un spectre plat et situé assez haut en fréquence par opposition avec /k/ qui présente un maximum unique au-dessus du second formant.
- les plosives voisées /b/, /d/, /g/ présentent un formant très bas.

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

2.4.3 Les sonnantes

2.4.3.1 Les nasales

Les nasales, qui sont toujours associées à une ou plusieurs voyelles, résultent de l'obturation du conduit bucal par les lèvres /m/ ou la langue /n/ et /ɲ/, et de l'ouverture du vélum qui ajoute le couplage des cavités nasales. Dans de nombreux cas, la présence d'une nasale est indiquée par l'importante nasalisation des voyelles adjacentes, et se caractérise par une brusque variation dans la continuité des transitions vocaliques à ses limites.

Acoustiquement, les nasales se distinguent des autres phonèmes par un premier formant très bas en fréquence (autour de 300 Hz) et la quasi-stationarité des formants pendant la durée de la nasale.

2.4.3.2 Les liquides

Comme les nasales, les liquides sont toujours associées à une ou plusieurs voyelles. Mais contrairement à ces mêmes nasales, elles ne présentent pas de stabilité acoustique sur le plan spectral car la coarticulation intervient toujours de façon essentielle dans la réalisation de /l/ et /R/ (figures 8 et 9). On distinguera donc les liquides en étudiant non pas les les valeurs et positions des formants, mais les transitions des formants avec les voyelles adjacentes qui sont en général atténuées et plus étendues en temps en comparaison avec celles des autres consonnes.

2.5 Classification phonétique

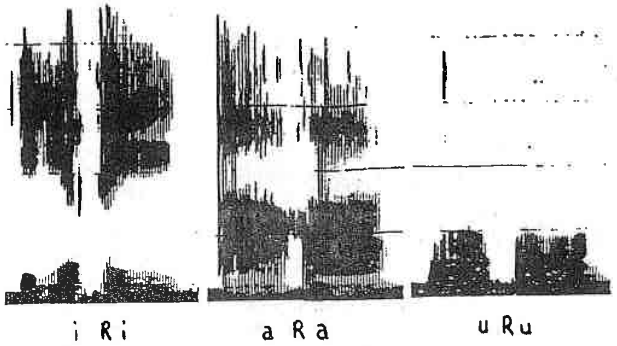


Figure 8 : la liquide /R/ en contexte (*)

Figure 9 : la liquide /l/ en contexte (*)

* Spectrogrammes realises par F. Lonchamp

2.5.1 Classification articulatoire

C'est la classification phonétique la plus utilisée pour caractériser les phonèmes. Elle repose essentiellement sur deux critères: le mode articulatoire et le lieu d'articulation ([MALB-72], [CART-74]). Le mode articulatoire est lié aux diverses sources d'excitation du conduit vocal, qui peuvent être utilisées simultanément à des degrés divers. Il peut être:

- sonore : vibration des cordes vocales
- sourd : absence de vibration des cordes vocales
- nasal : participation du conduit nasal à la production du phonème
- occlusif : fermeture partielle ou totale du conduit vocal suivi d'un relachement brusque
- fricatif : rétrécissement du passage de l'air en un point du conduit vocal
- latéral : passage de l'air sur les parois latérales du conduit vocal obstrué en son centre par la langue
- vibrant : la constriction du conduit vocal n'est pas permanente, et varie sous l'action de la pression de l'air expiré et de la tension musculaire qui ramène ensuite l'articulateur dans sa position initiale.

Le lieu d'articulation désigne la zone du conduit vocal qui participe à la formation du son. Il est désigné par les noms suivants:

- labial : niveau des lèvres
- dental : niveau des dents
- alvéolaire : niveau des alvéoles
- palatal : niveau du palais
- vélaire : niveau du voile du palais
- uvulaire : niveau de la luette
- apical : niveau de la langue

Cette classification "mode articulatoire-lieu d'articulation" n'est utilisée en fait que pour les consonnes, les critères mentionnés ne s'appliquant pas aux voyelles (ref. paragraphe 2.3). On préfère utiliser le caractère antérieur ou postérieur de l'articulation, son degré d'ouverture et la position des lèvres (arrondie ou écartée) pour caractériser les voyelles.

La figure 10 propose une classification articulatoire du sytème phonétique du français selon CARTON [CART-74]. Les indices renvoient au schéma de

VOYELLES

	Antérieures		Postérieures	
	Ecartees	Arrondies	Ecartees	Arrondies
Orales très fermées	i riz	y rue	********	u roue
Orales fermées	e ré	ø peu	********	o pot
Orale moyenne	********	ə le	********	********
Orales ouvertes	ɛ raie	œ peur	********	ɔ port
Orales très ouvertes	à patte		********	ɑ pâte
Nasales	ɛ̃ vin	œ̃ un	ɑ̃ blanc	ɔ̃ bon
Semi-voyelles	j hier	ɥ huit	********	w oui

CONSONNES

	Occlusives			
	Bi-labiales A1	Apico-dentales B2	Palatale D-E 5-6-7	Vélaires E 8-9
Sourdes	p passe	t toux	********	k cou
Sonores	b basse	d doux	********	g goût
Nasales	m masse	n nous	ɲ signe	********

	Fricatives				
	Labio-dentales A2	Alvéolaires C-D 5	Alvéolaire latérale B3	Prépalatale D5	Uvulaire F 9-10
Sourdes	f fer	s assis	********	ʃ chou	********
Sonores	v verre	z Asie	l la	ʒ joue	R rat

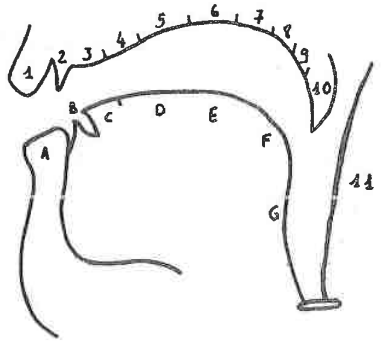


Figure 10 : Classification articulatoire du système phonétique du français d'après [CART-74].

l'appareil phonatoire et précisent le lieu d'articulation de chaque consonne.

2.5.2 Classification acoustique

Si, à l'évidence, la définition acoustique des phonèmes est primordiale pour la reconnaissance automatique, la grande variabilité des spectres caractéristiques d'un phonème ne facilite pas la désignation de formes distinctives. De ce fait, les principaux critères acoustiques utilisés sont basés sur la corrélation plus ou moins directe de la matière acoustique d'un phonème avec ses propriétés articulatoires. Tous les traits distinctifs proposés dans [JAKO-63b], [MALB-72] et [ROSS-77] ne sont pas toujours exploitables en décodage acoustico-phonétique, et nous ne retiendrons dans cette étude que les critères acoustiques suivants:

- consonnantique / non consonnantique : énergie totale réduite (élevée)
- voisé / non voisé : présence (absence) d'une excitation périodique de basse fréquence
- vocalique / non vocalique : présence (absence) d'une structure de formants nettement définie.

2.6 La coarticulation

La coarticulation est un facteur essentiel dans la variabilité de la réalisation d'un phonème. Selon les contextes d'émission d'un phonème, les caractéristiques acoustiques obtenues peuvent être très différentes (figures 8 et 9). Le degré de complexité du son à produire et le manque de synchronisation entre les articulateurs en sont les principaux responsables.

Pour illustrer ce problème, considérons les exemples présentés sur la figure 11. Le premier exemple caractérise la coarticulation de la langue: la langue ne participant pas à la production de la fricative /f/, elle prend aussitôt la position dela voyelle /o/ après avoir participé à l'articulation du son /l/. Résultat: il semble que ce soit /ilo/ qui ait été prononcé et non pas

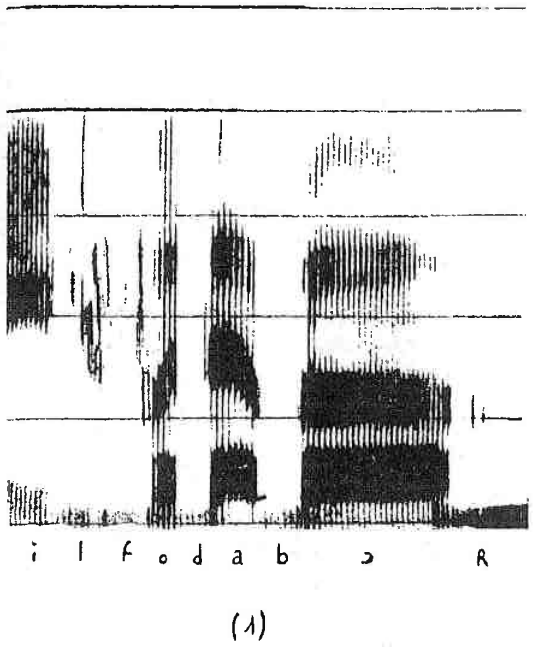


Figure 11

Exemples de coarticulation

/ilfo/ à la lecture du spectrogramme. Le second exemple caractérise la coarticulation des lèvres: la fricative /s/ étant un phonème alvéolaire, les lèvres ne participent pas à l'articulation de ce dernier et passent directement de la position de la voyelle /ə/ à celle de la voyelle /y/. Conséquence: le bruit de friction du /s/ se trouve décalé plus bas fréquentiellement.

Le niveau de perturbation des caractéristiques acoustiques des phonèmes est très variable, mais il arrive dans certains cas que l'influence du contexte entraine la perte ou la modification de certains traits distinctifs du phonème. Ces phénomènes, qui sont propres à une langue et en nombre limité, sont répertoriés sous le terme "règles d'assimilation". Par exemple, pour toute suite de deux consonnes, la seconde conserve toujours le trait de sonorité de la première.

CHAPITRE 3
LE DECODAGE ACOUSTICO-PHONETIQUE

3.1 Introduction

Le décodage acoustico-phonétique est l'étape du traitement automatique de la parole qui cherche à transformer le signal acoustique en une suite d'unités pseudo-phonétiques discrètes, c'est à dire en une suite de symboles équivalents ou non aux phonèmes. Cette suite est ensuite traitée par d'autres modules du système; analyseurs lexical, syntaxique et sémantique entre autres, dans le but de décoder le message contenu dans le signal. Ce schéma constitue la première approche du décodage acoustico-phonétique et est généralement désigné par les termes de "bottom up", approche ascendante ou approche guidée par les données. Inversement, le décodeur peut être assimilé à une procédure qui évalue la vraisemblance d'hypothèses sur le contenu du signal proposées par les modules de plus haut niveau, c'est l'approche descendante du décodage ou "top down".

Les performances du décodeur acoustico-phonétique ont une influence prépondérante sur celles de l'ensemble des modules d'un système de reconnaissance de la parole. Cette corrélation s'explique par le fait que le décodage constitue souvent la première étape du processus de reconnaissance, et en conséquence, par le biais de la propagation des erreurs, toute erreur commise à ce niveau est beaucoup plus coûteuse que celles issues des modules de niveau supérieur. De plus, le décodage est généralement le seul composant du système à traiter directement le signal acoustique. En dehors de l'analyseur prosodique, tous les autres composants interprètent le signal comme une suite d'unités discrètes telles que segments phonétiques, syllabes, mots ou phrases. Et la réduction du nombre de données qui résulte de cette discrétisation du signal, s'accompagne souvent d'une perte d'information [ZUE-78].

3.2 Unités de reconnaissance

Le rôle primordial du décodage acoustico-phonétique est de segmenter le continuum de parole en utilisant des indices de séparation présents dans le signal. Mais quelle unité de segmentation utiliser? La situation idéale serait celle où chaque segment correspondrait à un phonème. Malheureusement, un même phonème peut se réaliser non seulement en un segment, mais aussi quelquefois en une suite de plusieurs segments, ce qui complique les tâches de segmentation et d'interprétation du signal en terme de suite de phonèmes. Pour des applications n'exigeant que des vocabulaires limités, le mot est utilisé avec succés dans plusieurs systèmes de reconnaissance [GAGN-82], [DIMA-84]. On identifie alors directement les mots en les comparant à une représentation acoustique de référence. Le problème réside dans le fait que le nombre de références doit être au moins égal au nombre de mots à reconnaître, et pour des vocabulaires de taille plus importante (au delà de quelques centaines de mots), la prise en compte des phonèmes contextuels nécessite un nombre de références trop important à gérer. L'utilisation d'unités de segmentation de taille inférieure s'avère donc nécessaire.

Dans les systèmes de reconnaissance de la parole récents, on dénote une certaine tendance à choisir la syllabe comme unité de reconnaissance [FUJI-75], [MERM-75a], [DEMO-81], [PERE-81], et [MELO-82]. La localisation de la coarticulation en grande partie à l'intérieur de la syllabe et la décomposition naturelle du signal en syllabes en sont les principales raisons.

La qualité relativement bonne de la synthèse par diphones a également incité les chercheurs à utiliser cette unité comme unité de reconnaissance [RUSK-81], [ROSE-81], [WISW-81] et [MARI-81].

Intuitivement le phonème reste malgré tout l'unité de base des systèmes de reconnaissance de la parole continue à vocabulaire étendu. La caractéristique essentielle d'un phonème est pourtant son instabilité et sa variabilité. Le même phonème peut en effet être soumis à des influences très fortes et très diverses dues à la coarticulation, l'accentuation, le locuteur, le

débit d'élocution, etc... conduisant à une multitude de variantes appelées "allophones".

Ces difficultés ont conduit certains chercheurs à introduire un ensemble d'unités de reconnaissance plus important que celui des phonèmes et par là-même à considérer les différents allophones d'un phonème comme des unités distinctes. Ce type d'unités est généralement référencé sous le terme "phone", et peut être déterminé par des techniques de "clustering" [LELI-81].

Les avantages et inconvénients liés à l'utilisation de l'une de ces unités comme unité minimale de reconnaissance sont résumés dans le tableau 12 emprunté à [SHOU-80]. Mais une bonne solution est d'utiliser différentes unités dans un même système aux différentes phases de reconnaissance. C'est le cas du système de reconnaissance de parole continue implanté actuellement dans notre laboratoire [LAZR-83], qui utilise les syllabes pour la segmentation et les phonèmes pour la phase d'identification.

3.3 Paramétrisation et analyse acoustique du signal vocal

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

3.3.1 Représentations temporelles et fréquentielles du signal

Les différents types d'analyse acoustique requièrent un codage numérique de l'onde sonore. Le signal vocal est généralement échantillonné et numérisé à l'aide d'un convertisseur analogique/numérique, la fréquence d'échantillonage oscillant entre 6000 et 20000 Hz.

	AVANTAGES	INCONVENIENTS
MOT	- coarticulation incluse - indépendant de la langue - reconnaissance simple	- grands vocabulaires - adaptation du locuteur
SYLLABE	- facile à localiser (voyelle) - coarticulation incluse	- nombre total élevé - frontières mal définies
DIPHONE	- contient une partie de la coarticulation	- nombre total (1000) - problèmes de segmentation
PHONEME	- nombre total peu élevé - codage aisé des mots dans le lexique	- très dépendant du contexte - pas facile a localiser - algorithmes et règles complexes pour segmentation et reconnaissance
PHONE	- peut correspondre à un segment acoustique - facile à localiser et à reconnaitre	- nombre total peut être grand - peut dépendre du contexte

Tableau 12
Unités de base de la reconnaissance d'après [SHOU-80]

De nombreuses paramétrisations peuvent alors être obtenues à partir du signal numérisé. Certains paramètres comme la mesure de la densité des passages par zéro du signal et la fréquence fondamentale peuvent être directement extraits du signal dans le domaine temporel. Mais la représentation en fréquences du signal est la plus couramment utilisée. Elle est avantageuse à deux points de vue: d'une part, l'étude de l'audition montre que l'oreille effectue dans un premier temps une sorte d'analyse fréquentielle, d'autre part, cette forme d'analyse permet une représentation du signal vocal assez fidèle. Il est donc souvent intéressant de déterminer le spectre à court terme du signal de parole (figure 13).

Il existe plusieurs techniques pour obtenir ce spectre à court terme. Une des plus utilisées est l'analyse par prédiction linéaire ou analyse LPC [ITAK-68], [MAKH-72]. Cette technique s'appuie sur la modélisation du système phonatoire par un filtre numérique n-pôles qui est excité par des trains d'impulsions périodiques pour les sons voisés et aléatoires pour les sons non voisés. Les coefficients du filtre n-pôles spécifiant la fonction de transfert, l'estimation du spectre à court terme peut être réduit au problème de déterminer ces coefficients. Par suite,on obtient les fréquences formantiques, la fréquence fondamentale et la fonction d'aire du conduit vocal.

Parmi les autres types d'analyse couramment utilisés, on peut citer:

- l'analyse par transformée rapide de Fourier (FFT) [COOL-65]. Elle présente l'avantage de pouvoir obtenir une représentation fréquentielle du signal aussi fine qu'on le souhaite pour un temps de calcul raisonnable, mais la modulation du spectre obtenu par la fréquence fondamentale rend difficile certains traitements, et en particulier la localisation des maxima d'énergie (figure 14).
- l'analyse spectrale par bancs de filtres (vocodeurs à canaux), qui se caractérise par une compression importante de la quantité d'information nécessaire au codage du signal.
- l'analyse homomorphique ou cepstrale [SCHA-70]. L'analyse du cepstre, qui est la transformée de Fourier inverse du logarithme du module du spectre, permet à la

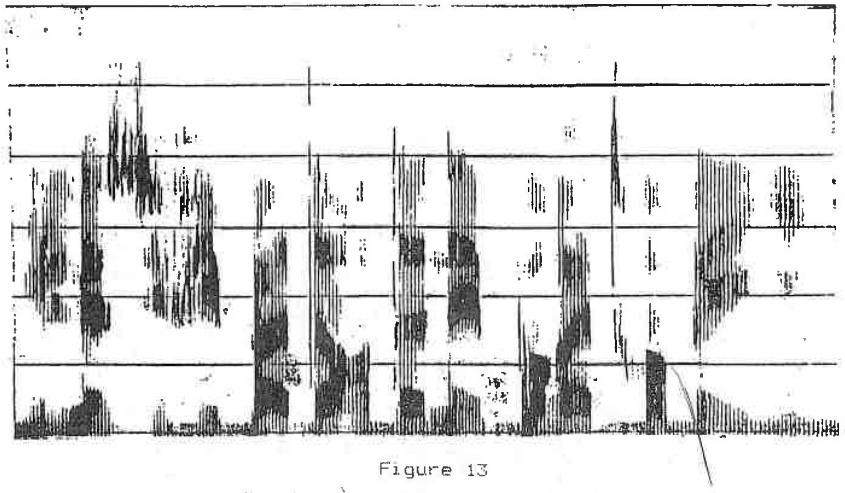


Figure 13

Spectre à court terme de la phrase
/il ne suffit pas d'ordonner pour être obéi/

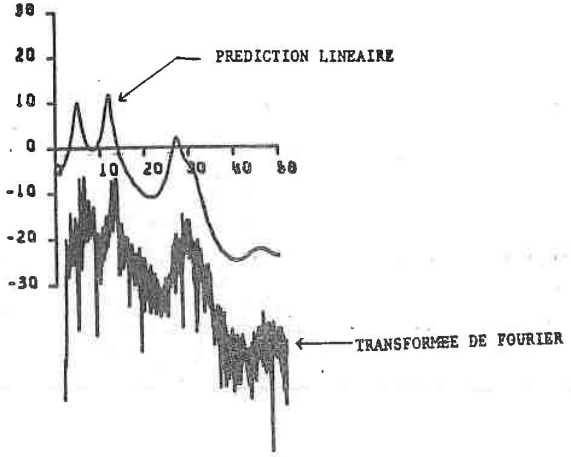


Figure 14

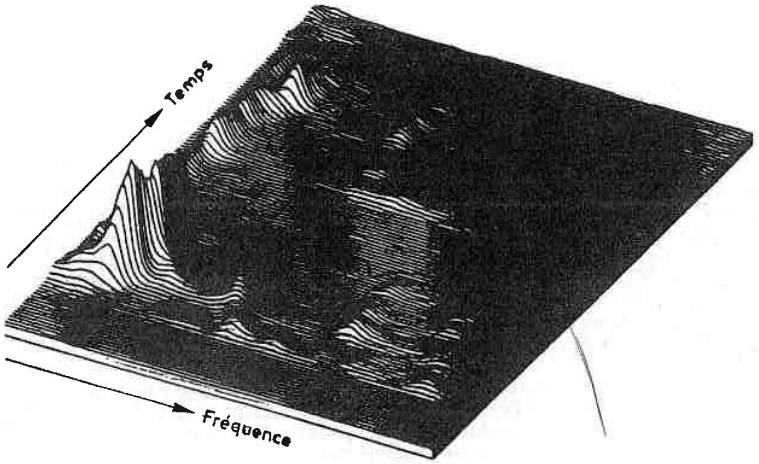
Comparaison des spectres par prédiction
linéaire et FFT

fois d'extraire la fréquence fondamentale et le spectre du conduit vocal tout en préservant une certaine économie de représentation: quelques coefficients seulement s'avèrent nécessaires pour représenter le signal vocal (figure 15a). Un inconvénient: c'est une analyse coûteuse en temps de calcul. Les différentes étapes du calcul cepstral sont schématisées sur la figure 15b.
Cette dernière analyse donnera de meilleurs résultats si le spectre est représenté à l'aide d'une échelle fréquentielle mel (étalement linéaire des fréquences en deçà de 1000 Hz, puis logarithmique au-delà), avant d'appliquer la transformation cepstrale sur le spectre [DAVI-80], [ELEN-82] et [DAME-83].

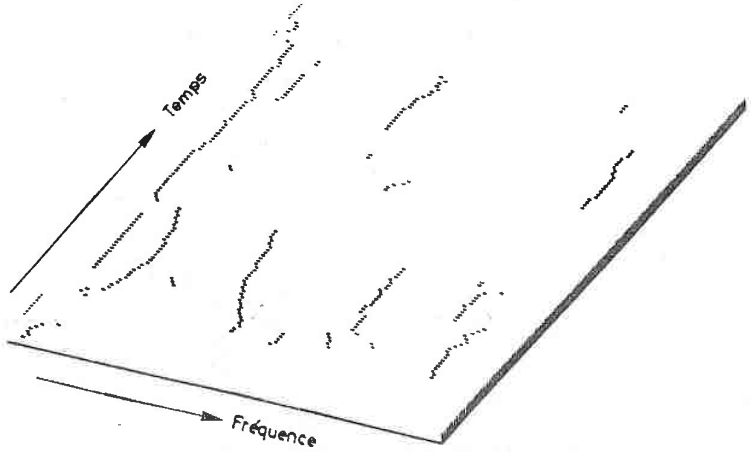
3.3.2 Paramètres utiles à l'analyse phonétique

A partir du signal lui-même ou à partir de ses représentations diverses, on peut dériver d'autres paramètres caractéristiques:

- les densités des passages par zéro du signal et de sa dérivée sont de bons paramètres pour détecter les sons fricatifs [HATO-81], et dans une moindre mesure pour évaluer les deux premières fréquences formantiques (quantités des passages par zéro du signal pour le premier formant et de sa dérivée pour le second formant).
- la courbe de variations de la fréquence fondamentale (F0), dont les techniques de détection et de mesure sont très nombreuses [RABI-76], est un paramètre important pour distinguer les sons voisés des sons non voisés. C'est d'autre part l'un des paramètres les plus importants de la prosodie.
- les différents critères énergétiques, courbe d'énergie totale et variations de l'énergie dans certaines bandes de fréquence, sont de bons indices pour segmenter le continuum vocal.
- les fréquences formantiques, qui sont reliées directement aux résonnances du conduit vocal, figurent bien sûr parmi les paramètres acoustiques les plus utilisés. D'une part, les voyelles et les consonnes sonores sont essentillement caractérisées par les valeurs des trois premières fréquences formantiques, et d'autre part, les transitions des formants peuvent être utilisées pour déterminer le lieu d'articulation, et par la suite la nature des phonèmes adjacents. Ces fréquences



Spectre du mot "table" obtenu par la méthode du cepstre.



Formants du mot "table" (méthode du cepstre).

Figure 15a
(d'après [MERC-82])

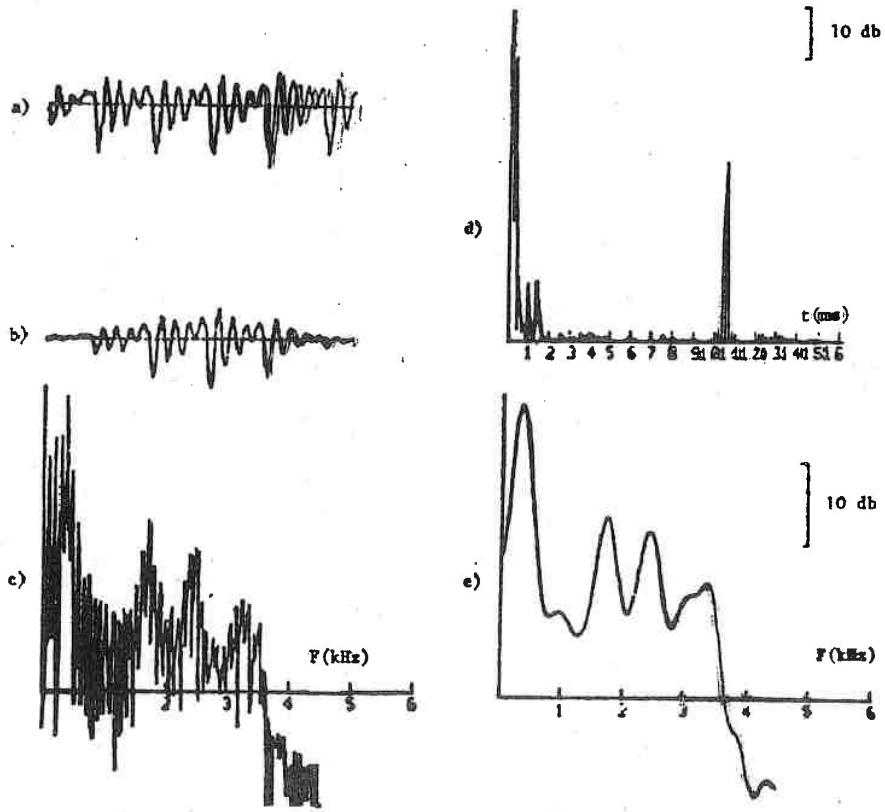


Figure 15b
Les différentes étapes d'une analyse cepstrale
d'après CARAYANNIS

a) Signal original
b) Après fenêtre de Hamming
c) Spectre brut par FFT
d) Cepstre
e) Spectre lissé

formantiques sont évaluées soit par analyse de prédiction linéaire, soit par transformée de Fourier, soit par analyse cepstrale (figure 15a).
Les autres paramètres fréquemment utilisés sont la dérivée du spectre pour la segmentation, les moments du spectre [ZWIC-79] ou encore les paramètres articulatoires [SHIR-78] et [LOCH-81].

Tous ces paramètres vont être utilisés par le décodeur acoustico-phonétique pour localiser et identifier la suite des sons dans le signal et pour décrire et prendre en compte les phénomènes de coarticulation.

