THESE

présentée à

L'UNIVERSITE DE NANCY I-FACULTE DES SCIENCES

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITE DE NANCY I EN INFORMATIQUE

par

Jean-Claude JUNQUA

Sujet de la thèse :

CONTRIBUTION A L'AMELIORATION DE LA ROBUSTESSE DES SYSTEMES DE RECONNAISSANCE AUTOMATIQUE DE MOTS ISOLES

Soutenue le 19 Mai 1989 devant le jury composé de :

MM.	**R. MOHR**	Président
	J.C. DERNIAME	Rapporteurs
	NNOU	
	TON	Examinateurs
	ON	
	J. MARIANI	
	J.M. PIERREL	
	H. WAKITA	

N° d'ordre:

THESE

présentée à

L'UNIVERSITE DE NANCY I-FACULTE DES SCIENCES

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITE DE NANCY I EN INFORMATIQUE

par

Jean-Claude JUNQUA

Sujet de la thèse :

CONTRIBUTION A L'AMELIORATION DE LA ROBUSTESSE DES SYSTEMES DE RECONNAISSANCE AUTOMATIQUE DE MOTS ISOLES

Soutenue le 19 Mai 1989 devant le jury composé de :

MM.	R. MOHR	Président
	J.C. DERNIAME G. PERENNOU	Rapporteurs
	M.C. HATON J.P. HATON J. MARIANI J.M. PIERREL H. WAKITA	Examinateurs

FOLLOW YOUR DREAMS

If while pursuing distant dreams
Your bright hopes turn to gray
Don't wait for reassuring words
Or hands to lead the way

For seldom will you find a soul
With dreams the same as yours.
Not often will another help you
Pass through untried doors.

If inner forces urge you
To take a course unknown,
Be ready to go all the way,
Yes, all the way alone.

That's not to say you shouldn't
Draw lessons from the best;
Just don't depend on lauding words
To spur you on your quest.

Find confidence within your heart
And let it be your guide.
Strive ever harder toward your dreams
And they won't be denied

Bruce B. Wilmer

Wouldn't it be wonderful if
science really were true ...

Richard Bach

Vers l'infini où tout se tient

A tout ceux qui m'ont
appris l'honnêteté du
coeur, avec une pensée
spéciale pour mon père.

Plan

Liste des Figures

Liste des Tables

Liste des Equations

Remerciements

Les remerciements sont aux individus un peu ce que la bibliographie est au travail présenté : on a toujours peur d'oublier quelqu'un! En quelques lignes je vais essayer de faire de mon mieux pour être le plus complet possible. Le travail rapporté dans ce document a été réalisé en grande partie dans le laboratoire Speech Technology Laboratory (STL) situé à Santa-Barbara en Californie. L'environnement de travail que ce soit technique ou humain que j'ai pu y trouver fut des plus favorables.

Au terme de ce travail je voudrais tout d'abord exprimer mes remerciements à ceux qui me font l'honneur et l'amitié de participer à mon jury,

Monsieur Jean-Paul HATON, Professeur à l'Université de Nancy I, aujourd'hui Directeur de Recherche à l'INRIA et mon responsable de recherche, qui est à l'origine du travail présenté dans ce document. Beaucoup apprécient ses qualités humaines et professionnelles. Je me bornerai à dire en quelques mots ce qui en nécessiterait davantage : merci encore, il y a des choses que l'on n'oublie pas!