3.4 L'analyse phonétique

Le processus de reconnaissnce phonétique peut être divisé en deux parties qui bien que distinctes n'en sont pas moins complémentaires. L'une des phases, appelée "segmentation", divise le signal vocal en unités phonétiques ou segments. La seconde phase consiste à identifier le ou les phonèmes correspondant à chaque segment: c'est "l'étiquetage". Il n'existe pas d'ordre prédéfini pour exécuter ces opérations. Certains systèmes segmentent dans un premier temps le signal, puis pour chaque segment extraient les paramètres acoustiques adéquats qui vont caractériser les traits phonétiques utilisés lors de l'étiquetage. D'autres systèmes par contre, découpent le signal en unités de même durée (10 ms par exemple), étiquettent chaque unité puis regroupent les unités de même étiquette pour constituer les segments.

3.4.1 La segmentation

La segmentation est un des problèmes les plus difficiles à résoudre. On peut distinguer deux procédures:

- la segmentation explicite [WEIN-75], [SCHW-76], [MERM-75b], [DEMO-80].
A partir des variations des paramètres tels que

l'énergie, la fréquence fondamentale, la dérivée du spectre, on peut décomposer le signal soit en pseudo-syllabes, soit en zones stables et instables très corrélées aux différents évènements acoustiques constituant les phonèmes, segmentation qui peut être affinée progressivement au moment de la reconnaissance.
Une procédure fréquemment employée [WOOD-76] est de détecter les pics et les vallées sur la courbe lissée de l'énergie contenue dans la plage fréquentielle [100Hz-500Hz], ce qui permet de séparer les régions sourdes (vallées) des régions sonores (pics). La même procédure est alors appliquée sur les zones sonores dans les plages fréquentielles [600Hz-3000Hz] et [3000Hz-5000Hz]. Les vallées caractérisent les consonnes sonores (nasales et liquides) et les consonnes sourdes qui n'auraient pas été détectées lors de la première passe. On pourra affiner la segmentation en utilisant les fréquences formantiques pour détecter les voyelles des liquides et nasales non éliminées sur le critère précédent, et la densité des passages par zéro du signal pour distinguer les fricatives des plosives. Le principal avantage de cette segmentation hiérarchisée est de diminuer la complexité de la reconnaissance en permettant un positionnement immédiat sur le segment à identifier, avec des frontières bien localisées. Ce type de segmentation ne sera efficient que si l'on a procédé au préalable à une analyse approfondie de la langue parlée à reconnaître, et on perçoit tout l'interêt qu'il peut y avoir à analyser et formaliser l'expertise humaine en lecture de spectrogrammes de parole pour réaliser un bon algorithme de segmentation. Le système de segmentation NOVOCA développé dans notre laboratoire [FOHR-85], dont le but est de trouver tous les noyaux vocaliques d'une phrase, est fondé sur la démarche de l'expert humain, ce qui explique en grande partie les performances obtenues: 98% de noyaux trouvés, moins de 5% d'insertions et moins de 2% d'omissions.

- la segmentation implicite [DIXO-77], [JELI-81], [HATO-81], [CALL-82].
Dans ce type de segmentation, c'est le processus d'identification échantillon par échantillon qui permet de localiser l'élément à reconnaître par regroupement des échantillons consécutifs ayant même étiquette. Ce processus est également très utilisé.

Le nombre d'erreurs engendrées lors de la segmentation (omissions, fusions, et insertions) est un problème important. Une solution intéressante est le concept de treillis de segmentation adopté dans le système de BBN [WOOD-76]. Lorsque les paramètres acoustiques ne permettent pas de segmenter de façon sûre une région du continuum vocal, plusieurs hypothèses sont alors proposées, chacune d'elles étant pondérée par un coefficient de vraissemblance. L'ensemble de ces hypothèses constitue le treillis de segmentation. Une autre solution serait d'acquérir un modèle de ces erreurs, modèle probalistique par exemple [BAHL-78], et de l'intégrer dans le lexique en même temps que les règles phonologiques de prononciation.

Il est à noter que la segmentation n'est pas toujours réalisée au niveau du phonème. Plusieurs systèmes de reconnaissance utilisent avec succés la syllabe ou la demi-syllabe comme unité de segmentation [FUJI-75], [DEMO-76], [RUSK-82]. Par rapport au phonème, l'unité syllabique présente l'intérêt d'être plus facile à localiser, et une grande partie des phénomènes de coarticulation y est intégrée. Par contre, les frontières d'une syllabe sont plus difficiles à déterminer avec précision, et le nombre élevé de syllabes dans une langue ne facilite pas leur identification. Un autre exemple de segmentation non phonémique est proposé dans le système HARPY où l'unité de reconnaissance est le phone [LOWE-76].

3.4.2 L'étiquetage

Le second module d'un analyseur phonétique est le module d'étiquetage, qui se donne comme but l'identification des unités issues de la phase de segmentation. On peut retenir essentiellement deux classes de méthodes d'identification qui prennent en compte ou non l'influence des phénomènes de coarticulation sur la réalisation acoustiques des phonèmes:

- les méthodes indépendantes du contexte phonémique.

Elles exploitent en général les techniques usuelles de reconnaissance des formes. Chaque unité de

reconnaissance est représentée par une ou plusieurs formes de référence qui sont regroupées au sein de vocabulaires ou dictionnaires. Ces références peuvent être extraites à partir d'une seule locution d'un locuteur donné, ou alors moyenner les formes extraites sur plusieurs locutions enregistrées par différents locuteurs. Dans le cas des phonèmes, elles peuvent aussi être synthétisées à partir d'un ensemble de règles décrivant une réalisation acoustique du phonème associé [KLAT-75], [COOK-76].

Pendant la phase d'identification, on extrait de chaque segment à reconnaître le même modèle paramétrique que celui qui a servi à construire les références, puis on compare ce modèle aux différents prototypes contenus dans les vocabulaires.

La représentation de ces références sous forme de spectres ou de cepstres rend ces systèmes entièrement dépendants du locuteur ayant réalisé l'apprentissage des références [HATO-81], et on observe une dégradation sensible des performances en reconnaissance multilocuteurs.

Il existe toutefois des méthodes permettant l'identification des unités à partir d'indices et de traits phonétiques, et qui ne prennent pas en compte l'environnement de ces unités. Le système initialement implanté au CRIN s'appuyait sur un modèle mixte d'identification: identification spectrale (centiseconde) et identification par les indices et les traits [LAZR-83].

- les mèthodes dépendantes du contexte phonémique.

Les méthodes dépendantes du contexte nécessitent généralement l'extraction de traits acoustico-phonétiques en s'appuyant sur un ensemble de connaissances sur la nature acoustique des différentes unités à identifier. On détermine ensuite l'identité de l'unité en utilisant des techniques statistiques ou bien par un raisonnement fondé sur des règles.

Cette approche présente plusieurs avantages. Nous avons vu que les différents sons sont caractérisés par des traits acoustiques différents, et il est raisonnable de penser que ces stratégies de reconnaissance sont plus à même d'identifier les segments de parole que les mèthodes indépendantes du contexte qui s'appuient sur une représentation uniforme et arbitraire du signal, tels les coefficients LPC et la représentation spectrale. De

plus, la description et l'identification des unités par des traits ont l'avantage de diminuer l'influence des différences acoustiques dûes aux locuteurs ou aux conditions d'enregistrement.
La prise en compte de l'environnement phonétique qui constitue la clef du succés de ces méthodes, en est aussi paradoxalement le talon d'Achille. Pour que les stratégies de reconnaissance soient statistiquement significatives, il est en effet nécessaire d'étudier l'influence de tous les contextes phonémiques possibles, et la constitution d'un corpus de données permettant une telle étude demeure un des problèmes majeurs actuels. Parmi les travaux qui ont été réalisés dans ce domaine, nous citerons ceux de Schwartz [SCHW-76], de Baker & Al [BAKE-83] pour l'anglais et du GRECO Communication Parlée [CARR-84] pour le français.

Toutes les stratégies de reconnaissance présentées font appel à des outils mathématiques couramment employés en reconnaissance des formes et en intelligence artificielle: comparaison de prototypes, programmation dynamique, analyses discriminantes, automates probabilistiques, ensembles flous, techniques de clustering, systèmes experts, etc... que nous présenterons en détail dans le chapitre suivant.

3.5 L'apprentissage

La phase d'acquisition des connaissances est d'une importance primordiale pour tout module de reconnaissance car elle conditionne les performances du système.
Il existe plusieurs façons de réaliser l'apprentissage qui sont parfois complémentaires. La méthode la plus simple d'apprentissage et de représentation des connaissances consiste, comme nous l'avons vu au paragraphe précédent, à mémoriser les formes de référence des unités à identifier: phonèmes, syllabes, diphones, mots, etc... Ces références sont en général extraites d'un corpus de phrases ou de mots soit de façon manuelle par l'intermédiaire d'un opérateur, soit de façon semi-automatique ou automatique à l'aide de programmes de segmentation et de programmes permettant d'étiqueter

chaque segment [JELI-80]. Ce type d'apprentissage doit présenter certaines qualités:
-la fiabilité et la robustesse des formes acquises pour tenir compte de la dérive de la voix, des conditions d'enregistrement, ...
- un nombre de références sélectionnées 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(s) réalisant l'apprentissage,
- la prise en compte des diverses variantes de prononciation des unités à reconnaître.
Parmi les différentes méthodes d'apprentissage qui s'appuient sur ce modèle et dont on trouvera une étude comparative pour les systèmes de reconnaissance de mots isolés dans [RABI-80], on retiendra les méthodes de "moyennage" et de "clustering" qui présentent l'avantage de fournir des systèmes indépendants du locuteur.

Le deuxième type de méthodes utilisées concerne surtout les systèmes qui s'appuient sur les traits acoustico-phonétiques pour identifier les unités segmentées. Il consiste à utiliser les traits comme unités de reconnaissnce et à rechercher l'ensemble des indices acoustiques et des paramètres associés susceptibles de représenter un trait. L'ensemble des connaissances est alors formalisé sous forme de règles liant indices acoustiques, traits et unités à reconnaître: c'est l'apprentissage par règles. Une telle représentation des connaissances a l'avantage d'être relativement peu dépendante du locuteur, bien que certains paramètres du système gardent une certaine dépendance vis-à-vis de ce dernier. Ainsi, certains locuteurs pourront utiliser un indice plutôt qu'un autre pour un trait donné.

3.6 Problèmes de normalisation et d'adaptation au locuteur

Les paramètres acoustiques utilisés dans l'analyse phonétique peuvent varier en fonction de plusieurs paramètres extra-linguistiques qui sont liés essentiellement aux conditions d'enregistrement et au locuteur. L'adaptation au locuteur, à la vitesse d'élocution et à l'environnement acoustique sont autant

de problèmes à prendre en compte si on veut obtenir des systèmes performants, et il est donc souhaitable de normaliser ces paramètres afin de minimiser leur influence dans le décodage.

3.6.1 Adaptation à l'environnement acoustique

La qualité du signal de aprole enregistré à partir d'un microphone dépend du placement et de l'orientation du microphone, de ses caractéristiques techniques, de l'acoustique du local où se déroule le dialogue avec la machine, et de la nature des bruits ambiants. Ces derniers constituent dans la plupart des cas la source principale des modifications acoustiques du message. Ils peuvent être permanents et réguliers (ventilation), intermittents (machinerie), aléatoires (portes, conversations) ou bien systématiques (réverbération). Différents schémas de normalisation ont été proposés dans ce domaine et sont exposés dans [REDD-76] et [BAKE-82].

3.6.2 Normalisation du signal

Hormis les personnes habituées à parler devant un microphone (animateurs, présentateurs, hôtesses ou ... chercheurs en parole), peu de locuteurs sont capables de maîtriser leur phrasé dans une situation donnée (vitesse d'élocution et intensité notamment). Or, l'état émotionel et physique du locuteur peut provoquer des modifications anatomiques du sytème phonatoire de courtes durées. Ainsi, la fréquence fondamentale et certains formants sont modifiés par des émotions comme la tristesse, la joie, la peur ou la colère [WILL-70]. Il en est de même pour l'état de fatigue 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 volontaire ou non de la voix. La normalisation du signal est généralement réalisée sur deux critères: l'amplitude et la durée. La normalisation en amplitude peut être obtenue de manière analogique à l'aide d'un circuit de contrôle automatique du gain pendant l'acquisition, ou

par logiciel en utilisant un facteur de normalisation basé sur l'énergie moyenne du signal par exemple.

Le problème de la normalisation en durée est plus délicat à traiter, car toute mesure de la durée des unités à reconnaître dépend étroitement de la vitesse d'élocution utilisée. Bien qu'il existe plusieurs métriques pour la vitesse d'élocution, la plupart des systèmes de reconnaissance n'en tiennent pas compte. Ces systèmes requièrent alors de la part du locuteur de s'exprimer sur un rythme comparable à celui employé par les autres utilisateurs, et forts de cette hypothèse, se contentent d'effectuer une normalisation temporelle à l'aide d'une méthode de comparaison dynamique.

3.6.3 Adaptation au locuteur

De tous les facteurs de variation considérés, il est probable que les différences anatomiques sont une des sources de variabilités dont l'influence sur les propriétés acoustiques des sons ont été le plus étudiées [STEV-71]. Ces différences sont conditionnées par l'age, le sexe et l'hérédité, concernent à la fois le système respiratoire, le larynx, le conduit vocal et les cavités nasales. Ainsi, les caractéristiques des cordes vocales (longueur, épaisseur, tension, etc...) et du système respiratoire (capacité des poumons) sont en grande partie responsables de la variabilité inter-locuteur de la fréquence fondamentale. De même, la taille et la forme des conduits vocal et nasal ont une grande influence sur les fréquences des formants.

Les habitudes de prononciation du locuteur ne sont pas à négliger dans la prise en compte des sources de variabilité des paramètres acoustiques. Par exemple, les habitudes ou les comportements phonatoires propres à chaque individu peuvent entrainer des altérations 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 des élisions, insertions, substitutions des phonèmes à la jonction des mots. Comme autres facteurs de variabilité importants, on citera les

différences de prononciation relatives aux dialectes (accents régionaux) et aux langues. L'articulation française qui est caractérisée par une tendance antérieure et dominée par l'articulation labiale, se trouve 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 [CART-74].

Toutes ces considérations tendent à prouver qu'il est impossible d'obtenir des prototypes et des indices indépendants du locuteur, et il semble donc important d'inclure des procédures d'adaptation au locuteur dans les systèmes de reconnaissance de la parole. L'adaptation peut se faire de diverses façons, qui dépendent notemment des techniques de reconnaissance mises en oeuvre dans le système, et de la paramétrisation du signal vocal (figure 16).

Dans certains systèmes, l'adaptation se fait pendant la phase de reconnaissance. La normalisation peut porter sur les formes à reconnaître par l'intermédiaire des fréquences formantiques [SCHW-71], [WAKI-77], [NEAR-77], [BOE-80] ou par recalage fréquentiel [MATS-79], ou bien dans l'ajustement des paramètres du système [SAMB-75]. D'autres systèmes réalisent l'adaptation au locuteur pendant la phase d'apprentissage. Chaque locuteur prononce quelques phrases représentatives d'où sont extraites automatiquement les formes de références des unités à reconnaitre. Cette phase d'acquisition automatique peut se présenter sous deux formes:

- un apprentissage automatique des formes de référence à l'aide d'algorithmes de cadrage [LECO-79], [NADA-81], [WAGN-81], [BRID-83] et [NEEL-83].

- la génération automatique des formes de référence 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 analyse statistique d'un grand ensemble de phonèmes multilocuteurs, qui permet d'établir des relations indépendantes du locuteur entre les différents phonèmes [FURU-80] et [DIBE-84].

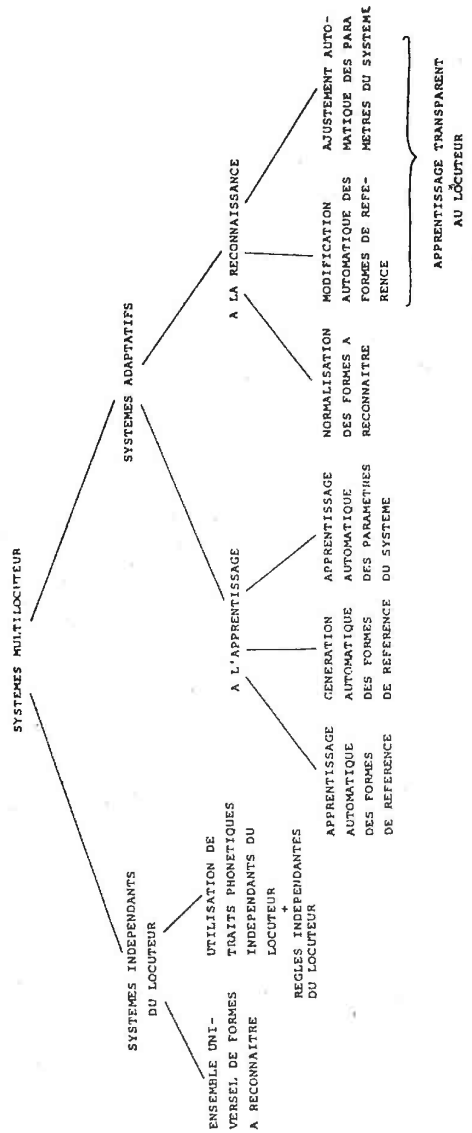


Figure 16 Les différentes approches de réalisation d'un système de la parole multilocuteur [PIST-84]

Ces deux types d'adaptation au locuteur ne sont pas exclusifs. On citera en exemple le système HARPY [LOWE-77] qui apprend dynamiquement les paramètres dépendants du locuteur pendant que ce dernier utilise le système.

CHAPITRE 4
ANALYSE PHONETIQUE ET R.F.I.A.

4.1 Introduction

Le module d'analyse phonétique est en lui-même un mini-système de reconnaissance: c'est le module d'interprétation du signal acoustique qui permet le passage de l'onde sonore au treillis phonétique. Et comme pour tout système de reconnaissance de la parole, on se trouve confronté à plusieurs choix:

- unités minimales à reconnaître (syllabe, diphone, phonème, traits...),
- modèle de représentation des connaissances (matrices, seuils, paramètres, ensembles de règles, objets structurés...),
- stratégies de reconnaissance mises en oeuvre (ascendante, descendante, dépendante ou non du contexte...),
- reconnaissance avec ou sans segementation préalable,
- techniques de normalisation utilisées.

Ces différents choix conduisent évidemment à toute une gamme de systèmes plus ou moins élaborés. Dans les paragraphes suivants, nous présenterons les principales options adoptées dans le domaine de l'analyse phonétique et essairons de dégager les tendances actuelles dans la représentation et l'utilisation des connaissances.

4.2 Analyse phonétique et reconnaissance des formes

4.2.1 Introduction

Historiquement, les méthodes de reconnaissance des formes ont guidé les premières stratégies d'identification adoptées en analyse phonétique. Généralement, les unités à reconnaître sont représentées sous une des diverses formes paramétriques que peut prendre le signal, et de ce fait, sont très influencées par la coarticulation. L'architecture de ces systèmes est souvent limitée à trois fonctions principales qui sont l'extraction des paramètres, la segmentation et la reconnaissance; cette architecture est en général mieux adaptée à la reconnaissance d'unités comme les mots, les

syllabes ou les demi-syllabes qui sont moins dépendantes du contexte.

Trois types de méthodes sont principalement utilisées en reconnaissance: les solutions heuristiques, les méthodes de corrélation et les analyses paramétriques. Leur inconvénient commun est la difficulté de constitution de dictionnaires de références.

4.2.2 Les solutions heuristiques

Dans l'approche heuristique, on est souvent limité en ce qui concerne les connaissances sur les relations liant les classes phonétiques et les traits acoustiques. Ainsi, il est possible d'expliciter le comportement de certains traits sur la réalisation d'un phonème particulier, et en même temps ignorer l'influence de ces mêmes traits sur les autres phonèmes. En conséquence, les stratégies de reconnaissance reposent généralement sur l'utilisation de systèmes à logique à seuil (SLS) pour déterminer la présence des traits acoustiques. Les différentes solutions proposées sont pondérées par des poids heuristiques appropriés, et ne sont retenus que le(s) phonème(s) présentant le(s) plus grand(s) poids. Cette approche, qui est utilisée pour identifier les consonnes dans le système KEAL [MERC-82] et dans le décodeur acoustico-phonétique de M. Lazrek [LAZR-83], est toutefois limitée du fait de son manque de formalisme mathématique, et ne permet pas d'obtenir de bonnes performances.

4.2.3 Les méthodes de corrélation

Dans ce type de méthodes, on calcule un taux de corrélation entre la forme à identifier et les différentes formes de références, et la référence présentant le score minimal est retenue pour étiqueter l'unité inconnue. Le calcul des taux de corrélation ou distances est généralement obtenu par comparaison de prototypes.

Le problème de ces méthodes réside d'une part, dans la variabilité de la vitesse d'élocution qui se traduit par l'obtention de formes vocales de longueur et de rythme différents lors d'élocutions d'un même mot pour un même locuteur et à fortiori pour des locuteurs différents. Ces distorsions se caractérisent au niveau temporel par des compressions et des déformations non linéaires de certains phonèmes en fonction du rythme et du débit de la parole (figure 17a). D'autre part, une autre cause majeure de variabilité réside dans l'altération des unités phonétiques en fonction du contexte, qui rend difficile la définition de formes de référence.

Un autre genre de difficultés - non inhérent à la prononciation - intervient lors de la phase d'analyse du signal vocal. En effet, il se peut que certains phonèmes de début ou de fin de forme ne puissent être détectés en raison de leur énergie trop faible. La forme obtenue subit alors une troncature qui peut par la suite fausser la reconnaissance. On dénomme ce genre de problème, erreur de détection parole/non parole.

Ainsi, comme on le voit, la tâche de l'algorithme de comparaison est difficile dans la mesure où il doit être à-même de tenir compte de toutes les distorsions possibles pour pouvoir effectuer une reconnaissance correcte. Toutes les difficultés que nous venons de citer ne sont pas encore résolues à l'heure actuelle. Toutefois, des solutions ont été proposées pour compenser les distorsions dues à la variabilité de la vitesse d'élocution et aux erreurs de détection parole/non parole de l'analyseur acoustique.

4.2.3.1 Le recalage temporel

Afin de compenser les distorsions introduites par les variations de la vitesse d'élocution, le recalage temporel se propose d'effectuer une synchronisation des échelles de temps de deux formes à comparer comme le montre la figure 17a.

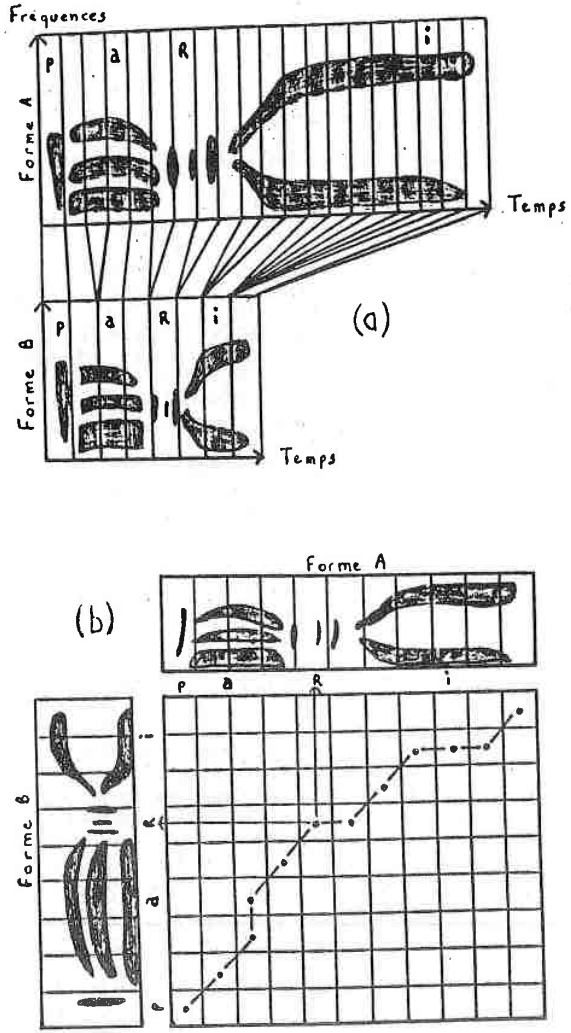


Figure 17 : Distorsions temporelles observées sur deux réalisations du mot /PARIS/ (a) et recalage temporel appliqué pour compenser les différences de durée et les variations non linéaires du rythme (b) [HATO-79]

Désignons par a(1),a(2),...,a(I) la suite des vecteurs fournis par l'analyseur acoustique caractérisant la forme A, I correspondant à la durée de la forme A, et de même par b(1),b(2),...,b(J) la suite de vecteurs caractérisant la forme B, J correspondant à la durée de la forme B. La stratégie optimale consiste à effectuer un recalage non linéaire de l'axe des temps optimal au sens d'une métrique d, c'est à dire à associer les échantillons a(i) et b(j) les plus semblables sans tenir compte de leurs positions respectives. Si l'on porte sur un axe horizontal les différents vecteurs de la forme A - un point représentant un vecteur - et sur un axe vertical les vecteurs de la forme B, une représentation de la fonction de recalage est un chemin dans le plan ainsi défini, et le problème consiste alors à rechercher le chemin optimal Go dans ce plan (figure 17b).

Soit G un chemin quelconque dans le plan (i,j). La fonctionelle qui est généralement associée à G et que l'on appelle distance entre A et B, est donnée par la relation:

$$(4.1)\quad D(G) = \frac{1}{N(P)} \sum_{k=1}^{K} d(i(k),j(k)).p(k)$$

où : - K est le nombre de points du chemin de recalage,
- d(i,j) est la distance locale entre le iéme vecteur de la forme A et le jéme vecteur de la forme B, les distances de Hamming, euclidienne et d'Ikatura pour les coefficients LPC sont les distances les plus couramment utilisées,
- p(k) est une fonction de pondération qui diffère suivant la transition locale (i(k-1),j(k-1)) ---> (i(k),j(k)),
- N(P) est un facteur de normalisation dont le rôle est de rendre D(G) indépendant de la longueur du chemin de recalage, on choisit généralement $N(P)= \sum_{k=1} p(k)$.

La fonction de recalage qui nous intéresse est celle qui minimise D(G). La solution du recalage temporel D(Go) est donc donnée par la relation suivante,

$$(4.2)\quad D(Go) = \min_{G}\left(\frac{1}{N(F)} \sum_{k=1}^{K} d(i(k),j(k)).p(k) \right)$$

qui est la définition formelle de la distance normalisée entre A et B.

Ainsi exposée, la méthode est prohibitive en temps de calcul, et on impose certaines restrictions à la fonction de recalage afin de limiter les temps de traitement.

4.2.3.2 Contraintes imposées aux chemins de recalage

Afin que la fonction de recalage respecte l'évolution dans le temps du signal vocal, celle-ci est soumise à des conditions de monotonie et de continuité exprimées par les relations suivantes:

$$(4.3)\quad i(k-1)<i(k) \quad \text{et} \quad j(k-1)<j(k) \qquad \text{(monotonie)}$$

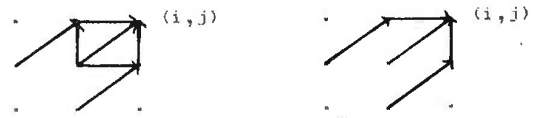
$$(4.4)\quad i(k)-i(k-1)<1 \quad \text{et} \quad j(k)-j(k-1)<1 \qquad \text{(continuité)}$$

Des relations (4.3) et (4.4), on peut déduire la propriété (4.5) qui lie les différents points c(k) du chemin de recalage:

$$(4.5)\quad c(k) = \begin{cases} i(k),j(k)-1 \\ i(k)-1,j(k)-1 \\ i(k)-1,j(k) \end{cases}$$

La fonction de recalage se doit aussi de ne pas effectuer des compressions ou des dilatations irréalistes. Dans cette optique, celle-ci est assujettie à des contraintes locales qui lui empêchent d'effectuer certains déplacements locaux. Ainsi, on interdit souvent par exemple les changements de direction orthogonaux et les déplacements consécutifs dans une même direction, si le sens du déplacement est horizontal ou vertical.

Cette restriction permet de réduire sensiblement le nombre de chemins.



Graphe normal Graphe réduit

Figure 18
Réduction du nombre de chemins par la contrainte de SAKOE & CHIBA [SAKO-78]

Les motifs visualisés par la figure 18 indiquent qu'un chemin de recalage ne peut aboutir au point (i,j) qu'en passant par (i-2,j-1) et (i-1,j), ou par (i-1,j-1), ou encore par (i-1,j-2) et (i,j-1). D'autre part, afin que la fonction de recalage tienne compte de l'ensemble des prélèvements de chacune des formes, des contraintes aux frontières lui sont imposées qui nécessitent la mise en correspondance des prélèvements initiaux et terminaux de chacune des deux formes. Ces contraintes sont exprimées par les relations (4.6) et (4.7).

$$(4.6)\quad i(1)=1 \text{ et } j(1)=1$$

$$(4.7)\quad i(K)=I \text{ et } j(K)=J$$

4.2.3.3 Résolution du problème par programmation dynamique

Le problème consiste à déterminer la distance normalisée entre A et B donnée par la relation (4.2). Cette équation peut être résolue par la programmation dynamique à l'aide du principe d'optimalité local introduit par Bellman [BELL-57]: "soit Gij le chemin

optimal joignant les points (1,1) et (i,j), alors pour tout point (i',j') appartenant à Gij, le chemin de recalage Gi'j' est optimal".

Désignons la distance cumulée optimale au point (i,j) associée à Gij par f(i,j). On a alors:

$$(4.8)\quad f(i,j) = \min_{K',i(k),j(k)} \left(\sum_{k=1}^{K'} d(i(k),j(k)).p(k) \right)$$

avec i(1)=1, j(1)=1, i(K')=i et j(K')=j.
D'après le principe d'optimalité local, il vient:

$$(4.9)\quad f(i,j) = \min_{i',j'} \left(f(i',j') + d((i',j'),(i,j)).p((i',j'),(i,j)) \right)$$

où (i',j') appartient à un voisinage de (i,j) défini par la contrainte locale utilisée.

Grâce à cette relation récursive, il est possible d'évaluer f(i,j) en tout point du plan (i,j), et par conséquent, la solution de l'équation (4.2) donnant le taux de dissemblance entre les formes A et B est donnée par l'équation (4.10):

$$(4.10)\quad D(Go) = D(A,B) = \frac{1}{N(P)} f(I,J)$$

Les relations (4.9) et (4.10) permettent de construire l'algorithme de programmation dynamique suivant:

1) Initialisation : f(1,1) = d(1,1).p(1)
2) Evaluation de f(i,j) pour $1 \leqslant i \leqslant I$ et $1 \leqslant j \leqslant J$
3) Détermination du taux de dissemblance D(Go)

Cet algorithme a fait l'objet de nombreuses variantes déterminées par le choix des contraintes locales imposées aux chemins de recalage et de la fonction de pondération associée. Parmi les différents algorithmes proposés, en particulier pour la

reconnaissance de mots isolés ou enchainés, nous citerons ceux de Sakoe & Chiba [SAKO-71], [SAKO-78], Rabiner [RABI-78], Myers [MYER-81] et Bridle & Nakagawa [BRID-82], [NAKA-83]. Pour une étude plus détaillée sur ce sujet, on pourra aussi se référer à [DIMA-84].

4.2.4 Les méthodes statistiques et paramétriques

Dans ces méthodes, on réalise une partition de l'espace des formes en autant de classes qu'il y a d'unités différentes à reconnaître. Chaque forme est décrite par N paramètres, et l'identification d'une forme se fait en déterminant son appartenance à une classe de formes parmi plusieurs classes, chaque classe pouvant être figurée par un nuage de points dans l'espace des paramètres. Le critère d'appartenance sera fondé par exemple sur un calcul de distance entre la forme inconnue et les centres de gravité des différentes classes possibles ou les voisins les plus proches.

Le but de ces méthodes est de trouver des fonctions discriminantes définissant des hypersurfaces séparatrices de nuages de points. On distingue généralement deux grandes familles de fonctions discriminantes:

1) les fonctions qui font abstraction de toute notion probabilistique et statistique. Une classe importante de ces fonctions discriminantes est constituée par les fonctions linéaires qui réalisent la séparation des formes à classer par apprentissage [MATO-83]. Le choix du type de distance utilisée (fonction de décision) déterminera la taille de la zone de rejet et donc la qualité de la séparation. La figure 19 illustre l'emploi de classificateurs linéaires pour séparer les classes w_1, w_2 et w_3. Le critère de séparation est défini par la fonction de décision suivante:

$$(4.11)\quad X \text{ inconnue} \in w_i \iff \begin{cases} d_i(X) > 0 \\ d_j(X) < 0 \quad \forall i \neq j \end{cases}$$

En cas de problème non linéaire, l'utilisation de classificateurs non linéaires est très couteuse, et il

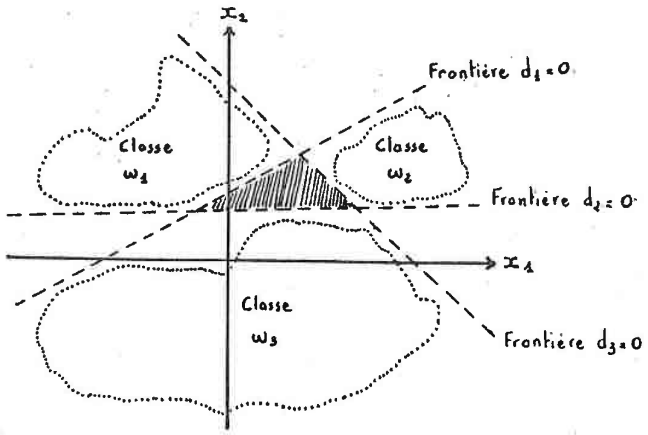


Figure 19
Separation a l'aide de classificateurs lineaires

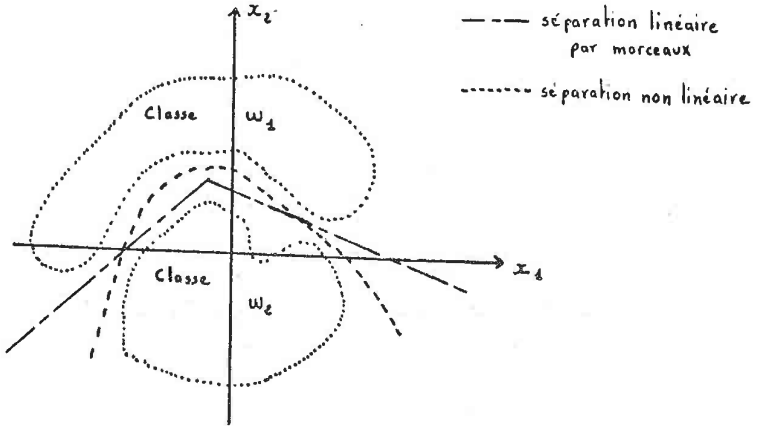


Figure 20
Separations lineaires par morceaux et non lineaire

est préférable de transformer le problème non linéaire en une suite de problèmes linéaires par application de l'algorithme de Mangassarian (figure 20):

Tant que Classe 3 = 0 Répéter

Séparation linéaire en 3 classes :
1) A 2) B 3) A.B

2) les fonctions qui s'appuient sur une distribution statistique des formes à classer. Ce sont les classificateurs de type Bayésien qui définissent la notion de distance comme la mesure de la probabilité qu'a une forme inconnue X d'appartenir à la classe w_i, i.e. $p(w_i/X)$. Cette probabilité est donnée par le théorème suivant:

$$(4.12) \quad p(w_i/X) = p(X/w_i).p(w_i).1/p(X)$$

En général, on ignore les valeurs des probabilités $p(X)$ et on émet l'hypothèse qu'elles sont équiprobables. La recherche de la distance optimale passe donc par la maximisation du produit $p(X/w_i).p(w_i)$. Si l'estimation du terme $p(w_i)$ peut être obtenue aisément à l'aide de méthodes statistiques classiques [ANDE-58], l'estimation du facteur $p(X/w_i)$, qui est typiquement une probabilité, est plus difficile. Parmi les techniques d'estimation de densité les plus utilisées, on peut retenir les méthodes paramétriques, et en particulier les distributions Gaussienne et de Poisson. Toutefois, des visualisations d'analyses en composantes principales ont montré que les nuages ne présentaient pas de formes suffisamment régulières pour pouvoir faire des hypothèses simples de fonction de densité, et il semble préférable d'utiliser des méthodes non paramétriques d'estimation de densité comme les estimateurs de Parzen [PARZ-62], [SOUD-85] ou les méthodes du plus proche voisin (NN et K-NN RULE), où on attribue à X la classe de son ou ses voisin(s) le(s) plus proche(s). Ainsi, parmi les systèmes utilisant les probabilités dans le module d'identification, le "Word Based Acoustic Processor" d'I.B.M. [BAHL-79] s'appuie sur un modèle Gaussien pour calculer les probabilités conditionnelles $p(X/w_i)$ alors que le décodeur

acoustico-phonétique de BBN [SCHW-76], [WOOD-76] & [COOK-77] utilise un modèle non paramétrique pour estimer ces probabilités.

Il est à noter que certains systèmes de reconnaissance qui utilisent ce type de méthodes 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 [BRIA-81],[DOLO-84] & [DIVO-85].

4.3 Modèle de reconnaissance synchrone

Les méthodes que nous venons de présenter se caractérisent par la constitution de dictionnaires de références propres à chaque locuteur, et par un codage des formes peu économique en environnement matériel (mémoire vive notamment).

Une variante de ces méthodes consiste à réduire la taille des dictionnaires de références par un codage vectoriel des échantillons extraits du signal: c'est le modèle de reconnaissance synchrone ou "centiseconde". Cette méthode consiste à découper le signal vocal en tranches égales de courte durée qui selon le type d'unité à reconnaître subiront un traitement différent pendant la phase d'identification.

Dans les systèmes de reconnaissance de mots ou de syllabes, on emploie les mêmes méthodes que celles présentées dans le paragraphe 4.2, l'interêt de la technique se limitant à la réduction de la taille des formes à comparer. Les algorithmes de comparaison dynamique implantés dans notre laboratoire [DIMA-84] & [BOYE-84] utilisent la représentation centiseconde comme codage du signal vocal.

La reconnaissance de phonèmes ou de phones procède de manière totalement différente [BAHL-78], [HATO-81] & [MERC-82]. Pour chaque échantillon centiseconde du mot ou de la phrase prononcé, on recherche les 3 phonèmes les plus proches en s'appuyant sur le calcul d'une distance entre le prélèvement et les références, et après avoir

établi une éventuelle préclassification sur des paramètres comme l'énergie et le voisement. L'identification des segments de même type se fait par un processus de décision majoritaire parmi les 3 phonèmes retenus pour chaque échantillon. Pour chaque phonème candidat, nous calculons son taux d'occurence dans le segment de la façon suivante:

$$(4.13) \quad Ti = \sum_{j=1}^{N} Pj(i)$$

où N est le nombre d'occurences du phonème i dans le segment, et Pj est une fonction de pondération de la position du phonème dans le treillis. Les phonèmes correspondant aux trois plus grandes valeurs Ti classées par ordre décroissant, constituent le triplet identifiant le segment. La figure 21 illustre l'utilisation de ce processus de décision majoritaire dans la reconnaissance de la phrase "C comme Célestin".

Par construction, les éléments représentatifs d'un segment donné sont homogènes et indépendants du contexte, et la méthode synchrone, qui nécessite toujours un apprentissage par locuteur, ne résout pas les problèmes contextuels et multilocuteurs du décodage acoustico-phonétique.

4.4 Analyse phonétique et intelligence artificielle

4.4.1 Introduction

Les techniques de reconnaissance que nous avons exposé sont mal adaptées à la reconnaissance d'unités phonétiques. Le problème contextuel est difficile à maîtriser: soit il est ignoré comme dans les modèles de reconnaissance synchrone, soit il est pris en compte mais nécessite alors la constitution de dictionnaires de références de taille très importante. Dans les deux cas, la solution consiste à réaliser un apprentissage par locuteur, et l'optimisation des techniques d'apprentissage constitue un pôle de recherche important dans la conception de systèmes de reconnaissance

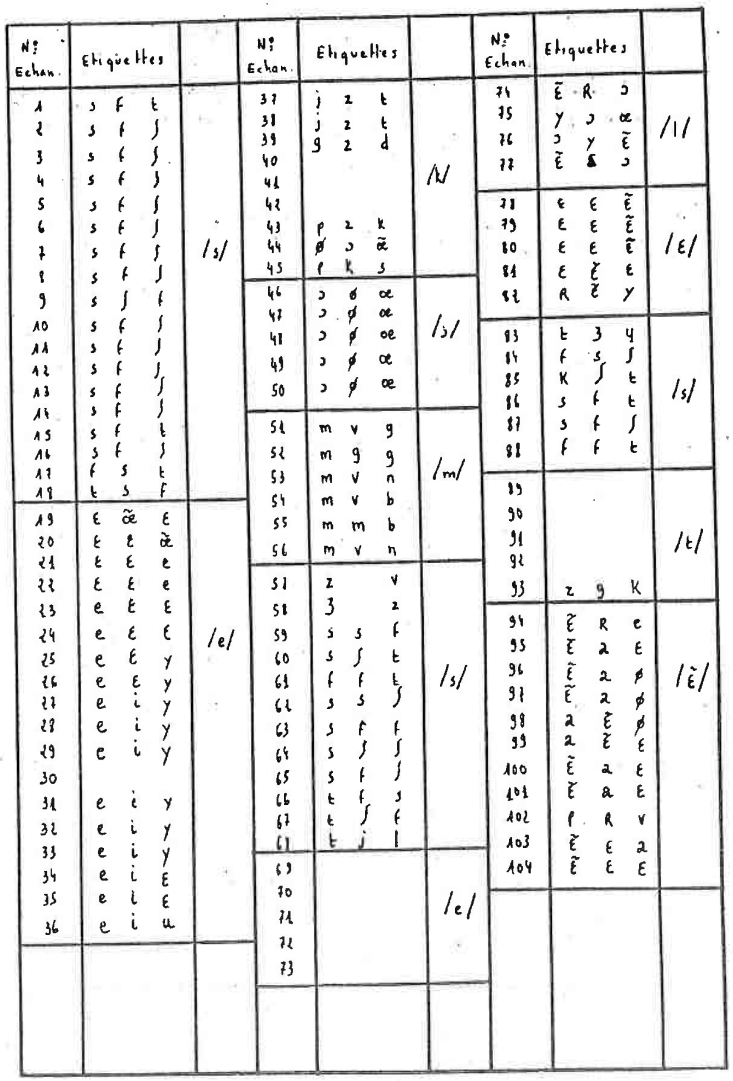


Figure 21

Reconnaissance centiseconde de la phrase

/C comme Celestin/

[MERC-82]

multilocuteurs performants.

Mais l'approche peut-être la plus prometteuse pour l'avenir reste la reconnaissance par traits phonétiques. Dans cette approche, les phonèmes, et donc les syllabes et les mots, sont représentés par un ensemble de traits phonétiques. Ainsi, indépendamment du contexte et du locuteur, le phonème /s/ peut être défini par l'ensemble {constrictif - sourd - alvéolaire}, et la voyelle /y/ par l'ensemble {orale - labiale - fermée - antérieure}. Il s'agit ensuite de déterminer un ensemble d'indices acoustiques et les règles, ou tout autre mécanisme de raisonnement, mettant en relation traits et indices. Ces indices seront évalués eux-mêmes à partir de paramètres acoustiques tels que les formants et leur évolution temporelle, les diverses énergies (totale, bandes de fréquence), les rapports entre les énergies hautes et basses fréquences, la fréquence fondamentale, etc...

En général, il est nécessaire de définir une hiérarchie sur les traits. En effet, certains d'entre eux sont plus stables et moins influencés par le contexte, et de ce fait sont plus facilement détectés; par contre, d'autres traits se caractérisent par une grande variabilité contextuelle (les indices du lieu d'articulation d'une occlusive par exemple). La figure 22 présente un tel système hiérarchisé.

Dans cette approche, un travail essentiel consiste à examiner une grande quantité de spectres et à rechercher les indices permettant de déceler la présence ou non d'un trait sur une partie du signal. L'apport des experts phonéticiens est évidemment capital.

Ces connaissances sont formalisées à l'aide d'outils propres à l'intelligence artificielle comme les systèmes experts, les ensembles flous, les systèmes à base de plans ou de frames, outils que nous développerons dans les paragraphes suivants. Les avantages potentiels de cette méthode de reconnaissance par traits sont nombreux, on retiendra la simplicité, l'universalité (tout phonème d'une langue peut être caractérisé par un ensemble de traits) et la dépendance plus faible vis à vis du locuteur. Les inconvénients sont quant à eux liés à la difficulté d'acquérir et de formaliser les connaissances nécessaires.

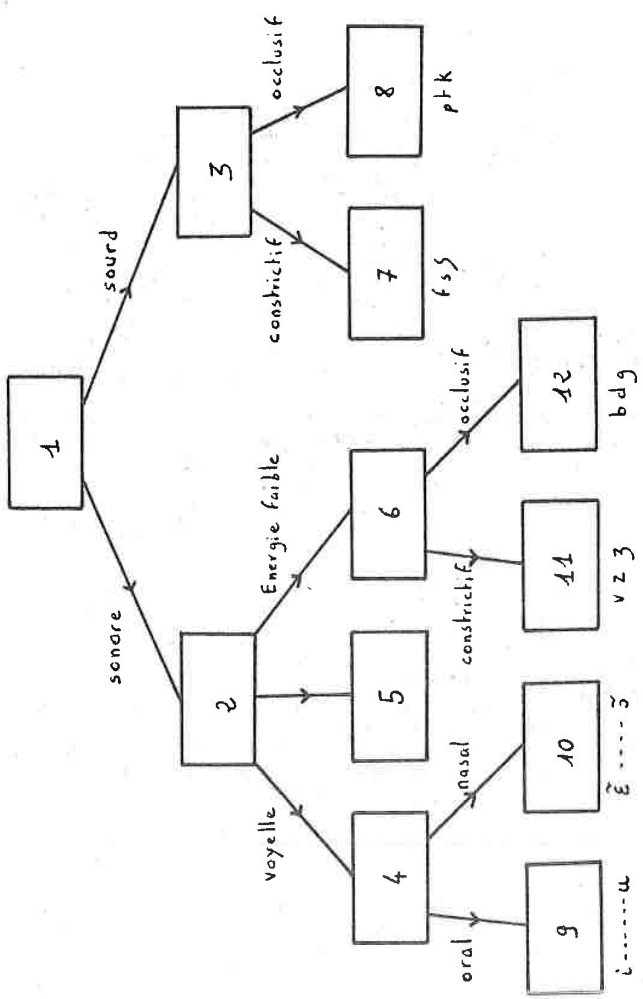


Figure 22 : Exemple de hiérarchie sur les traits

4.4.2 La théorie des ensembles flous

La complexité, l'imprécision, voire même le vague des définitions des concepts utilisés dans l'analyse de la parole ne permettent pas toujours de les traiter d'une manière exacte. La logique des mathématiques modernes s'applique parfaitement dans les sciences physiques rigoureusement définies, mais là où le comportement humain joue un rôle primordial, et c'est justement le cas de la communication parlée, les faits étudiés s'adaptent mal aux rigueurs d'une analyse stricte et trop précise.

Trois types d'inexactitude caractérisent l'analyse de la parole. Le premier est lié à la généralité, c'est à dire, un concept est appliqué aux divers éléments, même si ce n'est justifié que pour un nombre limité de cas. Citons comme exemple le point d'articulation qui est plus ou moins évident pour les occlusives, mais qu'on ne retrouve pas toujours pour les autres consonnes. Le second type d'inexactitude est lié à l'ambiguité que l'on rencontre dans la cas d'un terme ayant plusieurs significations, ou dans le cas d'une analyse de relations entre les indices acoustiques et les traits phonétiques. Le dernier type d'inexactitude concerne le caractère "flou" qui définit l'impossibilité de préciser les limites strictes des faits étudiés. En effet, il est bien connu que les relations entre les indices acoustiques et les traits phonétiques ne sont pas univoques et présentent un caractère plus ou moins approximatif. Dans ce cas, définir les limites strictes des valeurs de paramètres mesurés correspondant aux traits phonétiques donnés est une tâche presque impossible à réaliser. On ne peut que les spécifier d'une façon descriptive, ce qui d'ailleurs est fait, dans la plupart des cas, par l'homme qui préfère caractériser le monde extérieur en employant des mots tels que "grand", "petit", "moyen", "très grand", etc..., plutôt que des chiffres précis. On dit que les voyelles antérieures se caractérisent par une valeur élevée de la fréquence du deuxième formant au lieu de préciser les limites de variations de cette fréquence, par exemple, 1800 et 2800 Hz. Et c'est justement cette façon descriptive de caractériser les sons de la parole qui est pour l'homme tout à fait suffisante. Mais pour traiter de telles données à l'aide d'un calculateur, il faut les représenter sous forme numérique.

Ces trois types d'inexactitude sont traités d'une manière formelle et systématique par la théorie des ensembles flous qui tient compte, dans une certaine mesure, de l'évaluation subjective de l'homme [KAUF-73]. Cette théorie s'articule autour de la notion de "possibilité" [ZADE-78]. La possibilité est un indice permettant d'établir à quel point une chose est faisable, tandis que la probabilité est associée au concept de la fréquence de l'occurence. Ce qui est possible peut être improbable et ce qui est improbable ne doit pas être nécessairement non possible. A la différence de l'analyse statistique, dans le calcul des possibilités on peut largement tenir compte de l'état des connaissances à priori qui viennent de l'expérience humaine, et on est moins limité dans l'étude du phénomène.

L'exemple qui suit et qui est emprunté à [GUBR-82] illustre l'utilisation des ensembles flous pour formaliser les concepts de l'analyse phonétique. Considérons une classe de fréquences du deuxième formant, beaucoup plus grandes que 1000 Hz. Cet ensemble peut être défini de la manière suivante:

$$(4.14) \quad H = \left\{ F2 \;/\; F2 \gg 1000 \text{ Hz} \right\}$$

Il est évident que cet ensemble n'est pas bien défini, mais on peut le déterminer d'une manière subjective en utilisant une fonction d'appartenance dont l'allure est donnée par la figure 23 et définie comme suit:

$$(4.15) \quad h(F2) = \begin{cases} 0 & \text{pour } F2 < 1000 \text{ Hz} \\ \dfrac{F2 - 1000}{F2} & \text{pour } F2 > 1000 \text{ Hz} \end{cases}$$

La fonction h en mesurant le degré d'appartenance d'une fréquence quelconque F2 à H, quantifiera la possibilité qu'a F2 d'être beaucoup plus grande que 1000 Hz.

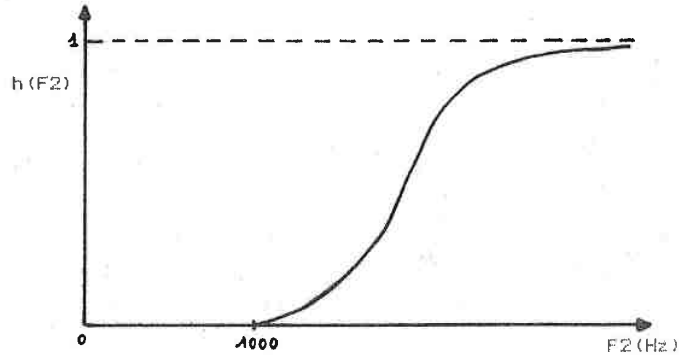


Figure 23

Ce formalisme utilisé en particulier par Wiezlak dans la reconnaissance de mots isolés [WIEZ-82] et par De Mori pour la compréhension de la parole continue [LAFA-80], se distingue des méthodes paramétriques par le fait que l'estimation de la possibilité d'un évènement ne nécessite pas la constitution d'un corpus de données aussi important que celui requis pour estimer des probabilités.

4.4.3 Les systèmes experts

Zue [ZUE-82] a montré que la représentation des signaux de parole par des spectrogrammes était très riche en informations phonétiques, et que des experts pouvaient transcrire phonétiquement ces spectrogrammes avec des taux de transcription de l'ordre de 70% à 85%. Il semblait donc naturel d'analyser et de modéliser cette expertise en lecture de spectrogrammes pour améliorer les algorithmes existants de décodage acoustico-phonétique, et la représentation et l'utilisation de l'expertise par une approche système expert s'imposait.

Les systèmes existants comprennent généralement:
- un ensemble de modules de pré-traitement dont le rôle est d'extraire du signal, pour chacun des segments de l'énoncé, des informations qui vont constituer la base de faits associée au segment;
- un ensemble de règles qui modélisent l'ensemble des connaissances nécessaires au domaine d'expertise: la base de connaissances;
- un interprète ou moteur d'inférences qui, à partir de la base de faits relative à un segment, des règles dont il dispose et qui formalisent le savoir de l'expert, et éventuellement de procédures d'analyse acoustique fine qu'il a la possibilité d'activer au cours de son raisonnement, détermine la nature exacte du segment.

Ce découpage du traitement correspond à la démarche de l'expert. Dans un premier temps, il jette un regard d'ensemble sur le spectrogramme pour évaluer certaines données comme la durée moyenne d'un segment par exemple. Ensuite, il commence le décodage proprement dit du continuum vocal, segment par segment, en s'appuyant sur des points d'ancrage constitués par les segments facilement identifiables et sur les éléments contextuels contenus dans les segments environnants.

Certaines caractéristiques de structures de contrôle sont communes à tout système expert [LAUR-82] & [GANA-85]:
- moteur d'inférences et représentation des connaissances de type formel (logique des propositions),
- stratégies par essais et erreurs: on peut revenir sur les choix et appliquer d'autres règles,
- raisonnement approximatif: lorsque plusieurs règles ont des conclusions communes, le poids de ces conclusions se trouve renforcé. On simule ainsi à la fois la pondération des conclusions exprimées par l'expert, et la notion de renforcement des hypothèses.