Monsieur Hisashi WAKITA, Président du laboratoire STL et Professeur au département de linguistique de l'Université de Californie à Santa-Barbara, qui me fait le grand honneur de participer à mon jury malgré l'éloignement. Ceci est pour moi une preuve supplémentaire de son intérêt à mon travail ainsi que de son amitié. J'ai beaucoup apprécié le temps que nous avons passé ensemble autant du point de vue professionnel que personnel et j'ai souvent été très impressioné par ses qualités qui vont beaucoup au delà de celles d'un chef d'entreprise,

Madame Marie-Christine HATON, Maître de Conférences à l'Université de Nancy I, qui a souvent été pour moi un modèle pour sa puissance de travail, sa gentillesse et sa disponibilité malgré les nombreuses tâches qu'elle assume. Sa présence dans mon jury est pour moi une grande joie,

Monsieur Roger MOHR, aujourd'hui Professeur à l'INPG de Grenoble, que je remercie particulièrement pour avoir bien voulu présider ce jury de thèse. J'en suis heureux à double titre qui tiennent à ses qualités professionnelles mais aussi et surtout pour la personne qu'il représente et que j'admire sur beaucoup de points,

Monsieur Guy PERENNOU, Professeur à l'Université Paul Sabatier a Toulouse, qui me fait un grand honneur d'avoir accepté de juger mon travail. Le côté toulousain qui me rappelle quelques souvenirs ..., sa personnalité ainsi que ses qualités professionnelles sont autant d'éléments que j'apprécie,

Monsieur Jean-Claude DERNIAME, Professeur à l'Université de Nancy I, qui a suivi mes premiers pas à l'informatique alors que j'étais élève ingénieur à l'ENSEM et que je remercie pour avoir accepté de rapporter sur mon travail,

Monsieur Joseph MARIANI, Directeur de Recherche au CNRS, que je remercie d'avoir accepté d'être un membre du jury malgré les nombreuses fonctions qui sont les siennes,

Monsieur Jean-Marie PIERREL, Professeur à l'Université de Nancy I, qui me fait l'honneur et l'amitié de participer à ce jury.

Je voudrais aussi associer dans une même pensée amicale,

Hynek HERMANSKY, anciennement à STL et aujourd'hui à USWEST dans le Colorado, avec qui j'ai eu le plaisir de travailler lors de mes débuts à STL. La partie optimisation de l'analyse PLP présentée au chapitre E.2 a été réalisée en commun,

Anne BOYER pour la relecture de quelques chapitres à une heure tardive,

Monsieur Jean-Pierre FINANCE, Professeur à l'Université de Nancy I et actuellement Directeur du CRIN, qui a contribué à rendre possible mon séjour à STL,

Gérald MASINI, mon "copain" de bureau qui m'a tenu au courant de l'évolution du CRIN pendant le temps que j'ai passé aux Etats-Unis et qui au fil des années est devenu un ami que j'apprécie beaucoup,

Odile MELLA à qui je dois énormément ...

Enfin, et ce ne sont pas les moindres, je voudrais remercier mes parents, ma famille et tous mes amis du CRIN de STL ou d'ailleurs qui m'ont aidé, encouragé et soutenu dans les périodes de doute. L'amitié qu'ils m'ont apporté va bien au delà de ce travail et est pour moi une source d'espoir en l'avenir.

Avant-Propos

L'étude de la parole fait appel à des domaines pluridisciplinaires et le contenu de cette thèse ne fait pas exception à la règle. La production de la parole a été étudiée par les *phonéticiens* et les *linguistes* et les mécanismes de l'oreille par les *physiologistes* et les *psychophysicistes*. Les *linguistes* ont apporté leur connaissance du language et les *ingénieurs* et les *physiciens* ont permis la réalisation de modèles fonctionnels. Toutes ces disciplines ont contribué à une meilleure compréhension de la parole.

Dans cette thèse nous nous sommes rendu compte que, pour construire des systèmes de reconnaissance automatique de la parole, il était nécessaire d'intégrer plusieurs sources de connaissances liées à des domaines divers. Aussi, nous nous sommes intéressés aux différents domaines qui nous ont paru approprié pour résoudre les problèmes posés. Plus particulièrement, les résultats obtenus en physiologie et psychoacoustique ainsi que la phonétique, le traitement du signal et l'intelligence artificielle ont fait l'objet de nos études. Au cours de ce travail nous sommes devenus convaincu que l'étude de la parole nécessitait l'intégration de différentes expertises afin de modéliser les problèmes rencontrés. Nous montrons dans cette thèse l'intérêt d'une telle approche.