Par contre, la spécificité du problème de la reconnaissance de la parole nécessite la mise en oeuvre de techniques originales, par exemple:
- la gestion d'arbres de solutions partielles en parallèle,
- la prise en compte d'un raisonnement temporel.

Parmi les différents systèmes experts en décodage acoustico-phonétique réalisés à ce jour, on citera SONEX [MEMM-84] & [STER-85], SERAC [GILL-84], les systèmes de Johanssen [JOHA-83] et de Johnson [JOHN-84], ainsi que le système APHODEX développé dans notre laboratoire [CARB-84] & [CARB-85].

4.4.4 Autres formalismes

Les formalismes de représentation des connaissances inspirés par les recherches sur la logique résolvent bien le problème du raisonnement, mais sont le plus souvent un peu faibles du côté de la description des objets manipulés et des relations existant entre eux. Parallèlement à l'approche système expert, il s'est développé des systèmes de décodage phonétique, influencés notamment par des études psychologiques, qui s'intéressent plus à la représentation de l'univers dans lequel se situent les problèmes (objets, relations, structures), qu'aux règles portant sur cet univers lui-même [DEMO-83], [DEMO-84] & [GREE-84]. Ils se rapprochent de ce fait des modèles développés par les spécialistes de bases de données. C'est dans ce contexte qu'ont été entrepris les travaux qui vont être détaillés dans les chapitres suivants et qui ont abouti à la réalisation d'un système de décodage acoustico-phonétique dont les structures de données et de contrôle s'articulent autour d'un formalisme à base de frames.

CHAPITRE 5
LES FRAMES, UN OUTIL POUR AIDER A RESOUDRE LE DECODAGE PHONETIQUE

5.1 Intoduction - Motivations

La compréhension de la parole est, pour l'essentiel, une activité humaine qui résiste à l'analyse, ce qui explique en partie les performances médiocres des systèmes monolocuteur actuels de compréhension d'énoncés prononcés de manière naturelle, ainsi que les difficultés supplémentaires rencontrées lorsqu'on impose au système d'accepter en entrée des locuteurs différents moyennant une phase d'apprentissage limitée.

Ces performances médiocres sont souvent dues à un mauvais décodage acoustico-phonétique qui peut s'expliquer par l'utilisation de méthodes de "pattern matching" qui sont peu adaptées à la reconnaissance multilocuteurs. Il semble donc nécessaire d'améliorer les performances des systèmes de décodage phonétique si l'on désire obtenir des systèmes de dialogue homme-machine efficaces.

Une solution intéressante consiste comme nous l'avons vu au chapitre précédent à utiliser les traits phonétiques pour décrire les phonèmes et à rechercher l'ensemble des indices acoustiques susceptibles de représenter un trait; et cette approche "intelligente" du problème apparâit comme très prometteuse [KLAT-79], [LEA-80] et [DEMI-83]. Dans le cadre du projet APHODEX mené conjointement par le CRIN et l'Institut de Phonétique de NANCY, l'amélioration des algorithmes de décodage acoustico-phonétique existants est réalisée à partir de la modélisation du savoir-faire d'un expert en lecture de spectrogrammes, F. Lonchamp. De ce fait, il semblait naturel d'opter pour une approche de type système expert dans lequel la base de connaissances phonétiques consiste en un ensemble de règles de production [CARB-84] et [CARB-85]. Cependant, l'implantation d'un système expert en reconnaissance de la parole pose des problèmes d'efficacité dans l'état actuel de la technique. En particulier, les contraintes de fonctionnement en temps réel sont quasi-impossibles à respecter, à la fois pour des raisons de puissance intrinsèque des machines (nombre d'inférences logiques par seconde), et pour des raisons

plus profondes concernant les principes mêmes de tels systèmes (contrôle de l'évolution temporelle de la base des faits entre autres).

L'indéterminisme important qui caractérise la reconnaissance de la parole continue implique l'utilisation de raisonnements approximatifs faisant largement appel au contexte et fonctionnant à plusieurs niveaux, ainsi que la possibilité de suivre plusieurs lignes de raisonnement simultanément, et de les abandonner si nécessaire. Enfin, une liaison efficace est indispensable entre le système de raisonnement et le monde physique dans lequel sont prélevées les données. Les faits sur lesquels s'appuie le raisonnement sont déduits de ces données par des procédures éventuellement complexes qu'il est souhaitable d'attacher aux connaissances. Tout cela conduit au développement de solutions originales: structures de contrôle manipulant des arborescences de solutions, représentations objets. Un système à base de connaissances fondé sur une représentation par "frame" [MINS-75], permet d'utiliser, en ce qui concerne le décodage phonétique, différents types d'analyses acoustiques et de contrôler le processus de reconnaissance par un système de plans d'action. Il facilite en outre le raisonnement à différents niveaux (i.e. la mise en oeuvre de connaissances et de méta-connaissances), et permet de structurer les connaissances en fonction des objets manipulés (par exemple les phonèmes).

En complément au projet APHODEX, nous avons développé un processus de décodage phonétique qui utilise ce formalisme à deux fins:
- pour représenter les connaissances et résultats issus du projet APHODEX,
- pour formaliser les stratégies complexes de décodage mises en oeuvre par l'expert; ces méta-connaissances sont regroupées dans une grammaire de frames constituée de règles de réécriture.

Le décodage phonétique est donc représenté par un langage défini sur des frames. Des approches similaires ont été adoptées par De Mori [DEMO-83] pour l'extraction de traits acoustiques, et dans un cadre plus général, par Green & Wood [GREE-84] avec le concept de "Speech Sketch" pour le décodage phonétique.

5.2 Description du formalisme

5.2.1 Principes de base

Le concept de frames est une structure déclarative de connaissance qui a été initialement introduite dans le traitement des langages naturels par Minsky [MINS-75]. Un "frame" est référencé par un nom et composé d'un ensemble de caractéristiques désignées sous le terme de cases ou "slots". Un slot est un support d'information concernant une connaissance élémentaire appelée "slot-filler" qui peut correspondre soit à un évènement, soit à une relation ou alors au résultat d'une procédure. Un frame représente un objet ou une situation typiques dans l'univers où l'on travaille.

Dans notre système, un frame correspond à la réalisation acoustique d'une unité phonétique ou d'un segment du continuum vocal, et les slots aux traits acoustiques qui caractérisent ces évènements phonétiques. L'instanciation des slots, qui correspond à la recherche des traits acoustiques, est liée à l'attachement procédural: des procédures sont associées à chaque slot dans le but de préciser au système les opérations à réaliser sur le signal vocal, opérations qui décrivent le type de représentation paramétrique du signal utilisée et la nature des indices acoustiques à extraire. Nous y reviendrons plus en détail dans le paragraphe suivant.

Tous les frames ne sont pas référencés dans le système par un nom. En fait, seuls les frames qui correspondent à un prototype d'une unité phonétique le sont; les différents frames temporaires créés pendant le processus de reconnaissance restent quant à eux "anonymes", et sont connus par le système sous le terme d'INSTANCE par opposition aux frames qui décrivent la réalisation type d'une unité phonétique appelés PROTOTYPES.

Les INSTANCES représentent un évènement acoustique (segment) à analyser et identifier. Les slots qui les caractérisent sont définis et instanciés au cours du processus de décodage sous contrôle d'un système organisé et structuré autour d'une grammaire de frames. Ce système de contrôle, qui contient les règles et

méta-connaissances nécessaires à l'analyse acoustique et au décodage phonétique, oriente et adapte le processus de l'analyse acoustique au type d'unité phonétique étudié. Ainsi, par exemple, si l'analyse des indices acoustiques déjà extraits semble indiquer que le segment décodé est une fricative sourde, les fréquences formantiques de l'unité phonétique ne semblent pas des indices très appropriés pour affiner l'identification de cette fricative.

Les PROTOTYPES caractérisent les réalisations les plus représentatives des unités phonétiques reconnues par le système. Chaque PROTOTYPE est constitué de slots qui regroupent les traits acoustiques typiques d'une unité, traits déterminés à partir des connaissances de l'expert phonéticien. Les slots d'un PROTOTYPE sont donc définis à priori par l'expert contrairement aux slots d'une INSTANCE qui sont instanciés par des procédures (slots-filler) au cours du processus de reconnaissance.

La figure 27 présente un exemple de chacun de ces deux types de frames pour le burst du phonème /p/.

PROTOTYPE /p/ burst		INSTANCE /p/ burst	
Voisement	0	Voisement	0
Durée	<40 ms	Durée	35 ms
Pic(s) d'énergie	[750Hz-2500Hz]	Pic(s) d'énergie	1500 Hz
Energie	<100	Energie	67

Figure 27
Exemples de frames PROTOTYPE et INSTANCE
pour le phonème /p/

5.2.2 L'attachement procédural

Bien que la représentation et l'organisation des connaissances dans les frames soient de type déclaratif, il est nécessaire de disposer dans de tels systèmes, de mécanismes de liaison efficaces entre les modules de raisonnement et le monde physique dans lequel sont prélevées les données: c'est l'attachement procédural. Dans notre système, l'attachement procedural constitue le mécanisme principal de contrôle du processus de reconnaissance par l'intermédiaire de procédures qui sont associées aux slots afin de dirriger et d'orienter la construction des frames INSTANCES.

On peut classer ces procédures en deux grandes classes, les "instancieurs" et les "sélecteurs" qui se réfèrent en grande partie aux "servants" et et "demons" du langage KRL [BOBR-77a]. Les instancieurs sont des procédures qui sont activées automatiquement par le système quand un indice acoustique est à extraire du signal (instanciation du slot correspondant). Par contre, les sélecteurs, qui sont activés uniquement sur demande, sont des fonctions de décision qui permettent de sélectionner l'ensemble des traits acoustiques caractéristiques du segment décodé. Le frame PHONEME proposé sur la figure 28 illustre l'utilisation de ces deux types de procédures. Ainsi, sur les slots VOYELLE, PLOSIVE et FRICATIVE, on trouve plusieurs instancieurs référencés par les mots-clé FILL...USING, et des sélecteurs précédés et identifiables per les mots-clé WITH...SELECT et OTHERWISE. Lorsque le frame PHONEME sera interprété, les instancieurs P-NOYAUX-VOCALIQUES, P-PLOSIVES et P-FRICATIVES seront automatiquement activés par le système afin de réaliser une pré-classification phonétique du segment analysé: voyelle, plosive, fricative ou autre type de phonème.

Lorsque les slots VOYELLE,PLOSIVE et FRICATIVE sont instanciés, les sélecteurs invoqués par les mots-clé SELECT et OTHERWISE vont orienter la suite du processus de décodage sur les frames PHON-VOYELLE, PHON-PLOSIVE, PHON-FRICATIVE et AUTRE-PHONEME en fonction de la pré-classification phonétique qui vient d'être réalisée.

Le système s'appuiera alors sur le frame invoqué pour continuer à créer et instancier les slots du frame INSTANCE courant.

```
Frame PHONEME
      Specialization-of ROOT
      Type Subframe
      Fill VOYELLE   Using P-NOYAUX-VOCALIQUES
      Fill PLOSIVE   Using P-PLOSIVES
      Fill FRICATIVE Using P-FRICATIVES
      With VOYELLE   Select PHON-VOYELLE
      With PLOSIVE   Select PHON-PLOSIVE
      With FRICATIVE Select PHON-FRICATIVE
      Otherwise AUTRE-PHONEME
End
```

Figure 28
Exemples d'instancieurs et de sélecteurs

En complément à ces deux formes d'attachement procédural que sont les instancieurs et les sélecteurs, notre système dispose d'une troisième catégorie de routines qui ont pour objet de vérifier la cohérence du raisonnement pendant le processus de reconnaissance: ce sont les contrôleurs. Par exemple, le contrôleur SPECIALIZATION-OF précise les liens devant exister entre les différents indices acoustiques extraits du signal et associés aux slots du frame INSTANCE courant.

5.3 Le raisonnement dans le processus de reconnaissance

Le processus de reconnaissance peut se résumer en deux phases successives. La première étape consiste à extraire les indices acoustiques pertinents du signal afin de construire le frame INSTANCE associé à l'unité phonétique étudiée. Durant la seconde étape appelée "Matching", un score de compatibilité est calculé entre le frame INSTANCE et les frames PROTOTYPES connus par le système. L'unité phonétique sera étiquetée par le

prototype qui aura obtenu le meilleur score. Toutefois, si aucun score n'atteint un seuil fixé à priori, le système peut décider de remettre en cause un ou plusieurs indices acoustiques associés aux slots du frame INSTANCE, et de réaliser de nouvelles analyses du signal plus pertinentes. Ce type de stratégie assure une meilleure identification des phonèmes analysés, les décisions quant à l'étiquetage des unités s'appuyant sur des bases de faits consistantes et cohérentes.

5.3.1 Le matching , moteur du processus de reconnaissance

Le raisonnement est dominé par un processus de reconnaissance dans lequel de nouveaux objets et évènements sont comparés à un ensemble de prototypes candidats connus par le système [MINS-75], [SCHAN-77] et [MOOR-73]. La partie clé de ce processus de reconnaissance est un comparateur de descriptions ou "matcher" qui sert de moteur d'inférences pour comparer les descriptions contenues dans les frames.

Dans notre système, le matcher compare deux entités à la fois: un frame PROTOTYPE et une forme spécifique à identifier, le frame INSTANCE. Or, ces deux objets s'articulent sur une structure interne complexe qui complique l'opération de matching. Cette complexité s'explique par la nature même du frame qui est divisé en un ensemble de slots, et par la syntaxe qui est utilisée pour décrire les slots-filler. Si les descripteurs du frame INSTANCE sont constitués uniquement de constantes numériques, il en est tout autrement pour les descripteurs des prototypes où la présence d'un ensemble de terminaux (=,<,>,#,[,-,]) esquisse un embryon de grammaire pour décrire l'information contenue dans les slots (exemple figure 27). La grammaire correspondante est relativement primaire puisque le nombre de terminaux est réduit (une dizaine), et qu'elle interdit toute imbrication et association des descripteurs entre eux.

Du fait de la complexité des structures manipulées, il est donc nécessaire de découper la comparaison en plusieurs sous-tâches, chacune d'elles ayant pour but de comparer l'information contenue dans un

des slots d'un frame PROTOTYPE avec le slot-filler correspondant du frame INSTANCE. L'utilisation d'outils syntaxiques pénalise l'utilisation de ces sous-tâches, et il est souhaitable de limiter le nombre de prototypes candidats au matching.

La solution proposée consiste à hiérarchiser les différents indices acoustiques utilisés pour caractériser les phonèmes, et d'orienter la construction du frame INSTANCE (choix des slots à instancier) en fonction de cet ordonnancement. Le processus de matching est alors déclenché au niveau des éléments terminaux de la hiérarchie. De ce fait, on opère un pré-décodage phonétique en ne retenant qu'un nombre limité de prototypes candidats d'une part, et on est assuré de retrouver une certaine homogénéité et cohérence entre les slots des frames INSTANCE et PROTOTYPE d'autre part, d'où une optimisation du fonctionnement du matcher.

Le résultat du processus de matching est fonction du pré-décodage réalisé pendant la construction du frame INSTANCE. Il ne peut se résumer au simple choix binaire "succès" ou "échec", qui, s'il est souvent bien adapté au traitement des langages naturels [BOBR-77b], [CHAR-78], ne peut s'appliquer de façon réaliste au domaine de la parole du fait de l'indéterminisme des objets manipulés (problèmes de segmentation et phénomènes contextuels entre autres). La méthodologie retenue est empruntée au raisonnement approximatif, et consiste à associer un score ou un coefficient de plausibilité aux conclusions, prennant en compte le degré de cohérence entre les slots des deux frames comparés. L'étiquetage est alors réalisé de la façon suivante:

- si plusieurs candidats ont un coefficient de plausibilité supérieur à un seuil fixé à priori, un processus de décision majoritaire est appliqué pour déterminer les trois phonèmes retenus pour identifier le segment analysé;
- si aucun score n'atteint le seuil fixé, le système peut décider soit de remettre en cause un ou plusieurs indices acoustiques associés aux slots du frame INSTANCE

et de réaliser de nouvelles analyses du signal plus pertinentes, soit de conclure qu'il ne dispose pas des outils nécessaires pour lever l'ambiguité issue du processus et d'affecter une ou plusieurs étiquettes par défaut au segment étudié.

5.3.2 Le contrôle du décodage : une grammaire de frames

Tout le système de contrôle du décodage phonètique est formalisé sous forme d'une grammaire de frames. Cette grammaire regroupe aussi bien l'extraction des indices acoustiques du signal et la construction du frame INSTANCE que les informations qui structurent le processus et l'étiquetage du signal. Une étude plus détaillée du système sera exposée dans le paragraphe suivant; nous nous contenterons de décrire dans cette section le formalisme du langage de contrôle.

La grammaire de frames consiste en un ensemble de règles de réécriture décrites sur la figure 29. On y retrouve les principes émis dans KRL [BOBR-77a], tant au niveau des structures de données (frames INSTANCE et PROTOTYPE) qu'au niveau des structures de contrôle, et en particulier l'attachement procédural et le processus de matching. Le formalisme proposé correspond à une synthèse entre la grammaire proposée par De Mori pour déterminer les traits acoustiques présents dans un continuum de parole [DEMO-83], et les langages à base de frames utilisés dans le traitement des langages naturels illustrés par le système GUS [BOBR-77b]. Notre grammaire contient actuellement 7 descripteurs de base. Ce nombre peut paraître très petit en comparaison de la richesse des langages à base de frames rencontrés dans le traitement des langages naturels ou en vision. Ceci peut s'expliquer par la relative simplicité des objets qui sont manipulés (unités phonétiques ou unités discrètes de parole), et des descriptions et attributs qui leur sont associés (indices acoustiques). Il est ainsi possible de travailler directement sur la structure d'entrée, contrairement aux domaines précités où il est nécessaire d'effectuer un ou plusieurs traitements préalables sur la structure d'entrée (analyses lexicale, syntaxique et sémantique dans le traitement des langages naturels).

```
                                                          K≥1
‹FRAME›          := ‹FRAME-HEADING› |(‹ORDER-LIST›) | ‹PERIOD›

‹FRAME-HEADING›:= ( Frame ‹NAME›)

‹ORDER-LIST›     := ( Specialization-of ‹NAME›)
                 := ( Type ‹TYPE-SORT›)
                 := ( Fill ‹NAME› Using ‹PROC›)
                 := ( Match With ‹PROTOTYPES›)
                 := ( Use-Ref From ‹NAME›)                K≥1
                 := ( With ‹EXPRESSION› Select ‹FRAME›)
                    ( Otherwise ‹FRAME›)

‹TYPE-SORT›      := Subframe
                 := Terminal

‹PROC›           := F-‹function›
                 := P-‹procedure›

‹PROTOTYPES      := (‹NAME› [ , ])

‹EXPRESSION›     := (‹PREDICAT›)
                 := ([ ( ] ‹EXPRESSION› ‹BOOL-OP› ‹EXPRESSION› [ ) ])

‹PREDICAT›       := ([ ( ] ‹NAME› ‹TEST-OP› ‹NAME› [ ) ])

‹BOOL-OP›        := And / Ior / Xor

‹TEST-OP›        := = / › / ‹ / #

‹PERIOD›         := End
```

Figure 29
Les règles de réécriture de la grammaire de frames

On peut classer les descripteurs du langage de contrôle en trois catégories:

- les "constructeurs" qui fabriquent les frames INSTANCES: FILL pour l'extraction des traits acoustiques; SELECT pour sélectionner les indices pertinents et orienter le choix des slots caractérisant chaque frame INSTANCE élaboré.

- les descripteurs décisionnels pour guider le choix de l'étiquetage: MATCH pour sélectionner les prototypes candidats; DEFAULT pour affecter une ou plusieurs étiquettes au segment étudié en cas d'ambiguité lors du processus; USE-REF qui indique la nécessité de faire appel à des techniques de reconnaissance des formes pour l'étiquetage. En effet, il peut arriver qu'une classe phonétique ne soit pas suffisamment caractérisée par un ensemble de traits acoustiques cohérent pour pouvoir lui associer un frame PROTOTYPE. Dans de tels cas, l'étiquetage est réalisé en comparant la représentation acoustique du segment étudié avec un ensemble de formes de référence précisé par le descripteur USE-REF. La flexibilité de l'approche par frames permet d'implanter facilement cette interaction entre techniques d'intelligence artificielle et de reconnaissance des formes.

- les descripteurs vérificatifs ou "méta-descripteurs" qui contrôlent le processus de décodage: TYPE pour préciser le niveau de décodage (SUBFRAME pour la phase construction du frame INSTANCE et TERMINAL pour invoquer le processus de matching); SPECIALIZATION-OF pour indiquer les liens possibles entre les différents traits acoustiques (ces liens sont particulièrement importants en cas de retour arrière).

Chacun de ces descripteurs est décrit de façon plus détaillée et illustré par des exemples sur la figure 30. Actuellement, le nombre des descripteurs est de 7, mais la modularité du système permet une totale évolutivité du langage.

Descripteur : FILL <Name> USING <Proc>
Attributs : un nom de slot et un identificateur de fonction ou de procédure
Utilisation : création et instanciation de slots
Exemples : FILL Voisement USING F-Calc-Pitch
FILL Voyelle USING P-Noyaux-Vocal

****

k>0

Descripteur : (WITH <Expression> SELECT <Frame>) OTHERWISE <Frame>
Attributs : un prédicat ou une conjonction de prédicats et des identificateurs de frames
Utilisation : sélection des indices caractérisant le frame INSTANCE
Exemples : WITH (Slot1 > Slot2) AND (Slot3 = 4)
OR (Slot4 > 50) SELECT Solution1
WITH (Slot1 = Slot2) AND (Slot3 # 4)
OR (Slot5 = 0) SELECT Solution2
OTHERWISE Solution3

****

Descripteur : MATCH WITH <Prototype>
Attributs : des identificateurs de frames PROTOTYPES
Utilisation : sélection des prototypes candidats au processus de matching
Exemples : MATCH WITH Proto-k,Proto-g
MATCH WITH Proto-m,Proto-n,Proto-gn

****

Descripteur : DEFAULT <String>
Attributs : une liste d'identification de phonèmes
Utilisation : étiquetage par défaut du segment étudié en cas d'ambiguité lors du processus de matching
Exemple : DEFAULT 'b-v'

Figure 30

Descripteur : USE-REF FROM <Name>
Attributs : un identificateur de vocabulaire de références
Utilisation : étiquetage par techniques de reconnaissance des formes entre la représentation acoustique du segment étudié et un ensemble de formes de références
Exemples : USE-REF FROM Fichier_Nasales
USE-REF FROM Contextes_lR

****

Descripteur : TYPE <Type-Sort>
Attributs : SUBFRAME ou TERMINAL
Utilisation : indication du niveau de décodage du segment analysé, SUBFRAME pour la phase constitution du frame INSTANCE et TERMINAL pour invoquer le processus de matching

****

Descripteur : SPECIALIZATION-OF <Name>
Attributs : un identificateur de frame
Utilisation : contrôle de la cohérence des traits acoustiques caractérisant le frame INSTANCE
Exemples : 1) FRAME Phonèmes_pbkg
SPECIALIZATION-OF Phonèmes_plosifs
2) FRAME Phon_1_contexte_y
SPECIALIZATION-OF Phonème_1
3) FRAME Voyelles_nasales
SPECIALIZATION-OF Voyelles

Figure 30
(suite et fin)

5.4 Une double approche de la reconnaissance phonétique

5.4.1 L'approche ascendante : le décodage acoustico-phonètique

C'est l'approche classique de la reconnaissance phonétique. Après avoir dans une première étape segmenté le continuum de parole en unités syllabiques puis phonétiques, le système détermine les traits acoustiques caractérisant chaque segment en appliquant hiérarchiquement un ensemble de régles phonétiques. La description acoustico-phonétique ainsi obtenue constitue le frame INSTANCE associé au segment considéré. Parallèlement, le système sélectionne un sous-ensemble de classes phonétiques qui sont acoustiquement les plus proches possibles de la forme inconnue. Ce sous-ensemble correspond aux frames PROTOTYPES. Durant le processus, le frame INSTANCE est comparé slot par slot aux différents PROTOTYPES retenus et un score de compatibilité est calculé pour chacun d'eux. L'étiquetage du segment est fonction des scores obtenus: si un ou plusieurs scores sont supérieurs à un seuil fixé à priori, un processus de décision majoritaire déterminera les phonèmes retenus dans le treillis phonétique; dans le cas contraire le système choisira de remettre en cause certains indices acoustiques (retour arrière à l'aide des descriptions indiquées par SPECIALIZATION-OF) et de réaliser de nouvelles analyses du signal plus pertinentes, ou alors d'affecter une ou plusieurs étiquettes par défaut (descripteur DEFAULT) au segment analysé, en fonction des scores de plausibilité issus du processus de matching.

Cette méthode est particulièrement efficace lorsqu'il n'y a pas d'erreurs de segmentation, comme dans le cas du traitement des langages naturels qui présentent des séparateurs évidents (ponctuation). Malheureusement, la situation est tout à fait différente en reconnaissance de la parole où les erreurs de segmentation sont inévitables. Nous proposons ici une nouvelle approche du problème de la segmentation qui est en étroite relation avec le formalisme utilisé, les frames. Une première segmentation du continuum vocal en unités phonémiques est réalisée sur la base de solides critères phonétiques. Dans le cadre du projet APHODEX, nous avons utilisé avec succés les noyaux vocaliques, les fricatives et les

plosives comme critères de segmentation. Cette décomposition en unités phonémiques est appliquée sur une des représentations paramétriques du signal, et est ensuite propagée aux autres représentations (notre ambition est de travailler sur la base d'analyses multi-critères). Les frontières ainsi déterminées ne sont pas définitives, elles peuvent être modifiées durant le processus de reconnaissance par les procédures associées aux indices acoustiques, i.e. les slots des frames. Cette étroite interaction entre règles, frames et procédures donne une méthode "segmentation par reconnaissance" efficiente.

Le décodage acoustico-phonétique peut finalement se résumer en trois phases succéssives et complémentaires:

- une segmentation par règles du continuum vocal en unités phonétiques;
- une caractérisation de chaque segment par un ensemble de traits acoustiques, pilotée par une grammaire de frames;
- un étiquetage des segments (et la modification des frontières établies lors de la segmentation si nécessaire) placé aussi sous le contrôle de la grammaire de frames.

5.4.2 L'approche descendante : la vérification phonétique

Dans la plupart des systèmes de reconnaissance de la parole continue actuels, l'information issue du processus de décodage acoustico-phonétique est considérée comme définitive et n'est plus remise en cause dans les modules de niveau supérieur. De ce fait, certaines connaissances d'ordre phonologique, prosodique ou syntaxique qui laissent supposer la présence d'insertions, d'élisions ou de substitutions dans le treillis phonètique restent ignorées, et faute de pouvoir être vérifiées et corrigées, ces erreurs sont propagées tout au long du processus de reconnaissance. Ceci est particulièrement vrai pour les mots monosyllabiques qui servent d'unités de liaison entre mots ou groupes de mots (conjonctions "et", "ou", pronoms "où", "en", "y", préposition "à", etc...). Caractérisés par une brièveté d'élocution très marquée, ils échappent quelquefois au

décodage acoustico-phonétique, et ils n'ont de chance d'être repérés dans le continuum vocal que par l'apport de connaissances d'ordre supérieur.

Cet aspect important du dialogue entre décodeur et modules de niveau supérieur est intégré à notre système. Lorsque certains éléments du treillis phonétique sont contestés par l'apport de connaissances de niveau supérieur, une vérification des hypothèses émises est demandée au module de décodage. Ces hypothèses concernent principalement les élisions et les substitutions de phonèmes, le problème des insertions étant traité par l'analyseur lexical. Mais contrairement au processus de décodage classique, il n'y a pas de phase de recherche de prototypes candidats puique le système connait à l'avance l'identité du phonème pressenti pour étiqueter le segment litigieux, et par là-même, les traits acoustiques qui vont caractériser le frame INSTANCE. Dans le processus de vérification, le but est de créer une instance du frame PROTOTYPE associé au phonème omis ou substitué. Pour ce faire, le système modifie sa stratégie d'application des règles phonétiques. Le contrôle en est toujours assuré par la grammaire de frames, mais le parcours des noeuds de la grammaire est inversé: le point d'entrée du système est maintenant le frame de type TERMINAL correspondant à la classe phonétique recherchée, et le pilotage de la construction du frame INSTANCE (création et instanciation des slots) est réalisé par chaînage arrière vers la racine par l'intermédiaire des descripteurs SPECIALIZATION-OF. L'instance ainsi créée, et le frame PROTOTYPE sont alors comparés par le matcher qui confirmera ou invalidera via le score de compatibilité la présence du phonème présumé dans le segment de signal analysé.

Un exemple d'utilisation de la vérification phonétique concerne l'accés lexical dans les systèmes de reconnaissance de grands vocabulaires. Dans le système de reconnaissance de 2000 mots fonctionnant actuellement dans notre laboratoire [MARI-84], [NOIR-85], deux processus sont menés en parallèle:
- la transcription phonétique du mot à reconnaitre sous forme d'une chaîne de phonèmes à réponses multiples,

- la description d'un mot en termes de classes phonétiques et la sélection d'une cohorte de mots satisfaisant à cette description.

Bien que ces deux processus interagissent fortement tant pour la segmentation que pour la caractérisation acoustico-phonétique du mot par l'intermédiaire de déductions effectuées par un système expert de décodage phonétique, il peut arriver que la description du mot en classes phonétiques diffère de la transcription phonétique par un ou plusieurs phonèmes. Par exemple, un /b/ peut être décodé comme plosive sourde ou un /R/ comme un noyau vocalique. Le module de reconnaissance du mot a alors peu de chances de trouver un candidat satisfaisant parmi la cohorte de mots sélectionnée à partir de la description en classes; et dans ce genre de situations, il est intéressant de disposer d'un module de vérification phonétique pour déterminer l'origine des erreurs, transcription phonétique ou description en classes, et effectuer les corrections adéquates.

5.4.3 Un exemple d'utilisation des frames

Afin de mieux évaluer l'intérêt des frames en décodage acoustico-phonétique de la parole, nous allons illustrer leur contribution dans la reconnaissance des plosives: /p/, /t/, /k/, /b/, /d/, /g/.

Les indices acoustiques utilisés généralement pour caractériser les plosives sont: le voisement, la durée de l'occlusion, l'énergie du burst, l'énergie spectrale et l'évolution des formants sur les segments contigus. Le tableau 31 présente le domaine de variation de ces indices pour chaque plosive tels qu'ils ont été déterminés dans le cadre du projet APHODEX.

Les valeurs mentionées sont utilisées pour construire les frames PROTOTYPES correspondants. On pourra noter que les frames intègrent les règles de l'expert pour le suivi des transitions des formants. Il nous faut maintenant écrire les règles de la grammaire de frames qui vont caractériser les frames INSTANCES dans le cas où le segment de parole laisserait supposer que c'est une plosive. Ceci revient à introduire une hiérarchie

	Durée	Pics d'énergie (KHz)	Evolution des Formants	Energie	Voisement
"p"	$< D_1$	[0.75, 2.5]	Règle 1	$< E_1$	Non
"b"	$< D_1$	[0.75, 2.5]	Règle 1	$< E_1$	Oui
"g"	$[D_2, D_3]$	[0.75, 2.5]	Règle 3	$> E_2$	Oui
"k"	$> D_3$	[0.75, 2.5]	Règle 3	$> E_2$	Non
"t"	$[D_1, D_2]$	>2.5	Règle 2	$[E_1, E_2]$	Non
"d"	$< D_2$	[4,5]	Règle 2	$[E_1, E_2]$	Oui

Tableau 31

Traits acoustiques caractérisant les sons plosifs

sur les indices à extraire du continuum vocal. En ordonnant les indices de la façon suivante: caractère plosif, pics d'énergie, énergie, transitions des formants, durée et voisement, le nombre de prototypes candidats au matching sera minimisé.

Il existe plusieurs manières de formaliser cet ordonnancement dans le langage de frames. Nous avons adopté celle-ci:

```
Frame PHONEME
      Specialization-of ROOT
      Type Subframe
      Fill VOYELLE    Using P-NOYAUX_VOCALIQUES
      Fill PLOSIVE    Using P-PLOSIVES
      Fill FRICATIVES Using P-FRICATIVES
      With VOYELLE    Select PHON-VOYELLE
      With PLOSIVE    Select PHON-PLOSIVE
      With FRICATIVES Select PHON-FRICATIVE
      Otherwise PHON-AUTRES
End


Frame PHON-PLOSIVE
      Specialization-of PHONEME
      Type Subframe
      Fill PICSENERGIE Using F-RECHERCHE_PICS
      Fill ENERGIE     Using P-CALC_ENERGIE
      With (PICSENERGIE>2.5) And (ENERGIE>E1) And
           (ENERGIE<E2) And (SUIVIFORMANTS=Règle2)
           Select PHON-td
      Otherwise PHON-pbkg
End


Frame PHON-td
      Specialization-of PHON-PLOSIVE
      Type Terminal
      Fill DUREE      Using F-CALC_DUREE
      Fill VOISEMENT Using F-CALC_PITCH
      Match With PROTOTYPE-t,PROTOTYPE-d
End
```

Les frames et le décodage phonétique

```
Frame PHON-pbkg
      Specialization-of PHON-PLOSIVE
      Type Subframe
      Fill DUREE Using F-CALC_DUREE
      With (DUREE<D1) And (ENERGIE<E1)
           Select PHON-pb
      Otherwise PHON-kg
End


Frame PHON-pb
      Specialization-of PHON-pbkg
      Type Terminal
      Fill VOISEMENT Using F-CALC_PITCH
      Match With PROTOTYPE-p,PROTOTYPE-b
End


Frame PHON-kg
      Specialization-of PHON-pbkg
      Type Terminal
      Fill VOISEMENT Using F-CALC_PITCH
      Match With PROTOTYPE-k,PROTOTYPE-g
End
```


CHAPITRE 6
L'IMPLANTATION DU SYSTEME : LFRAP

6.1 Description générale du système

Le système de reconnaissance acoustico-phonétique que nous avons implanté et baptisé LFRAP (Langage de Frames pour la Reconnaissance Acoustico-Phonétique), est constitué de modules correspondant chacun à une étape importante de l'analyse: traitement acoustique du signal et extraction des paramètres, segmentation, étiquetage et apprentissage. La figure 32 présente l'organisation générale de LFRAP; les flèches indiquent le sens du flux d'information entre les différents modules. Le résultat final est un treillis phonétique groupant les divers choix possibles pour étiqueter chaque segment.

Chaque module utilise les données et les hypothèses élaborées durant les étapes précédentes, en les corrigeant si nécessaire, dans le but de faire progresser la reconnaissance phonétique. Ainsi, la procédure d'identification détermine dans une première étape l'ensemble des traits acoustiques caractérisant chaque segment en appliquant un ensemble de règles phonétiques (module "hypothétiseur"), puis dans un deuxième temps, étiquette chaque segment en fonction des traits acoustiques déterminés par l'hypothétiseur (module "matcher"). Cet ensemble de règles est aussi utilisé pour la segmentation où une pré-classification en catégories phonétiques est réalisée (noyaux vocaliques, plosives et fricatives). Voici des exemples de règle employée:

R1 : Si (Occlusive=vrai) et (3000Hz<fréquence de l'énergie maximum du burst<4500Hz) et (contexte droit /y-u-w/)
alors /t/

R2 : Si (Occlusive=vrai) et (2500Hz<fréquence de l'énergie maximum du burst<3500Hz) et (contexte droit /i-e/)
alors /k/

Ces règles, qui constituent la base de connaissances phonétiques de notre système, ont été déduites des connaissances et du savoir-faire de F. Lonchamp en lecture de spectrogrammes, et sont actuellement testées par D. Fohr sur un corpus de cent

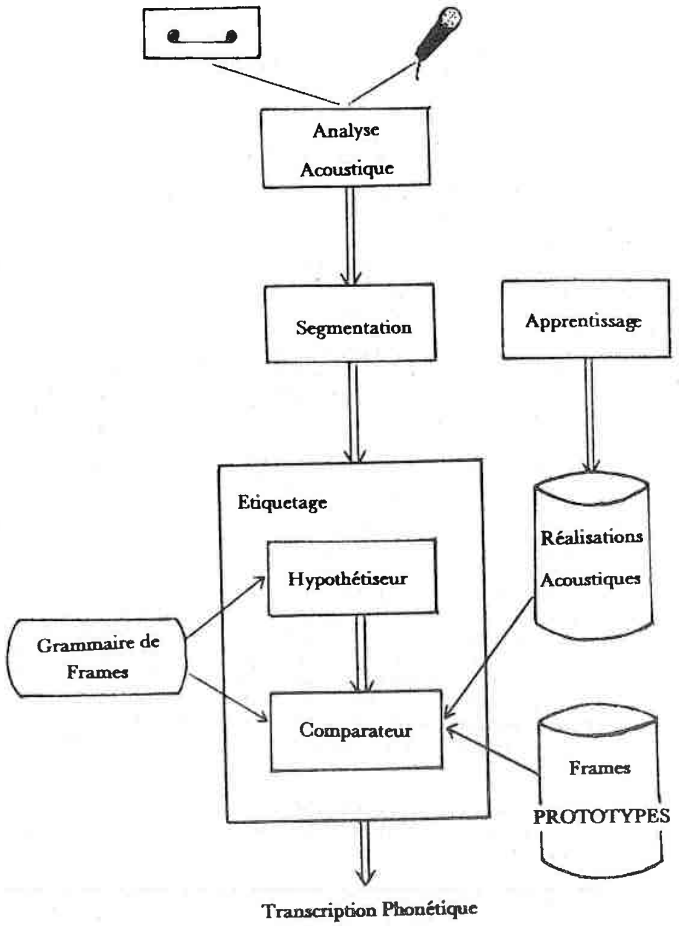


Figure 32

Organisation générale du système

phrases différentes (empruntées au corpus de Combescure) prononcées à un rythme naturel par dix locuteurs masculins non entraînés. On trouvera dans [CARB-83] et [CARB-84] une description détaillée de la méthodologie utilisée pour acquérir l'expertise.

Notre système, pour l'instant est implanté en PASCAL sur un mini-calculateur EXORMACS à base de microprocesseur 68000.

6.2 Analyse acoustique du signal et extraction des paramètres

Comme dans de nombreux systèmes de reconnaissance de la parole, plusieurs représentations paramétriques sont proposées. L'analyse acoustique est réalisée par un processeur NEC 7720 qui numérise le signal à 12 KHz, et qui outre la représentation temporelle du signal fournit aussi une analyse spectrale de celui-ci. Suivent ensuite des traitements cepstraux et prédictifs qui permettent d'extraire les paramètres présentés sur la figure 33.

6.2.1 Le vocoder à canaux

Le vocoder utilisé permet de séparer les fréquences du spectre vocal en N canaux d'énergie qui couvrent la bande 350-6500 Hz. Actuellement, nous utilisons 16 canaux qui donnent toutes les 10 ms l'énergie du signal dans les différentes bandes de fréquence dont la répartition est décrite sur le tableau 34. La visualisation du spectre obtenu est identique au spectrogramme: en abscisse est porté le temps, en ordonnée la fréquence, et l'amplitude relative à chaque fréquence est représentée par un code donnant l'intervalle où se trouve la valeur de l'amplitude. La figure 35 illustre la représentation d'une phrase à l'aide d'un tel spectrogramme.

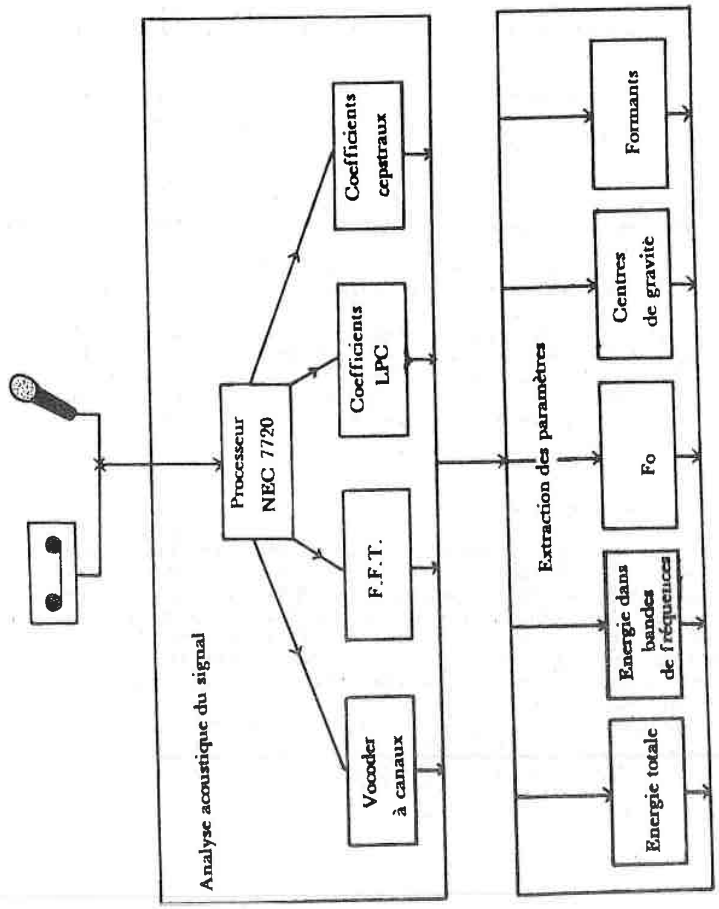


Figure 33

Analyse acoustique du signal et extraction des paramètres

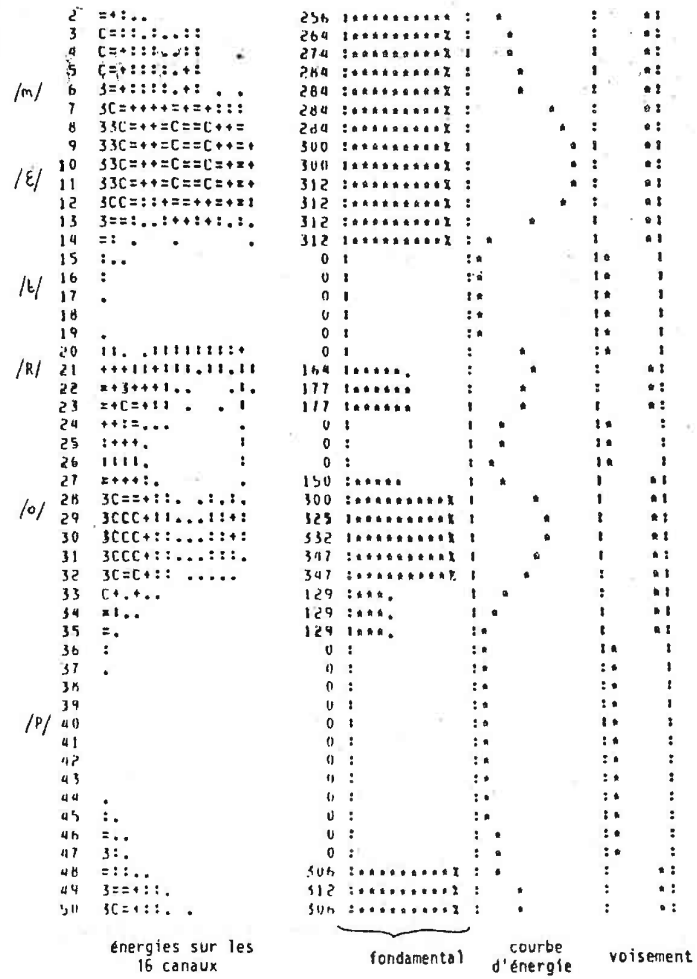
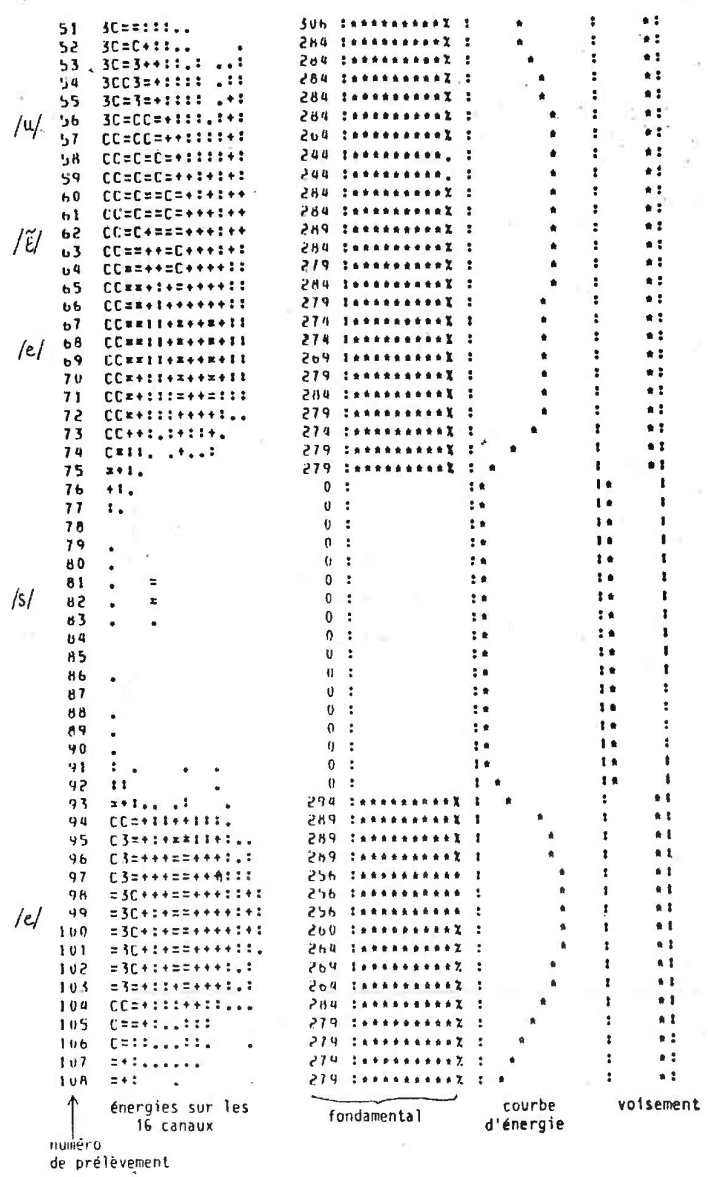


Figure 35

Spectrogramme de la phrase

/mettre au point/



Caractéristiques des filtres du vocodeur utilisé

No CANAL	BANDE PASSANTE	LARGEUR DE BANDE
1	350	200
2	550	200
3	750	200
4	950	200
5	1175	250
6	1450	300
7	1750	300
8	2050	300
9	2350	300
10	2650	300
11	2950	300
12	3300	400
13	3700	400
14	4100	400
15	4700	800
16	5900	1600

Tableau 34

6.2.2 Les coefficients cepstraux

Les coefficients cepstraux calculés sont des MFCC (Mel Frequency Cepstrum Coefficients). Le signal subit un filtre passe-bas à 5 KHz et est échantillonné à 10 KHz. Pour respecter l'échelle Mel, 24 filtres passe-bande sont simulés selon l'échelle proposée par Fant [FANT-73]: des fréquences espacées linéairement en dessous de 1000 Hz et un espacement logarithmique au-dessus (figure 36). Le spectre de Fourrier est calculé pour des segments de 64 points (6.4 ms) ou de 128 points (12.8 ms). Une fenêtre de Hamming d'au moins 256 points est utilisée pour sélectionner les données à utiliser (une taille inférieure à 256 donne de moins bons résultats). Les coefficients cepstraux sont alors calculés par transformation en cosinus, du logarithme

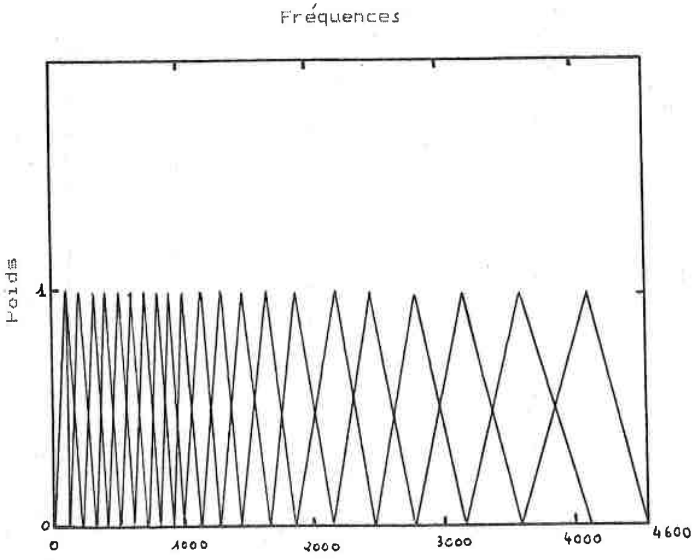


Figure 36 : Filtres générant les MFCC.

décimal du spectre des énergies:

$$(6.1)\quad MFCCi = \sum_{k=1}^{24} X_k \cos\left(i\left(k-\frac{1}{2}\right)\frac{\pi}{24}\right) \qquad i=1 \text{ à } M$$

où : - M est le nombre de coefficients spectraux
- X_k représente le logarithme décimal de l'énergie du kième filtre.

Les meilleurs résultats sont obtenus pour M proche de 10. Dans le programme de calcul qui a été implanté et qui s'inspire de celui qu'ont élaboré Chollet & Gagnoulet pour le système SERAPHINE [CHOL-82], M est égal à 8.

6.2.3 Les coefficients LPC

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:

$$(6.2)\quad S_n = -\sum_{k=1}^{p} a_k S_{n-k} + V_n$$

où les a_k représentent les paramètres de la fonction de transfert du conduit voacl et où V_n représente l'excitation. Si on prédit la valeur du signal S'_n à l'aide des p valeurs précédentes du signal de façon linéaire, on a:

$$(6.3)\quad S'_n = \sum_{k=1}^{p} \alpha_k S_{n-k}$$

l'erreur de prédiction étant $e_n = S_n - S'_n$.

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. Notons par E l'erreur quadratique moyenne.

$$(6.4)\quad E = \sum_{n} e_n^2 = \sum_{n} (S_n - S_n')^2$$

La minimisation de ce terme peut être réalisée par différentes méthodes de covariance et d'autocorrélation. Nous avons choisi pour notre système la méthode d'autocorrélation de Markel [MARK-72].

Appliquons une fenêtre de largeur N sur la partie de signal analysée. Nous avons alors:

$$(6.5)\quad S_n = \begin{cases} \text{quelques échantillons de signal}, & 0 \leqslant n \leqslant N-1 \\ 0, & n<0 \text{ et } n \geqslant N \end{cases}$$

En substituant (6.2) et (6.5) dans (6.4), et en posant L=N-1+p, nous obtenons:

$$(6.6)\quad E = \sum_{n=-\infty}^{+\infty} (S_n - S_n')^2 = S_0^2 + \sum_{n=1}^{L} \left(S_n - \sum_{k=1}^{p} a_k S_{n-k} \right)^2$$

La condition qui minimise E est obtenue en posant la dérivée partielle de E par rapport à chaque coefficient a_k, $1<k<p$, égale à 0. Le résultat obtenu peut être illustré par:

$$(6.7)\quad \sum_{k=1}^{p} a_k R_{|i-k|} = R_i \qquad \text{pour } 1 \leqslant i \leqslant p$$

où

$$(6.8)\quad R_i = \sum_{n=0}^{N-1-|i|} S_n S_{n+|i|}$$

est la fonction d'autocorrélation du signal S_n.

Pour minimiser la valeur de l'erreur, il suffit donc de calculer les coefficients d'autocorrélation R_k, $0 \leqslant k \leqslant p$, en utilisant la formule (6.8) et ensuite de résoudre les équations (6.7) pour déterminer les

coefficients de prédiction linéaire a_k, $1 \leqslant k \leqslant p$. L'erreur minimale E_p peut alors être calculée en substituant (6.7) dans (6.6):

$$(6.9)\quad E_p = -\sum_{k=1}^{p} a_k R_k + R_0$$

La méthode d'autocorrélation a été adoptée car, plus stable que la méthode de covariance, elle se prête mieux à un suivi de formants. On a conservé les mêmes conditions d'analyse que pour le cepstre (longueur de fenêtre, pas de déplacement et accentuation). Un nombre de coefficients égal à 14 paraît un bon compromis: au-delà on aboutit à une surdétection de formants et en-dessous à des manques manifestes.

6.3 La segmentation

6.3.1 La démarche poursuivie

De nombreux travaux concernant la perception de la parole tendent à montrer que les traits phonétiques constituent l'interface entre le signal acoustique et le phonème [ROSS-77]. Dans cette perspective, on saisira tout l'intérêt qu'il y a à partir des discontinuités acoustiques pour une segmentation du signal plus directement reliée à cette unité intermédiaire indispensable qu'est le trait phonétique. Toutefois, il convient de garder à l'esprit que les traits ne sont pas organisés linéairement mais simultanément, et présentent ainsi des caractéristiques de chevauchement. Celles-ci proviennent du fait que les traits s'organisent au sein de l'unité supérieure qu'est la syllabe.