Cependant, comme nous nous sommes intéressés à plusieurs domaines, nous sommes conscients que certains de nos travaux auraient peut-être nécessité plus d'approfondissement. Certaines idées developpées, ainsi que quelques-uns des résultats obtenus, méritent des études complémentaires. Nous avons délibérément choisi une telle approche au détriment sans doute d'une recherche plus thématique.

Une bonne partie de ce document est consacrée à la description des tests effectués et des résultats obtenus. Ceux-ci sont nombreux car nous avons voulu présenter nos travaux et les performances obtenues de façon précise.

Avant d'aller plus loin, nous tenons à souligner que l'écriture de cette thèse a bénéficié grandement des travaux de nombreuses personnes. Plus particulièrement, de nombreux articles et livres ont facilité l'écriture des premiers chapitres consacrés à l'introduction des connaissances nécessaires à la compréhension de nos travaux.

Tout au long de cette thèse, le lecteur pourra trouver de nombreuses figures. Ces figures contiennent des commentaires qui sont généralement en anglais pour des raisons de double emploi. Nous tenons à nous en excuser. Enfin, nous ne voulions pas que ce travail soit *seulement* scientifique. Aussi, afin de lui donner une note plus personnelle et plus poétique nous avons inclus quelques poèmes et citations. Comme le dit si bien une thèse lue récemment "*une thèse ce n'est pas la vie mais c'est la vie qui la permet*". La note personnelle que nous avons voulu lui ajouter va dans ce sens là.

RESUME

L'expérience montre que des systèmes de reconnaissance automatique de la parole qui donnent de bons résultats dans des environnements de laboratoire voient leurs performances se dégrader lorsqu'ils opèrent dans des conditions réelles. Ceci est lié à l'existence de nombreux problèmes souvent sous-estimés comme 1) la variabilité intra- et interlocuteur 2) l'existence de bruit de fond, et 3) la difficulté des vocabulaires étudiés. Les recherches liées à ces problèmes ont été regroupées sous la rubrique "amélioration de la robustesse des systèmes de traitement de la parole". Le travail présenté s'inscrit dans le cadre de "*l'amélioration de la robustesse des systèmes de reconnaissance de mots isolés*". Après une étude visant à déterminer le modèle d'analyse acoustique le moins sensible à la variabilité intra- et interlocuteur, les problèmes 2 et 3 sont étudiés. Certaines des méthodes développées sont ensuite appliquées à l'étiquetage automatique de la parole.

Initialement, un modèle auditif appelé, analyse par prédiction linéaire perceptivement fondée (*PLP*), est étudié. Après une évaluation de ce modèle par comparaison à d'autres modèles utilisés récemment, il est montré que la technique *PLP* permet d'obtenir d'aussi bons ou de meilleurs résultats que les autres modèles étudiés. Une optimisation de ce modèle, grâce à l'utilisation de caractéristiques spectrales dynamiques du signal de parole et d'une nouvelle mesure de distance, est ensuite proposée.

Une étude comparative de plusieurs modèles d'analyse acoustique en environnement bruité montre que ce modèle est néammoins très sensible au bruit. Ceci a amené le développement d'un modèle physiologique à synchronisation temporelle (*SLP*). Il est montré que ce modèle donne de meilleurs résultats que les autres modèles étudiés (sous certaines conditions) lorsque le rapport signal-sur-bruit est faible. Une extension de cette étude à de la parole prononcée dans du bruit (effet Lombard), montre que les changements phonétiques occasionnés par l'effort vocal dégradent beaucoup les performances. Dans un tel contexte, la technique *PLP* donne de meilleurs résultats que la technique de prédiction linéaire (*LP*).