La méthode de segmentation que nous avons choisie, est donc fondée sur l'identification d'évènements acoustiques de segments, qui correspondent aux macro-classes phonétiques suivantes: voyelles, plosives, fricatives et autre consonnes. Dans une première phase, nous effectuons un prétraitement pour obtenir la durée vocalique moyenne et une segmentation grossière qui

Nom	Résultats	Implantation
NOVOCA	Durée vocalique moyenne Noyaux voaciques+limites	Procédural
PLOSIF	Plosives + limites	Procédural
FRICATIF	Fricatives + limites	Procédural
FRAMINT	Frontières précises des phonèmes	Identification à base de frames

Figure 37
Synoptique de la segmentation dans LFRAP

correspond aux quatres grandes classes phonétiques précitées. Puis, dans un deuxième temps, nous utilisons l'ensemble des règles obtenues avec l'expert et contenues dans le langage à base de frames de LFRAP, pour identifier et segmenter finement. Ces règles peuvent modifier la segmentation, c'est à dire changer les limites d'un segment, scinder un segment en deux ou regrouper deux segments. La première phase est implantée sous forme procédurale alors que la seconde est intégrée au processus d'identification géré par le langage de frames (figure 37).

Notre but étant de développer un système de décodage acoustico-phonétique multicritères et multianalyses, la segmentation est réalisée sur les prélèvements vocoder, puis est ensuite propagée aux autres représentations paramétriques du signal.

Il est à noter que ce type de segmentation "hiérarchisée" tend à se développer de plus en plus. Outre les travaux menés au CRIN, citons les propositions destinées à segmenter et étiqueter la base de données acoustiques du GRECO Communication Parlée.

6.3.2 La détection des noyaux voacliques : NOVOCA

NOVOCA [FOHR-85] a pour but de trouver tous les noyaux vocaliques contenus dans une phrase et de déterminer la durée vocalique moyenne. Le critère, qui est utilisé pour reconnaître les noyaux vocaliques, est l'énergie contenue dans la bande de fréquences [250Hz-2350Hz] qui défavorise les sons ayant principalement de l'énergie en très basse fréquence (nasales par exemple) et ceux qui ont de l'énergie en haute fréquence (fricatives). Les noyaux vocaliques correspondent alors aux pics de la courbe d'énergie qui vérifient les critères suivants:

- le nouveau pic doit atteindre au moins 55% du pic précédent (deux noyaux successifs ne peuvent pas avoir des énergies trop différentes);
- la vallée de part et d'autre du pic est fonction de la hauteur du pic (plus un pic est important, plus la vallée doit être importante);
- au moins 50% des échantillons du noyau vocalique doivent être voisés.

Quand un pic vérifie tous ces critères, on recherche le début et la fin du noyau correspondant. On ne décrira que la recherche de la fin du noyau, car la recherche du début est symétrique.

1) A partir du pic, on recherche les points D,F et R qui correspondent respectivement aux bornes de la chute d'énergie, et au seuil à partir duquel l'énergie commence à remonter.

2) On trace le segment DF, et on cherche le point de la courbe d'énergie situé au-dessus de cette droite et qui, de plus, est à la plus grande distance de cette droite. Si le point trouvé se situe entre D+(F-D)/4 et F-(F-D)/4, c'est le marqueur de fin de noyau, sinon c'est (D+F)/2.

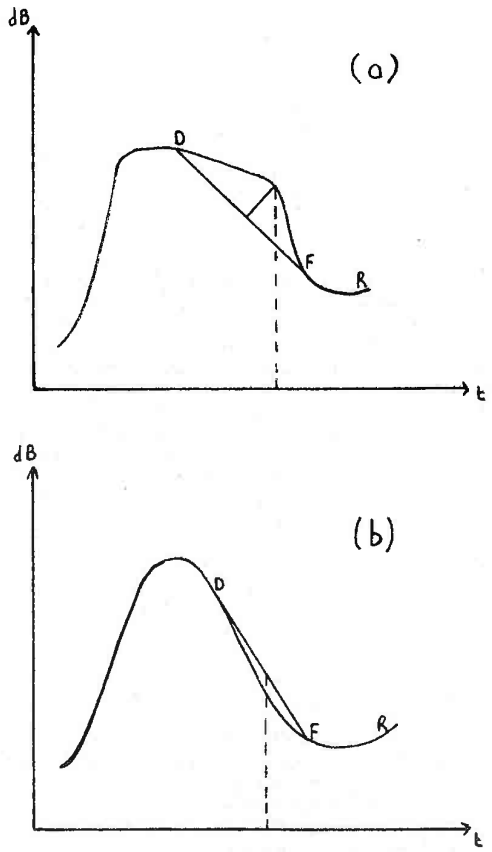


Figure 38
Exemples de courbes d'energie presentant une epaule (a) et sans epaule (b)

La figure 38 présente deux exemples.

6.3.3 La détection des plosives

Les occlusives sont caractérisées par la présence d'une zone de stabilité correspondant soit à la barre de voisement, soit au silence précédent l'explosion, suivie d'une zone de forte instabilité correspondant à l'explosion et à la transition vers la voyelle. La zone stable de l'occlusion est caractérisée par les indices suivants: énergies et centres de gravité des spectres peu élevés. La zone instable qui suit est caractérisée par une très forte variabilité spectrale, et par la présence d'un maximum d'énergie dans les 26 ms suivant le début de l'explosion.

L'algorithme de détection des plosives que nous utilisons est fondé sur ces considérations. Pour chaque prélèvement sont calculés:
- E : énergie dans la bande de fréquences [950Hz-6000Hz]
- MAXPV : maximum des pics vocaliques adjacents (énergie totale)

Seront alors considérés comme plosifs les prélèvements qui vérifient les critères suivants:
E < 30 dB ou MAXPV-E > 20 dB.

6.3.4 La détection des fricatives

Les fricatives se caractérisent sur le plan spectral par la présence de bruits de friction dans les bandes de fréquences aigües. Ainsi, le son /s/ a plusieurs pôles de bruits dont le plus important se situe autour de 5000 Hz. Notre algorithme de détection utilise ces critères fréquentiels pour localiser et délimiter les fricatives.

Dans un premier temps, on détermine pour chaque prélèvement PREL le maximum énergétique sur les quatres prélèvements adjacents PREL-2, PREL-1, PREL+1, PREL+2, et sur le prélèvement courant, MAX(PREL), afin de localiser et concentrer les pôles de bruits de friction. Puis dans

une deuxième phase, on utilise une détection par bandes de fréquences pour identifier les zones fricatives:

1) Calcul de: ENERG1 = E[3000-4000Hz] - E[250-1500Hz]
ENERG2 = E[4000-5000Hz] - E[250-1500Hz]
ENERG3 = E[5000-6000Hz] - E[250-1500Hz]
sur le prélèvement énergétique déterminé pour chaque prélèvement: MAX(PREL).

2) Si (ENERG1>15dB) ou (ENERG2>15dB) ou (ENERG3>15dB) alors PREL est fricatif.

6.3.5 Perspectives

Parallèlement à NOVOCA, nous avons testé l'algorithme de segmentation du système RAPACE proposé par L. Sauter [SAUT-84] & [FOHR-85]. Dans cet algorithme qui segmente le signal en demi-syllabes, les frontières interunités sont déterminées par les minima et les maxima successifs de l'énergie du signal dans la bande de fréquences 500-6000Hz. La suite d'extrémums de l'énergie recherchée doit vérifier les contraintes suivantes:
- la longueur de chaque demi-syllabe doit dépasser au minimum 40 ms,
- le rapport entre les valeurs de l'amplitude aux limites de chaque demi-syllabe doit dépasser un certain seuil,
- les maxima retenus doivent correspondre à une énergie suffisante.

Le choix de la demi-syllabe comme unité de segmentation permet d'éviter que plusieurs voyelles soient considérées comme un seul noyau vocalique. D'autre part, on tient compte du fait qu'en français, seules certaines suites de consonnes peuvent commencer ou terminer une demi-syllabe.

Cet algorithme a été évalué sur le même corpus que NOVOCA, et les performances respectives des deux méthodes sont sensiblement équivalentes: 95% de noyaux localisés avec RAPACE contre 97% avec NOVOCA, avec des taux d'insertion et d'omission inférieurs à 8%. Si NOVOCA a été retenu dans la version actuelle de LFRAP pour des

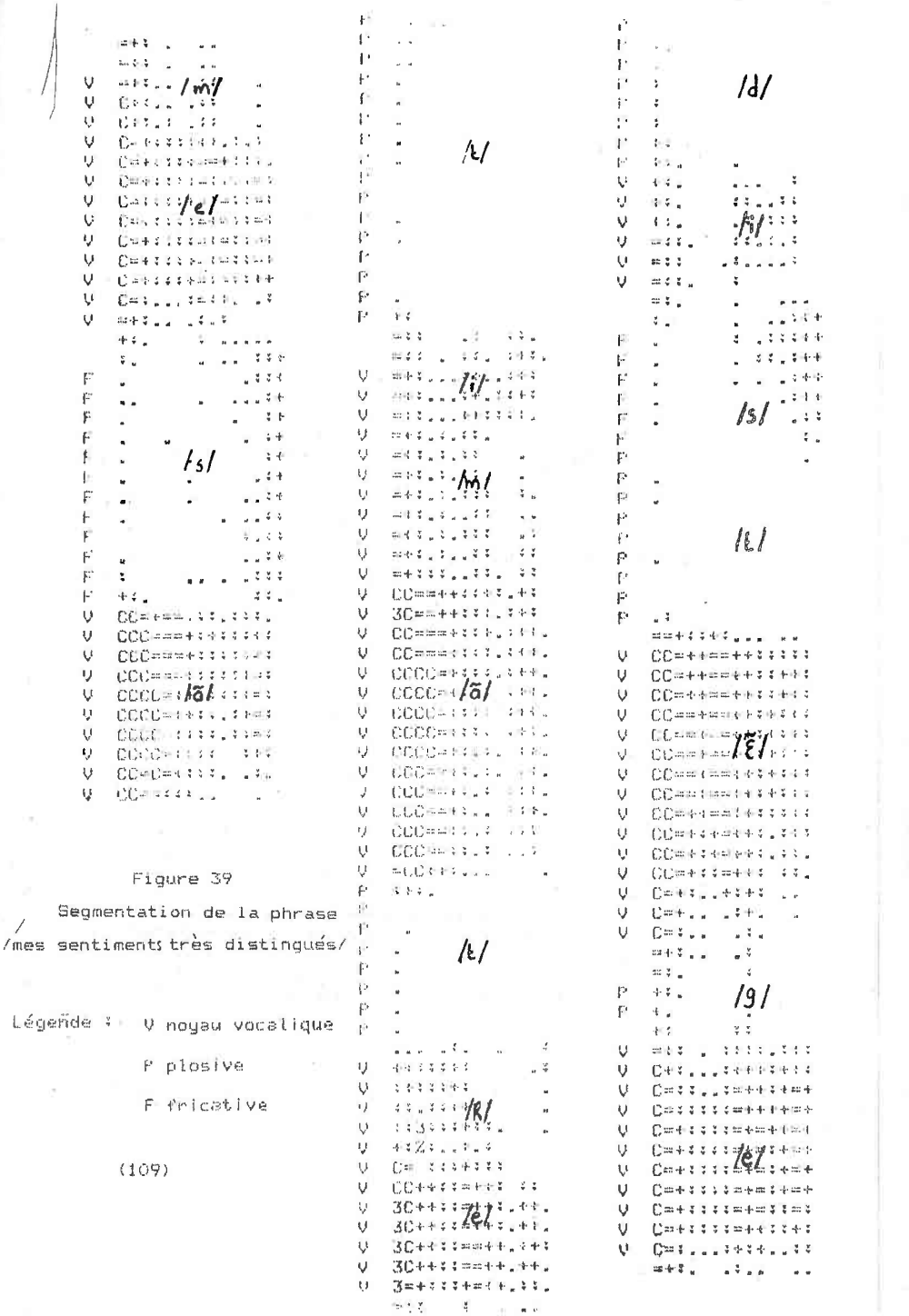


Figure 39

Segmentation de la phrase
/mes sentiments très distingués/

Légende : V noyau vocalique
P plosive
F fricative

raisons d'homogénéité de l'ensemble du module de segmentation (la figure 39 illustre la segmentation de la phrase "mes sentiments très distingués"), nous étudions actuellement dans quelle mesure leur utilisation conjointe pourrait permettre d'obtenir une segmentation syllabique de qualité encore meilleure.

6.4 L'étiquetage

6.4.1 La construction du frame INSTANCE : l'HYPOTHETISEUR

Ce module assure deux fonctions qui sont toutes deux supervisées par la grammaire de frames: la création de la description acoustique du segment étudié et la sélection des candidats au processus de matching.

L'algorithme qui est utilisé par LFRAP pour la construction du frame INSTANCE nécessite que la grammaire de frames soit représenté par un automate d'états finis $A(Q,V,\delta,q_0,F)$ où:
- Q est un ensemble de n+1 états, $Q=\{q_0,q_1,\ldots,q_n\}$, qui correspond aux différentes caractérisations acoustiques qui peuvent être associées à un segment,
- V est un vocabulaire qui regroupe l'ensemble des noeuds (frames) de la grammaire,
- $\delta \subset Q\times V\times Q$ est un ensemble de transitions: $\delta=\{t_1,t_2,\ldots\}$,
- q_0 est l'état initial de l'automate,
- $F\subset Q$ est l'ensemble des états finals, ceux-ci correspondent aux états définitifs que peut prendre le frame INSTANCE et qui seront comparés aux prototypes durant le processus de matching.

On suppose que les états de Q sont numérotés de manière que:

$$\forall\, v\in V \quad \forall\, q_i,q_j\in Q \quad (q_i,v,q_j)\in\delta \;\Longrightarrow\; i<j$$

D'autre part, les transitions de . sont numérotées de manière que:

si $t_i=(q_{i_1},v_i,q_{i_2})$ et $t_j=(q_{j_1},v_j,q_{j_2})$ alors $i<j \Longrightarrow i_1<j_1$.

Ces contraintes sont introduites dans le but de respecter la hiérarchie liant les indices acoustiques. La figure 40 présente l'automate correspondant à la grammaire formalisant la hiérarchie d'indices illustrée sur la figure 23.

Les frames de la grammaire, qui définissent les transitions entre les états de l'automate, constituent l'élément moteur de l'algorithme de construction du frame INSTANCE. Chaque frame est décomposé en atomes qui sont la représentation interne des descripteurs et slot-filler de la grammaire. Ces atomes sont implantés dans LFRAP en PCF-2, notation introduite par Sandewall pour résoudre les problèmes de représentation et d'implantation des frames dans le traitement des langages naturels [SAND-72]. La méthode utilisée distingue quatre types d'entités:
- les objets (frame INSTANCE et candidats au processus de matching),
- les actions qui regroupent instancieurs, sélecteurs et contrôleurs,
- les états (élements de Q),
- les propriétés (type de frame "SUBFRAME" ou "TERMINAL"),

qui, associées au sein de fonctions et de relations, constituent les atomes des frames.

LFRAP utilise les atomes suivants:
- CONS : objet x action --> objet
  crée et instancie un nouveau slot au frame INSTANCE (descripteur FILL)
- SUCC : état x action --> état
  assure la transition entre deux états de l'automate (descripteurs SELECT et SPECIALIZATION-OF)
- SELE : état x action --> objet
  sélectionne les candidats au matching ou de comparaisons par références acoustiques (descripteurs MATCH, DEFAULT et USE-REF)
- IS : objet x propriété x état
  indique que l'objet possède la propriété dans l'état (descripteur TYPE).

On peut noter que le signal et ses représentations paramétriques ne sont pas reconnus par le système en tant qu'objets afin de respecter le concept déclaratif des frames; ce sont les instancieurs qui se chargent de

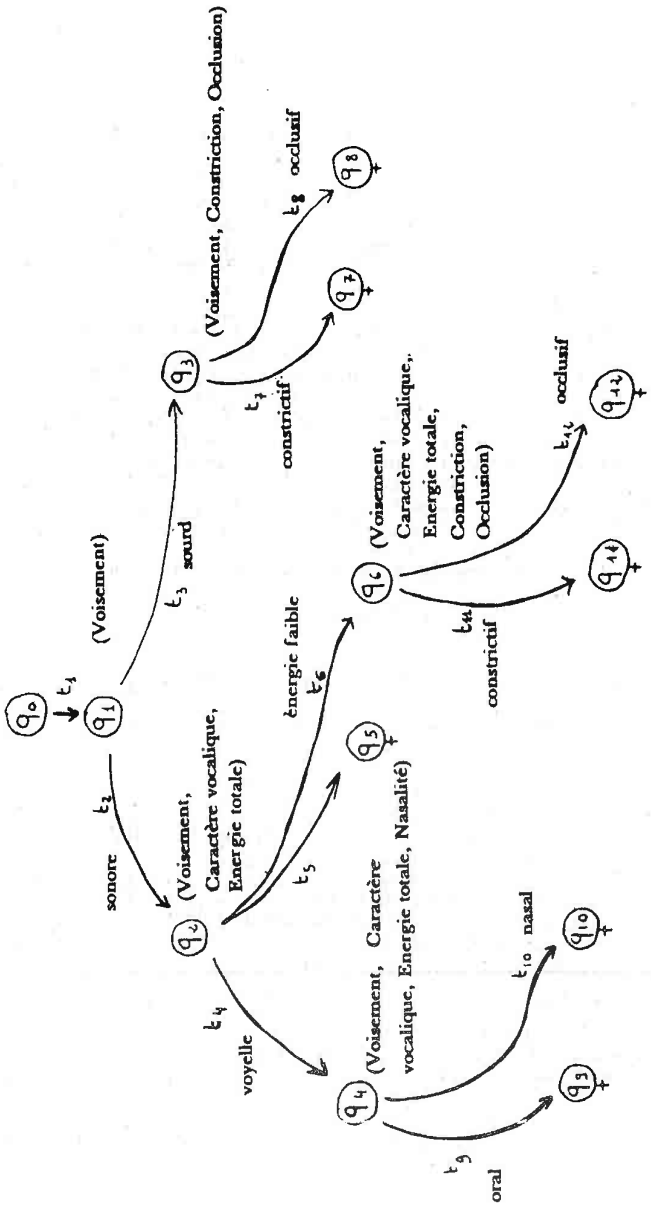


Figure 40.
Automate correspondant à la grammaire formalisant la hiérarchie d'indices illustrée sur la figure 22

réaliser l'interface entres frames et monde physique.

Ainsi formalisé en PCF-2, l'algorithme de construction du frame INSTANCE se traduit par l'application séquentielle de fonctions CONS et SUCC:

Initialisation: $q \leftarrow q_0$, $o_1 \leftarrow \emptyset$, $o_i \leftarrow \emptyset$

Construction :
```
Tant que IS(o1,"SUBFRAME",q) Faire
    (* N nombre de slots à créer *)
    Pour i=1 to N Faire o1 <-- CONS(o1,ai)
    q <-- SUCC(q,aN+1)
Pour i=1 to N Faire o1 <-- CONS(o1,ai)
```

Sélection au matching : $o_i \leftarrow SELE(q,a)$

Cette dernière opération SELE entraîne la création d'un treillis d'hypothèses noté dans l'algorithme o_i qui détermine la suite des traitements à réaliser, comparaison de frames ou comparaisons de réalisation acoustiques en fonction de la nature des éléments qui le composent.

6.4.2 Le comparateur

6.4.2.1 La comparaison de frames

Ce module se propose de comparer le frame INSTANCE avec l'ensemble des candidats qui ont été sélectionnés par l'hypothétiseur. Le frame INSTANCE sera considéré compatible avec le frame PROTOTYPE P si l'on peut établir qu' aucune des caractéristiques du frame INSTANCE ne contredit les propriétés acoustiques du phonème associé à P. La relation de compatibilité n'est pas une relation d'égalité au sens strict, et la fonction de décision qui s'appuie sur le théorie des nesmbles flous quantifie le résultat de la comparaison entre deux formes. Ce quantifieur, appelé "coefficient de plausibilité",

détermine le degré de compatibilité entre les deux formes.

La comparaison entre les deux frames est réalisée au niveau des slots par la fonction SLOT-BY-SLOT-MATCH. Considérons un slot commun aux deux frames. Si le contenu du slot du frame INSTANCE vérifie la relation d'ordre précisée par le frame PROTOTYPE, la compatibilité sera validée. Cette fonction est appliquée à tous les slots, et le coefficient de plausibilité correspondra au taux de slots compatibles entre les deux frames:

```
Taux = 0

(* N nombre de slots à comparer *)
Pour i=1 to N faire
   Si SLOT-BY-SLOT-MATCH(slot.)=vrai
   alors Taux = Taux + 1

Taux = Taux * 100 / N
```

La figure 41 présente un exemple de comparaison de frames.

6.4.2.2 La comparaison de réalisations acoustiques

La représentation paramétrique de référence est la représentation vocoder du signal. Chaque prélèvement est comparé aux références du dictionnaire retenu à l'aide de la distance de Hamming:

$$(6.11)\quad Dj = \sum_{j=1}^{N} \left| c(i) - R(j,i) \right|$$

où N est le nombre de canaux , c(i) la valeur du ième canal du prélèvement et R(j,i) la valeur du ième canal de la référence j; et il sera étiqueté avec les trois références les plus proches acoustiquement. Le taux de ressemblance qui détermine ce choix, est la plus petite distance entre le prélèvement à reconnaître et toutes les références:

INSTANCE		PROTOTYPE /p/	
Voisement	0	Voisement	=0
Durée	3.5	Durée	<4
Pics (Fréquences)	1.5	Pics (Fréquences)	[0.75,2.5]
Energie	67	Energie	<100

PROTOTYPE /t/		PROTOTYPE /k/	
Voisement	=0	Voisement	=0
Durée	[4,8]	Durée	>10
Pics (Fréquences)	>2.5	Pics (Fréquences)	[0.75,2.5]
Energie	[100,150]	Energie	>150

Figure 41
Identification d'une plosive sourde (matching)
Taux /p/ = 100 - Taux /t/ = 50 - Taux /k/ = 50

Prélèvements :		
1 a o ɔ		
2 a a o	T /a/ = 21	
3 a ɑ̃ o	T /ɑ̃/ = 10	
4 a ɛ̃ ɔ	T /o/ = 5	
5 a ɛ̃ o	T /ɔ/ = 4	
6 a ɑ̃ ɔ		
7 a ɔ̃ ɔ		

Triplet retenu: /a/ /ɑ̃/ /o/

Prélèvements :	
1 a o ɑ̃	
2 a a ɑ̃	T /ɑ̃/ = 13
3 a ɑ̃ ɔ	T /a/ = 12
4 ɑ̃ a a	T /o/ = 2
5 ɑ̃ a a	T /ɔ/ = 1
6 ɑ̃ a a	

Triplet retenu: /ɑ̃/ /a/ / /

Figure 42
Exemples d'identification par processus
de décision majoritaire

$$(6.12)\quad D = \min_{j} Dj \qquad j=1,\ldots,N$$

où J est le nombre de références du dictionnaire.

L'identification du segment est alors réalisé par un processus de décision majoritaire parmi les trois phones retenus pour chaque prélèvement. Pour chaque phonème candidat, nous calculons son taux d'occurence dans le segment de la façon suivante:

$$(6.13)\quad Tj = \sum_{i=1}^{I} Pi \qquad j=1,\ldots,J$$

où I est le nombre d'occurences du phonème j dans le treillis. Les trois plus grandes valeurs Tj classées par ordre décroissant déterminent les phonèmes qui forment le triplet identifiant le segment. Toutefois, on néglige les phonèmes non significatifs, i.e. qui ont un taux d'occurence faible. Cette hypothèse est vérifiée à l'aide du rapport Ti-1/Ti où i $\in$ [2,3]: si ce rapport est supérieur à un seuil fixé, le phonème associé à Ti est éliminé du treillis phonétique. La figure 42 illustre l'utilisation de ce processus de décision majoritaire.

6.4.2.3 L'identification du segment

Ce module constitue la dernière étape du processus de comparaison. Deux cas doivent être envisagés:

- la sélection des phonèmes candidats a été obtenue par comparaisons de réalisations acoustiques, et dans ce cas le triplet résultant de ces comparaisons est retenu pour étiqueter le segment.

- la sélection des phonèmes candidats a été réalisée par comparaisons de frames: le choix du triplet identifiant le segment analysé repose alors sur les scores de compatibilité calculés à cette étape. Si aucun des scores n'atteint un seuil S fixé à priori, on considère que le processus de reconnaissance a échoué.

Deux solutions peuvent être envisagées: dans le cas où les scores sont supérieurs à 50%, l'identification du segment est entérinée et le triplet retenu est défini par les phonèmes précisés par le descripteur DEFAULT, sinon le système reprendra le décodage phonétique du segment analysé en s'appuyant sur de nouveaux indices. Si par contre, un ou plusieurs scores (au maximum 3) se révèlent supérieurs à S, les phonèmes correspondants seront retenus dans le treillis phonétique pour étiqueter le segment courant.

6.5 L'apprentissage

La phase d'apprentissage consiste à créer et enrichier un dictionnaire phonémique dont chaque élément représente un phonème. Ces derniers sont représentés à l'aide de prélèvements vocoder, et les caractéristiques retenues pour chaque forme de référence sont les 16 canaux du vocoder. Ces 16 composantes décrivent pour chaque intervalle de 10 ms un point X dans $\mathbb{N}^{16}$, et l'ensemble de ces points constitue un nuage de $\mathbb{N}^{16}$ dont le centre de gravité Ω a pour coordonnées:

$$(6.14)\quad w_k = \sum_{i=1}^{N} \frac{X_{i,k}}{N} \qquad \text{avec } k=1,\ldots,16$$

où X_k est la kième composante de X et N le nombre d'échantillons ayant servi à l'apprentissage pour le phonème considéré.

Chaque phonème a une forme de référence vectorielle de 16 éléments $(X_1, X_2, \ldots, X_{16})$, et ce module détermine pour chaque phonème les composantes dans $\mathbb{N}^{16}$ des centres de gravité (vecteurs moyens) du nuage phonémique. Afin de tenir compte des variations contextuelles, les phonèmes sont représentés en moyenne par deux ou trois références.

L'apprentissage de formes de référence des phonèmes est manuel. Un programme interactif permet à l'utilisateur 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, et le locuteur désigne la partie du spectrogramme qui représente le phonème, si possible dans une zone de stabilité. Le programme calcule alors le prélèvement moyen en effectuant la moyenne des énergies sur chaque canal pour les prélèvements de la partie désignée (équation (6.14)).

Pour diminuer la dépendance vis-à-vis du locuteur des formes de référence, il est envisagé une adaptation automatique du système au locuteur en incorporant à LFRAP l'algorithme d'apprentissage automatique des formes de référence du locuteur élaboré par C. Pister dans le cadre de sa thèse de 3ème cycle [PIST-84].

6.6 Les utilitaires de LFRAP

LFRAP propose un certain nombre d'utilitaires pour faciliter la mise au point des procédures de reconnaissance du module d'étiquetage. Au niveau de la programmation du système de contrôle, la traduction source-->PCF-2 des frames de la grammaire s'accompagne d'une analyse morphologique et syntaxique qui met en évidence les erreurs de syntaxe ou d'orthographe ainsi que les erreurs de logique (descripteurs SELECT ou OTHERWISE dans un frame de type TERMINAL par exemple) caractérisant le programme source. Ces erreurs sont signalées au moyen de diagnostics explicites qui figurent dans le corps du texte immédiatement après le descripteur erroné, dans le listing des instructions générées pendant la traduction (figure 43).