Afin de prendre en compte le problème des vocabulaires difficiles, un système hybride : *ORION*, permettant la discrimination entre des mots acoustiquement similaires, a été développé. Grâce à l'utilisation de connaissances phonétiques, ce système améliore considérablement les performances obtenues à l'aide de systèmes conventionnels.

Enfin, un système de segmentation automatique, réalisé dans le cadre du système *ORION*, a été étendu à l'étiquetage automatique de la parole. La particularité de ce système est d'utiliser plusieurs sources de connaissances qui communiquent à travers un système "blackboard" administré par un contrôle hiérarchique.

Mots clés

Reconnaissance de la parole, mots isolés, robustesse, modèle d'analyse acoustique, mesure de distance, discrimination, source de connaissances, physiologie, psychoacoustique, phonétique, segmentation automatique.

Introduction

La parole est notre mode de communication privilégié. Pourtant les mécanismes sous-jacents à ce mode de communication sont encore mal connus. Depuis plus de quarante ans, beaucoup de chercheurs se sont intéressés à la reconnaissance de la parole par ordinateur. C'est un domaine ou la connaissance des mécanismes liés à la production et à la perception de la parole est essentielle. La conception d'un message dans le cerveau de l'émetteur jusqu'à la reception de celui-ci dans le cerveau du récepteur est possible grâce à de tels mécanismes. La parole est produite en utilisant les poumons et les cordes vocales (qui obstruent complètement ou partiellement l'air provenant des poumons) comme sources d'excitation et en coordonnant les organes articulatoires (langue, lèvres, mâchoires, etc). L'ensemble, modifiant la forme du conduit vocal, permet la production des différents sons. Ces sons, qui correspondent à une variation de pression provenant de la bouche de l'émetteur, atteignent les oreilles du récepteur qui les décodent pour reconstituer le message émis.

Les chercheurs ont étudié les mécanismes de production de la parole afin de les modéliser pour les appliquer à l'analyse et à la synthèse de la parole. Le principal objectif des études sur la perception de la parole est d'essayer de comprendre comment le système auditif humain convertit un signal acoustique continu en messages discrets. Une hypothèse généralement avancée est que le signal acoustique est mis en correspondance avec des codes phonétiques discrets appelés phonèmes. Cependant, une telle mise en correspondance n'est à l'heure actuelle pas très claire car la parole est très encodée, et une représentation discrète est très difficile à formuler. Une voie de recherche prometteuse, liée à ce problème de mise en correspondance, est l'étude de l'encodage de la parole au niveau des fibres nerveuses du système auditif.

Dans les différents travaux menés en traitement automatique de la parole, le signal de parole a été codé sous différentes formes. La forme de représentation utilisée influence beaucoup la conception des systèmes de traitement. D'un point de vue linguistique, la parole est représentée à l'aide d'une chaîne de symboles phonétiques discrets. Ces symboles, appelés phonèmes, sont supposés avoir d'uniques caractéristiques acoustiques et articulatoires. D'un point de vue physique, la parole est représentée sous la forme d'un signal temporel. Dans ce cas, aucune connaissance phonétique n'est prise en considération. Inversement, lorsque la parole est stockée par l'intermédiaire de paramètres fréquentiels, ce qui est souvent le cas en reconnaissance automatique de la parole, il est important de tenir compte des propriétés acoustiques du signal analysé ainsi que de la façon dont l'être humain les perçoit. Typiquement, les sons sont divisés en deux grandes catégories : les voyelles qui découlent d'un flot d'air ininterrompu dans le conduit vocal, et les consonnes, pour lesquelles le flot d'air est interrompu quelque part dans le conduit vocal. Comme le signal de parole varie continuellement, il est nécessaire de tenir compte des transitions entre sons. Ceci entraîne souvent la prise en considération de propriétés articulatoires. Certaines recherches se sont intéressées à la modélisation du conduit vocal [Wak73, Cok76] plutôt qu'à la modélisation de sa production acoustique. La forme du conduit vocal détermine les caractéristiques des sons produits.