Au niveau logique du raisonnement, LFRAP sait expliquer à la manière d'un système expert, les faits qui l'ont amené à opter pour telle solution plutôt qu'une autre dans la conception du treillis phonétique. Ces explications sont visualisées par l'intermédiaire des frames INSTANCES générés pendant le processus de reconnaissance, qui formalisent le décodage acoustico-phonétique de chaque segment de parole. Cette facilité permet de simplifier la mise au point des prototypes, et par conséquent la séparation des sons en classes phonétiques bien distinctes.

Figure 43

Exemple de listing d'interpretation de programme source

```
  Ligne     Chargeur et Interpreteur de Grammaire F.A.G.      Version 1.1 du 03/05/84      Page  1

     1      FRAME Phoneme
     2            SPECIALIZATION-OF Root
     3            TYPE SUBFRAME
     4            FILL Snovoca          USING Pnovoca
     5            FILL Splosif          USING Pplosif
     6            FILL Sfricat          USING Pfricat
     7            WITH (Snovoca=0)      SELECT Voyelle
     8            WITH (Splosif=0)      SELECT Plosive
     9            WITH (Sfricat=0)      SELECT Fricat
    10            OTHERWISE Autres
    11      END
    12      FRAME Plosive
    13            SPECIALIZATION-OF Phoneme
    14            TYPE SUBFRIME
**** Erreur  70 : Erreur Syntaxe Type
    15            FILL Picsener     USING Frecpics
    16            FILL Energie              Penergie
**** Erreur  30 : Erreur Syntaxe Fill
    17            WITH (Picsener>2.5) AND (Energie>E1) AND (Energie<E2) SELECT Phontd
    18            OTHERWISE Phonpbkg
    19      END
    20      FRAME Phontd
    21            SPECIALIZATION-OF Plosive
    22            TYPE TERMINAL
    23            FILL Duree      USING Fduree
    24            FILL Voise      USING Fpitch
    25            MATCH WITH ProtoF,ProtoD
    26      END
    27      FRAME Phonpbkg
    28            TYPE SUBFRAME
    29            FILL Duree USING
**** Erreur  30 : Erreur Syntaxe Fill
    30            WITH (Duree<D1) AND (Energie<E1) SELECT Phonpb
    31            OTHERWISE Phonkg
    32      END
**** Erreur  80 : Self ou Type Inexistant
    33      FRAME Phonpb
    34            SPECIALIZATION-OF Phonpbkg
    35            TYPE TERMINAL
    36            FILL Voise      USING Fpitch
    37            MATCH WITH ProtoF,ProtoB
    38      END
    39      FRAME Phonkg
    40            SPECIALIZATION-OF Phonpbkg
    41            TYPE TERMINAL
    42            FILL Voise      USING Fpitch          (Version avec erreurs)
    43            MATCH WITH ProtoK,ProtoG
    44      END


          ***   Nbre de Lignes Analysees :  44

          ***   Nbre de Frames           :   6

          ***   Nbre d'Erreurs           :   4
```

```
Ligne    Chargeur et Interpreteur de Grammaire F.A.G.      Version 1.1 du 03/05/84      Page 1

   1     FRAME Phoneme
   2           SPECIALIZATION-OF Root
   3           TYPE SUBFRAME
   4           FILL Snovoca        USING Pnovoca
   5           FILL Splosif        USING Pplosif
   6           FILL Sfricat        USING Pfricat
   7           WITH (Snovoca=0)    SELECT Voyelle
   8           WITH (Splosif=0)    SELECT Plosive
   9           WITH (Sfricat=0)    SELECT Fricat
  10           OTHERWISE Autres
  11     END
  12     FRAME Plosive
  13           SPECIALIZATION-OF Phoneme
  14           TYPE SUBFRAME
  15           FILL Picsener     USING Enecpics
  16           FILL Energie      USING Penergie
  17           WITH (Picsener>2.5) AND (Energie>E1) AND (Energie<E2) SELECT Phontd
  18           OTHERWISE Phonpbkg
  19     END
  20     FRAME Phontd
  21           SPECIALIZATION-OF Plosive
  22           TYPE TERMINAL
  23           FILL Duree      USING Fduree
  24           FILL Voise     .USING Fpitch
  25           MATCH WITH ProtoT,ProtoD
  26     END
  27     FRAME Phonpbkg
  28           SPECIALIZATION-OF Plosive
  29           TYPE SUBFRAME
  30           FILL Duree USING Fduree
  31           WITH (Duree<D1) AND (Energie<E1) SELECT Phonpb
  32           OTHERWISE Phonkg
  33     END
  34     FRAME Phonpb
  35           SPECIALIZATION-OF Phonpbkg
  36           TYPE TERMINAL
  37           FILL Voise      USING Fpitch
  38           MATCH WITH ProtoP,ProtoB
  39     END
  40     FRAME Phonkg
  41           SPECIALIZATION-OF Phonpbkg
  42           TYPE TERMINAL
  43           FILL Voise      USING Fpitch          (Version corrigée)
  44           MATCH WITH ProtoK,ProtoG
  45     END


       ***   Nbre de Lignes Analysees :  45

       ***   Nbre de Frames           :   6

       ***   Nbre d'Erreurs           :   0
```


CHAPITRE 7
RESULTATS EXPERIMENTAUX

7.1 Hypothèses de travail

Ce chapitre se veut une illustration de l'utilisation du formalisme des frames dans le domaine de la reconnaissance phonétique. Nous ne présentons pas d'algorithmes nouveaux dans la caractérisation acoustique des phonèmes, la finalité de ce travail n'étant pas la réalisation ou l'amélioration de procédures de décodage acoustico-phonétique, mais la mise en oeuvre d'un outil et d'un environnement favorables à la modélisation du savoir-faire d'un expert en lecture de spectrogrammes. Les travaux qui sont détaillés dans les paragraphes suivants, ont pour but de valoriser l'emploi de cette technique d'intelligence artificielle qu'est le formalisme par frames en décodage acoustico-phonétique d'une part, et d'évaluer les performances de notre système d'autre part. En effet, LFRAP est le module de décodage appelé à remplacer le décodeur acoustico-phonétique développé dans l'équipe [LAZR-83] et utilisé dans les différentes applications en reconnaissance automatique de la parole continue qui sont étudiées dans notre laboratoire, et en particulier les systèmes de reconnaissance de grands vocabulaires (2000 mots) [MARI-84] et de dictée automatique [CHAR-85].

Les performances de notre système vont être étudiées dans le cadre de la détection et l'identification des liquides, un des sous-problèmes les plus mal résolus actuellement. Deux approches ont été retenues pour ces tests: une méthode fondée sur les caractéristiques spectrales des sons /l/ et /R/ [CHAF-83], et une méthode multi-critères élaborée dans le cadre du projet APHODEX [LEMO-85]. Les phrases utilisées pour les tests sont empruntées aux corpus de Combescure [COMB-81] et de Charpillet [CHAR-85]. Elles ont été prononcées de manière naturelle par deux locuteurs non entraînés, et pour éviter une lecture des phrases, on a demandé à ces derniers de mémoriser chaque phrase, puis de la prononcer.

7.2 Le problème /l-R/

Les propriétés acoustiques des consonnes /l/ et /R/ du français ont été principalement analysées par Delattre qui s'est attaché à les caractériser du point de vue de leurs cibles acoustiques et des transitions de leurs formants [DELA-66] et [DELA-68]. A la lecture de quelques spectrogrammes (FIGURES 8 et 9), on s'apperçoit tout de suite que ces phonèmes peuvent se présenter sous des aspects très divers. Leur énergie, souvent moyenne, peut dans le cas de /R/ disparaître totalement après une consonne ou en cas de dévoisement. Le polymorphisme acoustique de ce phonème trouve ses origines principalement dans l'influence dans l'influence du contexte et de la position dans l'énoncé, ainsi que dans l'intermédiaire source vocale - source de bruit [CHAF-84]. Lorsque l'énergie est visible, on distingue généralement des formants leur donnant un air vocalique assez proche des voyelles /ɛ/,/a/ et /u/, et en contexte vocalique ceux-ci ont tendance à incliner les formants des voyelles adjacentes donnant au spectrogramme l'aspect d'un "X" caractéristique. Cette particularité acoustique permet de discriminer les phonèmes /l-R/ en étudiant la position et les transitions de leurs formants, et c'est l'approche qu'a suivi Lemoine dans ses travaux concernant la recherche d'indices de discrimination des sonnantes /m-n-l-R/ [LEMO-85]. Un autre point de vue consiste, comme le préconise Fant [FANT-60] à décrire ces consonnes d'après l'analyse de leur enveloppe spectrale. C'est cette approche qu'a choisi Chafcouloff dans ses travaux sur les liquides [CHAF-83].

Dans l'étude qui va suivre, ces deux méthodes vont être traitées successivement. Le but poursuivi n'est pas de comparer leurs performances respectives dans l'absolu, les auteurs ayant déjà précisé les limites de leurs méthodes, mais de mesurer la contribution des frames dans les résultats obtenus par LFRAP tant au niveau de la formalisation des algorithmes qu'au niveau des taux de reconnaissance.

7.3 Indices de discrimination proposés par Chafcouloff

7.3.1 Protocole expérimental

Les principales différences en termes de distance et d'énergie acoustique entre les pôles spectraux des consonnes /l/ et /R/ qui ont été déterminées par Chafcouloff sont illustrées sur la figure 44, et peuvent être résumées par les règles suivantes:

1) Si $P_2 - P_1 > 1000$ Hz alors /l/
2) Si $P_2 - P_1 < 1000$ Hz alors /R/
3) Si $A_1 - A_2 \geqslant 20$ dB alors /l/
4) Si $A_1 - A_2 \leqslant 15$ dB alors /R/
5) Si $A_1 - A_3 \geqslant 30$ dB alors /l/
6) Si $A_1 - A_3 \leqslant 20$ dB alors /R/

7.3.1.1 Extraction des paramètres

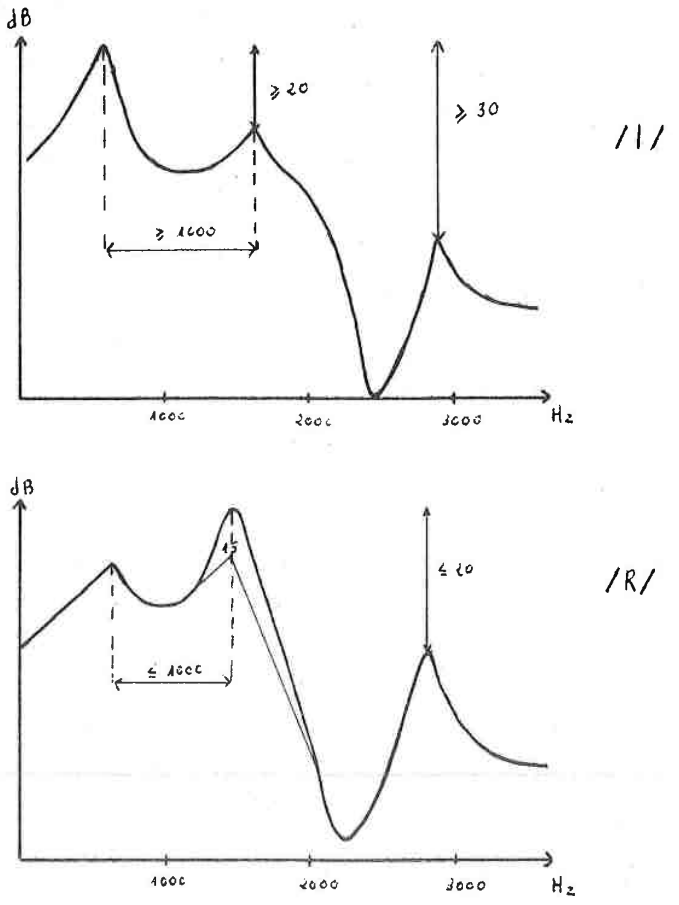
L'extraction des caractéristiques spectrales est réalisée par FFT. Les fréquences qui varient de 0 à 6000 Hz sont regroupées en 128 canaux, et le spectrogramme numérique obtenu donne toutes les 10 ms la valeur de l'énergie en dB pour chacun des 128 canaux. Afin d'éliminer tout spectre tourmenté, on procède à un lissage des prélèvements du spectrogramme.

7.3.1.2 Localistion des liquides

On recherche les segments ayant les configurations types décrites sur la figure 44, entre les noyaux voacaliques isolés pendant la phase de pré-traitement, sachant que les fricatives et les plosives ont été éliminées pendant cette même phase de pré-traitement. Afin de discriminer nasales et liquides, on recherche une paire de pics spectraux à 250 et 1000 Hz, ces fréquences étant considérées comme des invariants du murmure nasal des consonnes [FUJI-62]; et ne seront retenus comme liquides que les prélèvements ne présentant pas ce caractère nasal.

Figure 44

Enveloppe spectrale des consonnes /l-R/



7.3.1.3 Distinction /l-R/

Elle est réalisée par le recherche des 3 pôles spectraux P_1,P_2 et P_3, et l'application des règles de Chafcouloff sur chacun des prélèvements des segments retenus. La correction des frontières de chaque segment sera fonction de son homogénéité acoustique et de la durée vocalique moyenne.

7.3.2 La formalisation des connaissances

La séquence de traitements présentée ci-dessus permet de dégager deux types de connaissances: les connaissances proprement dites sur les liquides /l/ et /R/ qui sont formalisées dans LFRAP par deux frames PROTOTYPES, et les métaconnaissances qui permettent de localiser et caractériser une liquide à l'intérieur du continuum vocal et qui sont formalisées par un programme de contrôle à base de frames. C'est ce programme de contrôle qui déterminera la base de faits du système, ensemble de traits phonétiques réunis dans le frame INSTANCE qui sera comparé aux frames PROTOTYPES associés aux phonèmes /l/ et /R/ afin de déterminer la nature du segment analysé.

7.3.2.1 Le programme de contrôle

Remarque : nous ne traitons que la partie concernant les liquides.

```
(* Préclassification voyelle, fricative, plosive *)

Frame PHONEME
      Specialization-of ROOT
      Type subframe
      Fill VOYELLE   Using P_NOYAUX_VOCALIQUES
      Fill PLOSIVE   Using P_PLOSIVES
      Fill FRICATIVE Using P_FRICATIVES
      With VOYELLE   Select PHON-VOYELLE
      With PLOSIVE   Select PHON-PLOSIVE
      With FRICATIVE Select PHON-FRICATIVE
      Otherwise AUTRE-PHONEME
```

```
End

(* Discrimination liquides / nasales *)

Frame AUTRE-PHONEME
      Specialization-of PHONEME
      Type subframe
      Fill NASALITE  Using  F_NASALITE
      With NASALITE  Select PHON-NASALE
      Otherwise PHON-LIQUIDE
End

(* Distinction /l-R/ *)

Frame PHON-LIQUIDE
      Specialization-of AUTRE-PHONEME
      Type terminal
      Fill DP1       Using P_POLES(1,2,Fréquence)
      Fill DP2       Using P_POLES(1,2,Energie)
      Fill DP3       Using P_POLES(1,3,Energie)
      Match With PROTOTYPE-l,PROTOTYPE-R
End
```

où la procédure P_POLES(Param1,Param2,Param3) détermine l'écart en fréquences (Param3 = "Fréquence") ou en dB (Param3 = "Energie") entre les deux pics Param1 et Param2.

7.3.2.2 Les frames PROTOTYPES

```
PROTOTYPE-l                    PROTOTYPE-R

   VOYELLE    = 0              VOYELLE    = 0
   PLOSIVE    = 0              PLOSIVE    = 0
   FRICATIVE  = 0              FRICATIVE  = 0
   NASALITE   = 0              NASALITE   = 0
   DP1        > 1000           DP1        < 1000
   DP2        ⩾ 20             DP2        ⩽ 15
   DP3        ⩾ 30             DP3        ⩽ 20
```


7.3.3 Résultats obtenus

Ils sont résumés sur le tableau 45. On notera que les taux de reconnaissance obtenus corroborent les conclusions et réserves émises par Chafcouloff à propos des ces indices, à savoir que les configurations spectrales présentées sur la figure 44 ne constituent pas un invariant acoustique des sons /l/ et /R/, la coarticulation entraînant de nombreuses variations allophoniques. Ainsi les principales omissions ne résultent pas de confusions avec d'autres sons vocaliques mais d'une mauvaise discrimination entre liquides, notamment en contexte /u/. Ces règles simples nous permettent tout de même d'atteindre un score global de 72%, et on peut envisager de meilleures performances en adoptant des critères d'analyse plus fins, qui seront faciles à intégrer en raison du clivage qui existe entre le système de contrôle et les algorithmes de reconnaissance.

	Taux de Reconnaissance	Omissions	Insertions
/l/	66%	34%	7%
/R/	78%	22%	<5%

Tableau 45

7.4 Indices de discrimination proposés par Lemoine

7.4.1 Protocole expérimental

Menés dans le cadre du projet APHODEX, les travaux de Lemoine s'appuient sur la position et les transitions des formants pour essayer de dégager des configurations typiques des phonèmes /m-n-l-R/. Il a ainsi pu classer

l'ensemble de ces sonnantes en 5 groupes différents:

- configuration C1: pics dans les bandes [0-500Hz] et [2000-3000Hz]

- configuration C2: pics dans les bandes [0-500Hz], [1000-2000Hz] et [2000-3000Hz]

- configuration C3: pics dans les bandes [0-500Hz], [500-1000Hz] et [2000-3000Hz]

- configuration C4: un pic large dans la bande [0-1000Hz] et un pic dans la bande [2000-3000Hz]

- configuration C5: un pic large dans la bande [0-1000Hz] et des pics dans les bandes [1000-2000Hz] et [2000-3000Hz].

Une fois déterminé le type de configuration d'un prélèvement, on l'affecte à l'une des classes /l/, /R/ ou /m-n/ en utilisant les remarques suivantes:
1) un pic large entre 0 et 1000 Hz est caractéristique d'un /R/ voisé.
2) un pic entre 0 et 500 Hz et rien entre 500 et 2000 Hz est caractéristique d'une nasale.
3) deux pics entre 0 et 1000 Hz sont caractéristiques d'une nasale et parfois d'un /R/ voisé.
4) tout son sourd est considéré comme un /R/, les /l/ sourds étant trop peu nombreux pour les caractériser.

Pour les configurations de type C1, C4 et C5, ces remarques s'appliquent directement donnant les affectations /m-n/ pou C1, et /R/ pour C4 et C5. En ce qui concerne les configurations ambigües de type C2 et C3, on utilise des caractéristiques plus fines sur la position ou la hauteur des formants:

```
* C2: Si (pic entre 1000 et 2000 Hz) > 1400 Hz
     alors /l/
     sinon si (pic entre 0 et 500 Hz) inexistant
              ou < 350 Hz
           alors /m-n/
           sinon /l/
```



```
* C3: Si (pic entre 0 et 500 Hz) > 280 Hz
     alors /l/
     sinon si (pic entre 0 et 500 Hz) >
              (pic entre 500 et 1000 Hz)
           alors /m-n/
           sinon /l/
```

7.4.1.1 Extraction des paramètres

L'extraction des caractéristiques spectrales est réalisée par FFT. Les fréquences qui varient de 0 à 6000 Hz sont regroupées en 128 canaux, et le spectrogramme numérique calculé donne toutes les 10 ms la valeur de l'énergie en dB pour chacun des 128 canaux. Afin d'éliminer tout spectre tourmenté, on procède à un lissage des prélèvements du spectrogramme.

Une méthode de morphologie mathématique permet d'extraire les pics possibles de chaque spectre. Elle consiste à calculer le minimum de la courbe E=h(f), représentant le spectre, sur une fenêtre glissante de taille fixée à l'avance. On obtient ainsi une nouvelle courbe E'=h'(f) sur laquelle on réitère le procédé en prenant cette fois les maxima, ce qui donne la courbe E"=h"(f). La différence entre les courbes E" et E permet de localiser les pics: ils se situent aux endroits où cette différence est non nulle (figure 46). Ne sont cependant retenus que les pics qui vérifient un écart d'amplitude d'au moins 5 dB entre le sommet et les vallées de droite et de gauche.

On classe alors les bons pics localisés en 4 groupes suivant la position de leur sommet par rapport aux bandes 0-500Hz, 500-1000Hz, 1000-2000Hz et 2000-3000Hz. Arrivés à ce stade, on dispose en règle générale d'au plus un pic par bande. Dans les quelques cas contraires, on moyenne les pics de chaque bande afin de se ramener au cas général. En effet, deux pics voisins ne sont différentiables sur un spectrogramme que s'ils sont séparés d'au moins 350 Hz. Il est donc inutile de considérer deux pics dans une bande lorsque l'oeil n'en distingue qu'un seul.

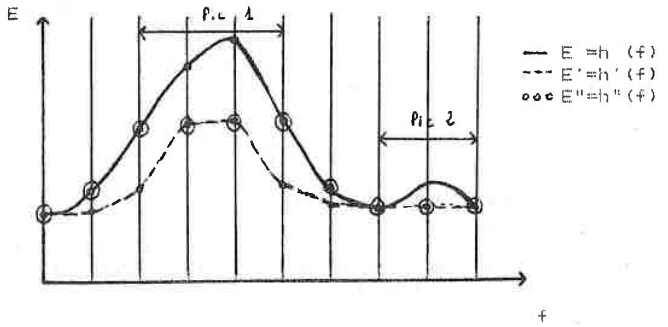


Figure 46 : Extraction des pics spectraux
(fenêtre de largeur 3)

Le dernier traitement concerne les pics des bandes 500-1000Hz et 1000-2000Hz. L'étude de spectrogrammes ayant montré que ces pics pouvaient facilement disparaître, leur existence est vérifiée en les comparant à ceux des bandes 0-500Hz et 2000-3000Hz qui sont quasiment toujours présents. Enfin on définit les composantes C4 et C5 en vérifiant la position des pics dans les bandes 0-500Hz et 500-1000Hz: s'ils débordent sur la bande voisine, on considère que l'on a à faire à un pic large, et les composantes C4 ou C5 sont validées.

7.4.1.2 Localisation des sonnantes

Les segments situés entre les noyaux vocaliques, qui n'ont pas été classés dans les classes fricatives ou plosives pendant la phase de prétraitement, sont considérés comme sonnantes.

7.4.1.3 Distinction /l/,/R/ et /m-n/

Elle est réalisée par la recherche des formants sur chaque spectre, et par application des règles d'affectation proposées par Lemoine en fonction de la configuration spectrale définie. La correction des frontières de chaque segment est comme dans l'étude précédente fonction de son homogénéité acoustique et de la durée vocalique moyenne.

7.4.2 La formalisation des connaissances

7.4.2.1 Le programme de contrôle

Remarque : nous ne traitons que la partie concernant les sonnantes.

```
(* Pré-classification voyelle, plosive, fricative *)

Frame PHONEME
      Specialization-of ROOT
      Type subframe
      Fill VOYELLE   Using P_NOYAUX_VOCALIQUES
      Fill PLOSIVE   Using P_PLOSIVES
      Fill FRICATIVE Using P_FRICATIVES
      With VOYELLE   Select PHON-VOYELLE
      With PLOSIVE   Select PHON-PLOSIVE
      With FRICATIVE Select PHON-FRICATIVE
      Otherwise AUTRE-PHONEME
End

(* Discrimination inter-sonnantes *)

Frame AUTRE-PHONEME
      Specialization-of PHONEME
      Type subframe
      Fill VOISEMENT Using P_CALC_PITCH
      Fill PIC1      Using P_PICS(0,500)
      Fill PIC2      Using P_PICS(500,1000)
      Fill PIC3      Using P_PICS(1000,2000)
      Fill PIC4      Using P_PICS(2000,3000)
      Fill PIC5      Using P_PICS(0,1000)
      With (VOISEMENT = 0) ior (PIC5 > 0)
```

```
            Select PHONEME-R
        Otherwise PHONEME-lmn
  End

  (* Phonème /R/ retenu *)

  Frame PHONEME-R
        Specialization-of AUTRE-PHONEME
        Type terminal
        Match With PROTOTYPE-R
  End

  (* Discrimination entre /l/ et /m-n/ *)

  Frame PHONEME-lmn
        Specialization-of AUTRE-PHONEME
        Type terminal
        Fill DP Using P_POLES(1,2,Energie)
        Match With PROTOTYPE-l,PROTOTYPE-mn
        Default "m-n"
  End
```

où :
- la procédure P_PICS(Param1,Param2) vérifie la présence d'un pic et d'un seul dans la bande de fréquences [Param1,Param2]. Si ce pic existe, sa position exacte est retournée, sinon la procédure retourne la valeur 0.
- la procédure P_CALC_PITCH indique le voisement du prélèvement: 1 pour prélèvement voisé et 0 pour prélèvement non voisé.
- la procédure P_POLES(Param1,Param2,Param3) est la même procédure que celle utilisée dans l'étude précédente.

7.4.2.2 Les frames PROTOTYPES

PROTOTYPE-l		PROTOTYPE-mn		PROTOTYPE-R	
VOYELLE	=0	VOYELLE	=0	VOYELLE	=0
PLOSIVE	=0	PLOSIVE	=0	PLOSIVE	=0
FRICATIVE	=0	FRICATIVE	=0	FRICATIVE	=0
VOISEMENT	=1	VOISEMENT	=1	VOISEMENT	[0,1]
PIC1	>280	PIC1	[0,350]	PIC1	=0
PIC2	>0	PIC2	>0	PIC2	=0
PIC3	>1400	PIC3	<1400	PIC3	>0

PIC4	>0	PIC4	>0	PIC4	>0
PIC5	=0	PIC5	=0	PIC5	>0
DP	<0	DP	>0		

7.4.3 Résultats obtenus

Résumés dans le tableau 47, ces résultats confirment les difficultés rencontrées dans l'identification des phonèmes /l/. Bien qu'ils aient une configuration stable, on n'en retrouve finalement que 60%. Certes, l'assimilation des /l/ sourds à la classe /R/ contribue à la médiocrité de ce taux de reconnaissance, mais la principale raison à ces mauvais résultas se trouve être l'ambigüité et l'imprécision des critères C2 et C3 dans la discrimination des classes /l/ et /m-n/. Il convient donc d'introduire des critères plus fins, comme la hauteur relative des pics ou leur largeur par exemple, afin de limiter ces insertions qui entraînent la perte de certains /l/.

	Taux de Reconnaissance	Omissions	Insertions
/l/	60%	40%	2%
/R/	86%	14%	10%
/m-n/	75%	25%	18%

Tableau 47

7.5 Conclusions

Ces deux études autorisent plusieurs commentaires:

- la simplicité et la clarté de la transcription des algorithmes de reconnaissance à l'aide des structures dont dispose LFRAP;

- la modularité du processus de raisonnement aussi bien dans l'extraction des indices acoustiques (procédures paramétrées) que dans le programme de contrôle (stratification en frames correspondant chacun à un noeud de la hiérarchie des indices);

- l'abstraction qui régit le processus de raisonnement: la conception du programme de contrôle se caractérise par le fait que l'on retarde le plus longtemps possible la prise de décisions à chaque étape de la reconnaissance, on procède par raffinement progressif;

- la fiabilité du système: les taux de reconnaissance obtenus dans chacune des méthodes sont comparables à ceux qui ont été obtenus par leurs auteurs.