Au cours de ce travail, le signal de parole a été représenté à l'aide de paramètres fréquentiels ou de propriétés acoustiques et phonétiques. Des connaissances sur la façon dont l'être humain perçoit les sons ont été utilisées. Par contre, nous n'avons pas manipulé le signal de parole sous la forme de ses propriétés articulatoires. Ce modèle de représentation est plus utilisé en synthèse qu'en reconnaissance de la parole.

Cette thèse est organisée en sept parties. Les parties A, B, et C sont consacrées à des rappels sur les propriétés acoustiques et phonétiques de la parole dans le cadre de l'anglais américain, ainsi qu'aux mécanismes de perception, aux techniques d'analyse, aux mesures de distance et plus généralement à la reconnaissance automatique de mots isolés. La partie D présente les outils développés dans le cadre de ce travail. En particulier, un système modulaire de reconnaissance de mots isolés (*ORION*) intégrant différentes sources de connaissances, un logiciel d'analyse et de traitement de la parole (*STAR*), et un algorithme de segmentation automatique (*SAIPH*) sont détaillés. Ensuite, dans la partie E, les études et les réalisations effectuées pour améliorer la robustesse des systèmes de reconnaissance de mots isolés sont présentées. Après une étude comparative visant à déterminer le modèle d'analyse acoustique le moins sensible à la variabilité intra- et interlocuteur, une optimisation du modèle d'analyse perceptivement fondé *PLP* est proposée. Il est montré que l'utilisation de caractéristiques spectrales dynamiques et d'une nouvelle mesure de distance adaptée à la technique *PLP* améliore les scores de reconnaissance. Toutefois, une étude comparative de plusieurs modèles d'analyse acoustique en environnement bruité montre que ce modèle est très sensible au bruit. Ceci a amené le développement d'un modèle physiologique à synchronisation temporelle (*SLP*). Ce modèle donne de meilleurs résultats que les autres modèles étudiés (sous certaines conditions) lorsque le rapport signal-sur-bruit est faible. Une extension de cette étude à de la parole prononcée dans du bruit (effet Lombard), montre que les changements phonétiques occasionnés par l'effort vocal dégradent beaucoup les performances. Dans un tel contexte, la technique *PLP* donne de meilleurs résultats que la technique de prédiction linéaire (*LP*). Afin de prendre en compte le problème des vocabulaires difficiles, un système hybride : *ORION*, permettant la discrimination entre des mots acoustiquement similaires, a été développé. Grâce à l'utilisation de connaissances phonétiques, ce système améliore considérablement les performances obtenues à l'aide de systèmes conventionnels. La partie F décrit un système d'étiquetage automatique de la parole qui utilise plusieurs sources de connaissances communiquant à travers un système "blackboard" administré par un contrôle hiérarchique. Enfin, la partie G résume les points importants de ce travail tout en proposant des extensions.

La Parole

Alors un lettré dit : Parlez-nous de la Parole,
Et il répondit, disant :
Vous parlez lorsque vous cessez d'être en paix avec vos pensées;
Et lorsque vous ne pouvez rester davantage dans la solitude de votre coeur vous vivez dans vos lèvres, et le son est un divertissement et un passe-temps.
Et dans une large part de vos discours, la pensée est à moitié assassinée.
Car la pensée est un oiseau de l'espace, qui dans une cage de mots peut ouvrir ses ailes et ne peut voler.

Khalil Gibran

PARTIE A
NATURE ET PERCEPTION DES SONS

Après avoir abordé certaines caractéristiques acoustiques et phonétiques des sons de la parole, une représentation très utilisée en parole appelée *spectrogramme* est introduite. L'ensemble des sons peuvent être décrits à partir d'une telle représentation.

Ensuite, une introduction aux mécanismes auditifs et à la perception de la parole est présentée. Après une brève description du système auditif périphérique, les mécanismes, souvent inclus dans les modèles auditifs, qui résultent de travaux menés dans des domaines liés à la psychoacoustique et à la physiologie, sont énoncés.

Le but des descriptions se trouvant dans cette partie est de donner au lecteur les définitions des termes utilisés dans ce document. De plus amples détails sur les sujets abordés pourront être trouvés dans les articles ou livres référencés.

Chapitre A.1 ACOUSTIQUE ET PHONETIQUE

Là-bas, dans la lumière du soleil,
vivent mes plus hautes aspirations.
Je ne les atteindrai peut-être pas mais
je peux lever la tête et en voir la beauté,
y croire et m'efforcer de suivre la voie
qu'elles me montrent.

Louisa May Alcott

A.1.1 Introduction

Les sons d'un langage donné sont généralement décrits à l'aide d'unités appelées phonèmes. Un phonème peut être défini comme étant l'unité minimale telle qu'un changement de phonème résulte en un changement, au niveau du lexique, du mot auquel il appartient. Les variations de prononciation que l'on peut observer entre plusieurs locuteurs sont appelées *accents*. Lorsque ces variations sont aussi dues à des différences de syntaxe, de vocabulaire, et de morphologie, on les appelle *dialectes*.

Dans ce chapitre, les phonèmes de l'anglais américain sont présentés ainsi que les caractéristiques acoustiques de ces différents phonèmes. Les études décrites dans ce document ont été réalisées sur de la parole produite par des locuteurs américains car toutes les bases de données qui étaient disponibles lors des travaux présentés étaient en anglais américain. Cependant, nous verrons par la suite que les conclusions avancées ne sont probablement pas liées (ou très peu) au langage utilisé. Toutefois, avant de présenter les voyelles et les consonnes de l'anglais américain les phonèmes du français sont décrits rapidement en insistant plus particulièrement sur quelques différences entre les deux langages.

A.1.2 Quelques éléments de phonétique française

Il existe trente à quarante phonèmes pour la langue française, selon les locuteurs et la région. Ceux-ci peuvent être rangés en différentes catégories qui peuvent varier suivant les propriétés (articulatoires, acoustiques, etc) auxquelles on se réfère. La table 1 présente[1] une classification possible souvent utilisée.

[1] Certains symboles sont des approximations des symboles réels à cause des limitations du système de traitement de textes utilisé.

Phonèmes	Classe	Exemple
a	voyelle orale	patte
a	voyelle orale	pâte
i	voyelle orale	riz
y	voyelle orale	rue
ɔ	voyelle orale	port
o	voyelle orale	pot
ə	voyelle orale	le
ɛ	voyelle orale	raie
e	voyelle orale	ré
ø	voyelle orale	peu
œ	voyelle orale	oeuf
u	voyelle orale	roue
ã	voyelle nasale	blanc
ɔ̃	voyelle nasale	bon
ɛ̃	voyelle nasale	vin
œ̃	voyelle nasale	un
j	semi-consonne	iode
	semi-consonne	lui
w	semi-consonne	moi
p	plosive	passe
t	plosive	toux
k	plosive	cou
b	plosive	basse
d	plosive	doux
g	plosive	goût
m	nasale	masse
n	nasale	nous
μ	nasale	signe
f	fricative	fer
s	fricative	assis
ʃ	fricative	chou
v	fricative	verre
z	fricative	Asie
ʒ	fricative	joue
l	liquide	la
r	liquide	rat

Table 1 Les phonèmes du français.

Dans cette liste de phonèmes notons les voyelles nasales et le fait qu'en français moderne il n'exite pas de diphtongues. Ce sont d'importantes différences par rapport à l'anglais. Le français, à l'inverse de l'anglais (ou de l'allemand) n'est pas un langage à accent d'intensité. Dans ces langages les voyelles qui ne sont pas accentuées sont souvent réduites, c'est-à-dire articulées de façon moins extrême. Au niveau du français l'articulation est plus uniforme. Il faut toutefois tenir compte de l'accent d'insistance optionnel en début de mot.