LFRAP peut donc être considéré comme un véritable environnement de programmation orienté vers la reconnaissance phonétique, et de ce fait, répond bien à l'objectif que nous nous sommes fixé: un système de décodage acoustico-phonétique pour et par des phonéticiens.

Conclusion

Nous proposons dans le cadre de cette recherche un nouveau formalisme pour représenter et utiliser les compétences d'un expert phonéticien en décodage acoustico-phonétique de la parole continue: les frames.

Dans notre système, nous représentons l'ensemble des connaissances concernant un segment phonétique, les différents "slots" d'un frame contenant les indices acoustiques qui caractérisent ce segment.

Après la phase de segmentation, l'identification d'un segment s'effectue de la manière suivante:
- d'abord, le frame correspondant au segment est instancié à l'aide des procédures attachées aux différents slots;
- ensuite, le segment est identifié par comparaison du frame instancié avec un sous-ensemble de prototypes de référence; un score de compatibilité est calculé pour chaque prototype de référence comparé.

Extraction des indices, sélection des prototypes et comparaison (matching) sont controlés par les méta-connaissances inclues dans la grammaire de frames qui permet de mettre en oeuvre des raisonnements très élaborés. Par exmeple, si aucun des scores de compatibilité n'est acceptable, le système pourra recalculer certains indices ou modifier le poids relatif des indices dans le calcul des scores de compatibilité (cf. la notion d'indice plus ou moins net, d'une part, plus ou moins significatif d'autre part).

La grammaire de frames contient donc trois types de règles:
- des règles de construction d'instances de prototypes: extraction de paramètres acoustiques, sélection d'indices acoustiques pertinents;
- des règles de décision: par exemple, des règles de choix des prototypes de référence qui interviendront dans la phase d'identification du segment (matching);

- des méta-règles définissant la stratégie globale de décodage phonétique.

L'interêt essentiel de cette modélisation, qui met en oeuvre une méthode ascendante d'analyse, approche classique en décodage acoustico-phonétique, mais aussi une méthode d'analyse descendante, réside dans le fait qu'elle autorise les interactions indispensables entre les processus de segmentation et d'identification phonétique, grâce aux relations qui existent entre les notions de règle, de prototype et de procédure.

Plus généralement, la mise en oeuvre de langages définis sur des frames permet une imbrication étroite entre deux modes de représentation: règles et objets.

Nous poursuivons actuellement nos recherches dans deux directions parallèles:
- enrichissement de la base de connaissances acoustico-phonétiques à partir de la compétence de l'expert (informations utilisées, démarche), dans le cadre du développement du projet de système expert en décodage acoustico-phonétique APHODEX,
- développement de structures de contrôle élaborées, en particulier amélioration de l'interaction segmentation - identification, en vue de l'implantation d'un système de décodage phonétique encore plus efficace.

BIBLIOGRAPHIE

[ANDE-58] : Anderson T.W. 'An introduction to multivariate statitical analysis', New-York, Wiley, 1958.

[BAHL-78] : Bahl L.R., Baker J.K., Cohen P.S., Cole A.G, Jelinek F., Davies B.L. & Mercer R.L. 'Automatic recognition of continuously spoken sentences from a finite grammar ', IEEE ICASSP-78, Tulsa, pp 418-421, 1978

[BAHL-79] : Bahl L.R., Baker J.K., Cohen P.S., Cole A.G., Jelinek F., Davies B.L. & Mercer R.L. 'Recognition results with several experimental acoustic processor', IEEE ICASSP-79, Washington, pp 249-251, 1979

[BAKE-82] : Baker J.K. & Al 'Guidelines for performance assessment of speech recognizers', Acts of the Workshop on Standardization for Speech I/O Technology, Gaitghersburg, 1982

[BAKE-83] : Baker J.K., Pallet D.S. & Bridle J.S. 'Speech recognition performance assessment and available databases', IEEE ICASSP-83, Boston, 1983

[BELL-57] : Bellman R. 'Dynamic Programming', Princeton, Univ. Press., 1957

[BOBR-77a]: Bobrow D.G. & Winograd T. ' An overview of KRL, a Knowledge Representation Language', Cognitive Sciences No 1, pp 3-46, 1977

[BOBR-77b]: Bobrow D.G., Kaplan R.M., Kay M., Norman D.A., Thomson H. & Winograd T. 'GUS : a frame driven dialog system', Artificial Intelligence No 8, pp 155-173, 1977

[BOE-80] : Boe J.L., Abry C. & Corsi P. 'Les problèmes de normalisation interlocuteurs. Méthodes d'ajustement aux limites', 11 èmes JEP, Strasbourg, 1980

[BOYE-84] : Boyer A. 'Etude d'algorithmes de reconnaissance de mots enchainés', Rapport de DEA, Université de NANCY 1, 1984

[BRIA-83] : Briant N. & Flocon B. 'SYRIL : système temps réel de reconnaissance de mots isolés', Colloque EUPSICO, Erlangen, 1983

[BRID-82] : Bridle J.S., Brown N.D. & Chamberlain R.M. 'An algorithm for connected word recognition', IEEE ICASSP-82, Paris, pp 899-902, 1982

[BRID-83] : Bridle J.S. & Chamberlain R.M. 'Automatic labelling of speech using synthesis-by-rule and non linear time alignment', Speech Communication, No 2, pp 187-189, 1983

[CALL-82] : Callec A., Monne S., Querre M., Travarain O. & Mercier G. 'Automatic segmentation of phonetic units and training in the KEAL speech recognition system', IEEE ICASSP-82, Paris, pp 2000-2003, 1982

[CARB-83] : Carbonell N., Haton J.P., Lonchamp F. & J.M. Pierrel 'Elaboration d'un système expert pour le décodage phonétique de la parole', Speech Communication, No 2-3, pp 231-233, 1983

[CARB-84] : Carbonell N., Fohr D., Haton J.P., Lonchamp F. & J.M. Pierrel 'An Expert System for the Automatic Reading of French Spectrograms', IEEE ICASSP-84, San Diego, March 84

[CARB-85] : Carbonell N., Damestoy J.P., Fohr D., Haton J.P., Lonchamp F. & Pierrel J.M. 'Techniques d'intelligence artificielle en décodage acoustico-phonétique', 14 èmes JEP, Paris, pp 299-303, 1985

[CARR-79] : Carré R., Haton J.P. & Liénard J.S. 'Reconnaissance et synthèse de la parole - Etat de la recherche du développement', Les synthèses du SESORI, Septembre 1979

[CARR-84] : Carré R., Descout R., Eskénazi M., Mariani J. & Rossi M. 'The french language database : defining, planning and recording a large database', IEEE ICASSP-84, San Diego, 1984

[CART-74] : Carton F. 'Introduction a' la phonétique du français', Collection ETUDES No 303, Bordas, pp 22-25, 1974

[CHAF-83] : Chafcouloff M. 'A propos des indices de distinction des sons /l-R/ en français', Speech Communication, No 2, pp 137-139, 1983

[CHAF-84] : Chafcouloff M. 'Le polymorphisme acoustique du /R/ en français', 13 èmes JEP, Bruxelles, pp 165-166, 1984

[CHAR-78] : Charniak E. 'On the use of framed knowledge in language comprehension', Artificial Intelligence No 11, pp 225-265, 1978

[CHAR-85] : Charpillet F. 'Un système de reconnaissance de parole continue pour la saisie de textes lus', Thèse de doctorat d'université en informatique, Université de NANCY 1, 1985

[CHOL-82] : Chollet G.F. & Gagnoulet C. 'On the evaluation of speech recognizers and data bases using a reference system',

[COMB-81] : Combescure P. 'Vingt listes de dix phrases phonétiquement équilibrées', Revue d'acoustique 14, No 56, 1981

[COOK-76] : Cook C.C. 'Word verification in a speech understanding system', IEEE ICASSP-76, pp 553-556, 1976

[COOK-77] : Cook C.C. & Schwartz R. 'Advanced acoustic techniques in automatic speech understanding', IEEE ICASSP-77, Hartford, pp 663-666, 1977

[COOL-65] : Cooley J.W. & Tuckey J.W. 'An algorithm for the machine calculation of complex Fourrier series', Math. Comp. No 19, 1965

[DAME-83] : Damestoy J.P. 'Les performances des cepstres en reconnaissance de la parole', Rapport de DEA, Université de NANCY 1, 1983

[DAVI-80] : Davies S.B. & Mermelstein P. 'Comparison of parametric representation for monosyllabic word recognition in continuous spoken sentences', IEEE Trans. ASSP-28, No 4, pp 357-366, August 1980

[DELA-66] : Delattre P. 'Studies in french and comparative phonetics', Mouton, The Hague, 1966

[DELA-68] : Delattre P. 'From acoustic cues to distinctive features', PHONETICA, Vol. 18, pp 198-230, 1968

[DEMO-76] : De Mori R., Laface P. & Piccolo E. 'Automatic detection and description of syllabic features in continuous speech', IEEE Trans. ASSP October, pp 365-379, 1976

[DEMO-80] : De Mori R. & Giordano G. 'A parser for segmenting continuous speech into pseudo-syllabic nuclei', IEEE ICASSP-80, Denver, pp 876-879, 1980

[DEMO-83] : De Mori R. 'Extraction of acoutic cues using a grammar of frames', Speech Communication, No 2, pp 223-225, 1983

[DEMO-84] : De Mori R. & Mong Y.F. 'A system of plans for connected speech recognition', Proc. of the Conf. of American Association for Artificial Intelligence, Austin, 1984

[DEMO-85] : De Mori R., Rossi G. & Sun J. 'Multispeaker computer recognition of ten connectedly spoken letters', IEEE ICASSP-85, Tampa, pp 1225-1228, 1985

[DIBE-84] : Di Benedetto M.G. & Lanaro A. 'How to avoid vowel normalization in identification of vowels in continuous speech', IEEE ICASSP-84, San Diego, March 1984

[DIMA-84] : Di Martino J. 'Contibution à la reconnaissance globale de la parole : mots isolés et mots enchainés', Thèse de docteur ingénieur en informatique, Université de NANCY 1, 1984

[DIVO-85] : Divoux P. 'Utilisation de la classification ascendante hiérarchique en reconnaissance de mots isolés multilocuteurs', 5 ème congrés AFCET-RFIA, Grenoble, pp 309-318, 1985

[DIXO-77] : Dixon N.R. & Silverman H.F. 'The 1976 modular acoustic processor (MAP)', IEEE Trans. ASSP October, pp 367-379, 1977

[DOLO-84] : Dologlou Y. & Dolmazon J.M. 'Classification des sons au moyen de la prédiction linéaire et d'un mod`le du système auditif périphérique', 13 èmes JEP, Bruxelles, pp 141-142, 1984

[ELEN-82] : Elenius K. & Blomberg M. 'Effects of emphasizing transitional or stationary parts of the speech signal in discrete utterance recognition system', IEEE ICASSP-82, Paris, pp 535-538, 1982

[FANT-60] : Fant G. 'Acoustic theory of speech production', Mouton, The Hague, 1960

[FANT-73] : Fant G. 'Speech sounds and features', Cambridge MA, MIT Press, pp 32-83, 1973

[FOHR-85] : Fohr D., Haton J.P., Lonchamp F. & Sauter L. 'Méthodes de segmentation syllabique en reconnaissance de la parole', 14 èmes JEP, Paris, pp 164-167, 1985

[FUJI-62] : Fujimura O. 'Analysis of nasal consonants', JASA 34, pp 1865-1875, 1962

[FUJI-75] : Fujimura O. 'The syllabe as a unit of speech recognition', IEEE Trans. ASSP-23, pp 82-87, February 1975

[FURU-80] : Furui S. 'A training procedure for isolated words recognition systems', IEEE Trans. ASSP-28, No 2, 1980

[GAGN-82] : Gagnoulet C. & Couvrat M. 'SERAPHINE : a connected word recognition system', IEEE ICASSP-82, Paris, pp 887-889, 1982

[GANA-85] : Ganascia J.G. 'La conception des systèmes experts', La Recherche, No 170, pp 1142-1151, 1985

[GILL-84] : Gillet & All 'SERAC : un système expert en reconnaissance acoustico-phonétique', 4 ème congrés AFCET-RFIA, Paris, 1984

[GREE-84] : Green P.D. & Wood A.R. 'Knowledge-based speech understanding : towards a representation approach', Proc. ECAI, 1984

[GUBR-82] : Gubrynowicz R. 'Application de la théorie des sous-ensembles flous à l'analyse et la reconnaissance automatique de la parole', Rapport DT/LAA/TSS/RCP/101, CNET Lannion, 1982

[HATO-79] : Haton J.P. & Liénard J.S. 'La reconnaissance de la parole', La Recherche, No 99, 1979

[HATO-81] : Haton J.P. & Sanchez C. 'Méthodes synchrone et asynchrone en reconnaissance phonétique de la parole', Actes du séminaire 'Processus d'encodage et de décodage phonétiques', Toulouse, pp 144-155, 1981

[HATO-83] : Haton J.P. 'Reconnaissance des formes et intelligence artificielle', Cours DEA informatique, Université de NANCY 1, 1983

[ITAK-68] : Itakura F. & Saito S. 'Analysis synthesis telephony based on the maximum likelihood method', Proc. 6th ICA, paper C-5-5

[JAKO-63a]: Jakobson R., Fant G. & Halle M. 'Preliminaries to speech analysis', Cambridge MA, MIT Press, 1963

[JAKO-63b]: Jakobson R. 'Essai de linguistique générale', Edition de Minuit, 1963

[JELI-80] : Jelinek F., Mercer R.L. & Bahl L.R. 'Continuous speech recognition : statistical methods', Summer school on speech recognition CISM-IBM Italy, 1980

[JELI-81] : Jelinek F. 'Self-organized continuous speech recognizer', NATO ASI, Bonas, July 1981

[JOHA-83] : Johanssen J., Mac Allister J., Michaele T. & Ross S. 'A speech spectrogram expert', IEEE ICASSP-83, Boston, 1983

[JOHN-84] : Johnson S.R., Connoly J.H. & Edmonds E.A. 'Spectrogram analysis : a knowledge-based approach to automatic speech recognition', Proc. of Expert Systems 84, CUP, 1984

[KAUF-73] : Kaufman A. , 'Théorie des sous-ensembles flous', Vol. 1-5, Masson, Paris, 1973 à 1977

[KLAT-75] : Klatt D.H. 'Word verification in a speech understanding system', in 'Speech recognition by machine: a review', Proc. IEEE, pp 321-341, 1975

[LAFA-80] : Laface P. & De Mori R. 'Use of fuzzy algorithms for phonetic and phonemic labelling of continuous speech', IEEE Trans. PAMI, Vol. 2, pp 136-148, 1980

[LAUR-82] : Laurière J.L. 'Représentation et utilisation des connaissances', TSI No 1, pp 25-42, 1982

[LAZR-83] : Lazrek M. 'Décodage acoustico-phonétique en compréhension automatique de la parole continue', Thèse de 3 ème cycle, Université de NANCY 1, 1983

[LECO-79] : Lecorre C. & Vives R. 'Un programme de cadrage pour l'adaptation au locuteur en reconnaissance automatique de la parole', 3 ème congrés AFCET-RFIA, 1979

[LELI-81] : Lelièvre A. 'Classification hiérarchique automatique pour la création de phones ou de mots de référence en reconnaissance de la parole', 12 èmes JEP, Montréal, 1981

[LEMO-85] : Lemoine E. 'Recherche de paramètres discriminatoires pour les consonnes L, R, M et N dans un énoncé de parole continue', Rapport de DEA, Université de NANCY 1, 1985

[LOCH-81] : Lochsmidt B.F. 'Acoutic phonetic analysis using an articulary model', NATO ASI, Bonas, July 1981

[LOWE-76] : Lowerre B.T. 'The HARPY speech recognition system', PH. D. Thesis, Carnegie Melon University, Pittsburg PA, 1976

[LOWE-77] : Lowerre B.T. 'Dynamic speaker adaptation in the HARPY speech recognition system', IEEE ICASSP-77, Hartford, pp 788-790, 1977

[MAKH-72] : Makhoul J.I. & Wolf J.J. 'Linear prediction and the spectral analysis of speech', Bolt Beranek & Newman Inc., Report 2304, Cambridge MA, 1972

[MALB-72] : Malberg B. 'Phonétique française', Hermods Malmo Suède, End Edition, 1972

[MARI-81] : Mariani J. 'Reconnaissance de la parole continue par diphones', Actes du séminaire Processus d'encodage et de décodage phonétiques , Toulouse, 1981

[MARI-84] : Mari J.F. & Haton J.P. 'Some experiments in automatic recognition of a thousand words vocabulary', IEEE ICASSP-84, San Diego, March 1984

[MARK-72] : Markel J.D. 'Digital inverse filtering - a new tool for formant trajectory estimation', IEEE Trans. AU-20, pp 129-137, 1972

[MATS-79] : Matsumoto H. & Wakita H. 'Frequency warping for non uniform talker normalization', IEEE ICASSP-79, Washington, pp 566-569, 1979

[MELO-82] : Meloni H. 'Contribution à la recherche sur la reconnaissance automatique de la parole continue', Thèse de docteur d'état es sciences, Université d'AIX-MARSEILLE 2, 1982

[MEMM-84] :
Memmi D., Eskénazi M., Mariani J. & Stern P. 'SONEX : système expert en lecture de spectrogrammes', Rapport ATP Intelligence artificielle, 1984

[MERC-82] : Mercier G. 'Acoustic-phonetic decoding and adaptation in continuous speech recognition', in Automatic Speech Analysis and Recognition , J.P. Haton Ed., pp 69-99, 1982

[MERM-75a]: Mermelstein P. 'A phonetic-context controlled strategy for segmentation and phonetic labelling of speech', IEEE Trans. ASSP-23, pp 79-82, February 1975

[MERM-75b]: Mermelstein P. 'Automatic segmentation of speech into syllabic units', JASA 58, pp 880-883, 1975

[MINS-75] : Minsky M. 'A framework for representing knowledge', in 'The Psychology of Computer Vision', pp 211-277, P.H. Winston Ed., 1975

[MOOR-73] : Moore J. & Newell A. 'How can MERLIN understand ?', in Knowledge and Cognition , L. Gregg Ed., Hillsdale, 1973

[MYER-81] : Myers C. & Rabiner L.R. 'A level building dynamic time warping algorithm for connected word recognition', IEEE Trans.ASSP 29, April 1981

[NADA-81] : Nadas A. & Al 'Continuous speech recognition with automatically selected acoustic prototypes obtained by either bootstrapping or clustering', IEEE ICASSP-81, Atlanta, 1981

[NAKA-83] : Nakagawa S.I. 'A connected spoken word recognition method by o(n) dynamic programming pattern matching algorithm', IEEE ICASSP-83, Boston, pp 296-299, 1983

[NEAR-77] : Neary T. 'Phonetic features systems for vowels', Doct. Diss., University of Connecticut, 1977

[NEEL-83] : Neel F., Eskénazi M. & Mariani J. 'Cadrage automatique pour la constitution de dictionnaires d'entités phonétiques', Speech Communication, No 2, pp 193-195, 1983

[NOIR-85] : Noiré J. 'Reconnaissance automatique de mots isolés pour de grands vocabulaires', Thèse de docteur ingénieur CNAM, Nancy, 1985

[PARZ-62] : Parzen E. 'On estimation of a probability density function and mode' Ann. Math. Stat., Vol. 33, 1962

[PERE-81] : Perennou G. & Decalmes M. 'Le décodage au niveau phonologique dans ARIAL II', Actes du séminaire 'Processus d'encodage et de décodage phonétiques', Toulouse, 1981

[PIST-84] : Pister C. 'Adaptation au locuteur par apprentissage automatique - Application à un système de reconnaissance automatique de la parole', Thèse de 3 ème cycle, Université de NANCY 1, 1984

[RABI-76] : Rabiner L.R., Cheng M.J., Rosenberg A.E. & Mac Gonegal C.A. 'Comparative performance stydy of several pitch detection algorithms', IEEE Trans. ASSP-24, pp 399-417, October 1976

[RABI-78] : Rabiner L.R. & Rosenberg A.E. 'Considerations in dynamic time warping algorithms for discrete word recognition', IEEE Trans. ASSP-26, No 6, December 1978

[RABI-80] : Rabiner L.R. & Wilpon J.G. 'A simplified, robust training procedure for speaker trained, isolated words, recognition systems', JASA 68, pp 1271-1276, 1980

[REDD-76] : Reddy D.R. 'Speech recognition by machine : a review', Proc. IEEE, pp 501-531, Avril 1976

[ROSE-81] : Rosenberg A.E., Rabiner L.R., Levinson S.E. & Wilpon J.G. 'A preliminary study of the use of demi-syllabes in automatic speech recognition', IEEE ICASSP-81, Atlanta, pp 967-970, 1981

[ROSS-77] : Rossi M. & De Cristo A. 'Proposition pour un modèle d'analyse de l'intonation', 8 èmes JEP, Aix-en-Provence, 1977

[RUSK-81] : Ruske G. & Schotola T. 'The efficiency of demi-syllabes segmentation in the recognition of spokken words', IEEE ICASSP-81, Atlanta, pp 971-974, 1981

[RUSK-82] : Ruske G. 'Automatic recognition of syllabic speech segmentation using spectral and temporal features', IEEE ICASSP-82, Paris, pp 550-553, 1982

[SAKO-71] : Sakoe H. & Chiba S. 'A dynamic programming optimization for spoken word recognition', IEEE Trans. ASSP, No 26, pp 43-49, 1971

[SAKO-78] : Sakoe H. & Chiba S. 'Dynamic programming algorithm optimization for spoken word recognition', IEEE Trans. ASSP 26, No 1, pp 43-49, Februar 1978

[SAMB-75] : Sambur M.R. & Rabiner L.R. 'A speaker independant digit recognition system', The Bell System Technical Journal, pp 81-102, January 1975

[SAND-78] : Sandewall E. 'An approach to the frame problem, and its implementations', Machine Intelligence No 7, pp 195-204, 1972

[SAUT-84] : Sauter L. 'RAPACE : un système de reconnaissance analytique de la parole continue', 4 ème congrés AFCET-RFIA, Paris, pp 89-98, 1984

[SCHA-72] : Schafer R.W. 'A survey of digital speech processing techniques', IEEE Trans. AU-20, 4, pp 28-37, 1972

[SCHA-77] : Schank R. & Abelson R. 'Scripts, plans, goals and understanding', Lawrence Erlbaum Assoc., 1977

[SCHW-71] : Schwartz R.M. 'Automatic normalization for recognition of all speakers', S. D. Thesis MIT, Cambridge MA, 1971

[SCHW-76] : Schwartz R.M. 'Acoustic-phonetic experiment facility for the study of continuous speech', IEEE ICASSP-76, Philadelphia, pp 12-14, 1976

[SCHW-76] : Schwartz R.M. & Zue V. 'Acoustic phonetic recognition in BBN SPEECHLIS', IEEE ICASSP-76, Philadelphia, pp 21-24, 1976

[SHIR-78] : Shirai K. & Honda H. 'Feature extraction for speech recognition based on articulary model', Proc. 4th ICJPR, pp 1064-1068, 1978

[SHOU-80] : Shoup J. 'Phonological aspects of speech recognition', in 'Trends in Speech Recognition', W. Lea Ed., Prentice Hall, 1980

[SOUD-85] : Soudoplatoff S. 'Classification de phonèmes par méthodes non paramétriques', 14 èmes JEP, Paris, pp 308-310, 1985

[STER-85] : Stern P.E., Eskénazi M. & Memmi D. 'Elaboration d'un système expert en lecture de sonagrammes', 14 èmes JEP, Paris, pp 295-298, 1985

[STEV-71] : Stevens K.N. 'Sources of inter and intra-speaker variability in the acoustic properties of speech sounds', 7th Int. Cong. Phon. Sciences, pp 206-227, 1971

[WAGN-81] : Wagner M. 'Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms', IEEE ICASSP-81, Boston, pp 1156-1159, 1981

[WAKI-77] : Wakita H. 'Normalization of vowels by vocal tract length and its application to vowel identification', IEEE Trans. ASSP-25, No 2, pp 183-192, April 1977

[WEIN-75] : Weinstein C.J., Mac Candless S.S., Mondshein L.F. & Zue V. ' A system for acoustic phonetic analysis of continuous speech', IEEE Trans. ASSP 23, pp 54-67, February 1975

[WIEZ-82] : Wiezlak W.N. & Gubrynowicz R. 'Articulatory description of speech signal in isolated word recognizer', IEEE ICASSP-82, Paris, pp 925-930, 1982

[WILL-70] : Williams C.E., Stevens K.N. & Hecker M.H.L. 'Acoustical manifestations of emotional speech', JASA 47, 1970

[WISW-76] : Wiswanathany R., Makhoul J. & Schwartz R.M. 'Medium and low bit rate speech transmission', NATO ASI, Bonas, July 1981

[WOOD-76] : Woods W.A. 'Speech understanding systems : final report', Vol. 1-5, BBN Report No 3438, AD Nos 1:A035165, 2:A035166, 3:A035167, 4:A035277, 5:A035278, December 1976

[ZADE-78] : Zadeh L.A. 'Fuzzy sets as a basis for a theory of possibility', Fuzzy sets and systems, No 1, pp 3-28, 1978

[ZUE-76] : Zue V. 'Acoustic caracteristics of stop consonants : a controlled study', S. D. Thesis MIT, Cambridge MA, 1976

[ZUE-78] : Zue V. & Schwartz R.M. 'Acoustic processing and phonetic analysis', in 'Trends in Speech Recognition', W. Lea Ed., Prentice Hall, 1978,

[ZUE-82] : Zue V. 'Acoutic phonetic knowledge representation : implications from spectrogram reading experiments', in 'Automatic Speech Analysis and Recognition', J.P. Haton Ed., D. Reidel, 1982

[ZWIC-78] : Zwicker E., Terhardt E. & Paulus E. 'Automatic speech recognition using psychoacoustic models', JASA 65, pp 487-498, 1978

NOM DE L'ETUDIANT : DAMESTOY Jean-Paul

NATURE DE LA THESE : Doctorat de l'Université de NANCY I en Informatique

NANCY, le - 9 JAN. 1986

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

Le Président
R. MAINARD

RESUME

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

Nous proposons dans cet exposé une nouvelle méthode pour représenter et utiliser les connaissances dans le décodage phonétique de la parole continue. Notre système utilise la notion de prototype ou frame, initialement introduite pour le traitement des langages naturels, pour formaliser l'expertise d'un phonéticien en lecture de spectrogrammes de parole.

L'intéret des frames se situe aussi bien dans l'utilisation de différents types d'analyses acoustiques et de traits phonétiques que dans le controle du processus de reconnaissance, plus sophistiqué que dans les systèmes classiques.

Mots clé : Décodage phonétique
Onde acoustique de parole
Représentation des connaissances
Frames
Prototypes