A.1.3 Les voyelles de l'anglais américain

Les voyelles de l'anglais américain forment un continuum souvent représenté à l'aide d'un graphe [Lad82] qui précise leur qualité. Il n'y a pas de frontières distinctes entre les différentes voyelles. Ces frontières sont très influencées par les différences entre locuteurs. Afin que les phonéticiens puissent décrire les voyelles, Jones proposa [Jon57] les *voyelles cardinales* qui représentent des voyelles de référence. La table 2 présente[2] les voyelles de l'anglais américain.

[2] Certains symboles sont des approximations des symboles réels à cause des limitations du système de traitement de textes utilisé.

Phonème	Arpabet	Classe	Exemple
i	IY	voyelle	beat
I	IH	voyelle	bit
ϵ	EH	voyelle	bet
æ	AE	voyelle	bat
a	AA	voyelle	bob
$\supset$	AO	voyelle	bought
ϖ	UH	voyelle	book
u	UW	voyelle	boot
$\wedge$	AH	voyelle	but
$\ni$	ER	"schwa"	bird
∂	UR	"schwa"	neighbor
aϖ	AX	"schwa"	about
χ	IX	"schwa"	roses
α^y	AY	diphtongue	my
$\supset^y$	OY	diphtongue	boy
e^y	EY	diphtongue	bait
o^w	OW	diphtongue	boat
α^w	AW	diphtongue	down

Table 2 Les voyelles de l'anglais américain (diphtongues incluses). La notation Arpabet est une représentation qu'il est possible d'écrire sans fonte particulière.

Les voyelles de l'anglais américain sont légèrement différentes des voyelles de l'anglais britannique mais la plupart des relations entre voyelles sont conservées (voir Ladefoged [Lad82]).

Les voyelles sont voisées (les cordes vocales vibrent) sauf lorsqu'elles sont chuchotées. Ce sont généralement les phonèmes de plus grande énergie. Leur durée peut varier de 50 à 300 ms pour de la parole énoncée normalement. Elles se distinguent principalement grâce à la position fréquentielle des trois premiers formants qui correspondent à des résonances du conduit vocal. La table 3 indique les valeurs moyennes des formants pour un dialecte particulier de l'anglais et des locuteurs particuliers.

Formant	i	I	ϵ	æ	a	$\supset$	ϖ	u
F1	280	400	550	690	710	590	450	310
F2	2250	1920	1770	1660	1100	880	1030	870
F3	2890	2560	2490	2490	2540	2540	2380	2250

Table 3 Fréquence des trois premiers formants pour huit voyelles de l'anglais américain du "Mid-West" (d'après Ladefoged [Lad85])

A cause des différences de forme et de longueur du conduit vocal, les valeurs des formants varient beaucoup entre locuteurs. Il y a cependant une large intersection entre ces valeurs pour différents locuteurs. Celle-ci est telle qu'il est possible de distinguer des voyelles qui possèdent les mêmes premier et deuxième formants (F1 et F2) lorsqu'elles sont prononcées par des locuteurs différents [PB52]. Cela signifie que d'autres caractéristiques, comme la position des autres formants, la fréquence fondamentale, ou la largeur de bande des formants, rendent possible l'identification.

En tenant compte d'observations obtenues grâce à des tests perceptuels utilisant des voyelles syntétiques, Carlson et al. [CGF70] ont suggéré une représentation des voyelles à l'aide de deux pics fréquentiels (F1, F2'). De plus, Carlson et al. [CFG75] ainsi que Bladon [Bla83] ont proposé une formule empirique pour calculer F2' à partir des valeurs des quatre premiers formants. Chistovich et al. [CSL78] ont représenté les voyelles à l'aide des principaux pics du spectre de fréquence obtenus en deux étapes : extraction des pics et intégration auditive. Ils ont rapporté que, lorsque les pics des stimuli étaient séparés par moins de *3.5 Bark*, ces stimuli pouvaient être représentés par seulement un pic dont la position était donnée par le centre de gravité des deux pics originaux. Lorsque les pics étaient séparés par plus de 3.5 Bark, une représentation à l'aide d'un pic n'était plus possible. Cependant, différents points de vue existent sur la question. Lonchamp [Lon88] développa des arguments contre l'idée de l'intégration à large bande et de l'existence du concept F2'. De plus, dans le cadre de la parole continue, le problème est beaucoup plus complexe car il faut alors tenir compte de la variation inhérente à la localisation des pics.

En anglais un "*schwa*" est une variante d'une voyelle qui diffère simplement de celle-ci par sa qualité. Quant aux *diphtongues* (qui peuvent être regroupées en deux catégories : AY, OY, EY, et OW, AW), elles sont caractérisées par une transition d'un premier état stable vers un deuxième. Elles se distinguent par leurs états stables mais aussi par la pente et la direction des transitions. Cependant, pour de la parole continue, la présence des états stables est difficile à détecter. Notons, enfin, qu'une grande partie de l'accent d'un étranger parlant anglais est due à une inadéquation dans l'utilisation des diphtongues.

A.1.4 Les consonnes de l'anglais américain

La table 4 liste les consonnes de l'anglais américain regroupées par classes suivant la manière dont elles ont été articulées. Cette division fait apparaître six classes : *semi-consonnes*, *liquides*, *nasales*, *fricatives*[3], *plosives*, et *affriquées*. La division aurait pu être différente. En particulier, les fricatives auraient pu être divisées en fricatives faibles et fricatives fortes.

[3] Le phonème /h/ aurait aussi pu être classifié comme une voyelle chuchotée. Une variante voisée de ce phonème est souvent présente lorsqu'il se trouve entre deux voyelles.

Phonème	Arpabet	Classe	Voisé	Exemple
j	Y	semi-consonne	oui	you
w	W	semi-consonne	oui	wit
l	L	liquide	oui	let
r	R	liquide	oui	rent
m	M	nasale	oui	met
n	N	nasale	oui	net
ŋ	NX	nasale	oui	bang
h	HH	fricative	-	hat
f	F	fricative	non	fat
θ	TH	fricative	non	thin
s	S	fricative	non	sat
š	SH	fricative	non	shut
v	V	fricative	oui	vat
δ	DH	fricative	oui	that
z	Z	fricative	oui	zoo
ž	ZH	fricative	oui	azure
č	CH	affriquée	non	church
ǰ	JH	affriquée	oui	judge
p	P	plosive	non	pet
t	T	plosive	non	ten
k	K	plosive	non	kit
b	B	plosive	oui	bet
d	D	plosive	oui	den
g	G	plosive	oui	get

Table 4 Les consonnes de l'anglais américain.

Dans cette table le phonème WH (en notation Arpabet), qui est la contrepartie non voisée du phonème /w/ et qui se trouve par exemple dans certaines prononciations du mot "which", a été délibérément omis. De plus, certaines variantes phonétiques, appelées *allophones*, tels les "*flaps*" (caractérisés par la rencontre de deux articulateurs comme dans

le mot "batter") ou les *plosives glottales* (définies plus loin dans cette section), n'ont pas été explicitées pour des raisons de simplicité.

Les *semi-consonnes* et les *liquides* sont des consonnes très similaires aux voyelles car elles sont caractérisées par un signal temporel périodique et intense, ainsi que par une énergie importante dans les premiers formants. Souvent appelées *semi-voyelles*, elles sont plus faibles que des voyelles car elles sont produites à l'aide d'un conduit vocal moins ouvert. Le spectre de fréquence qui caractérise les semi-consonnes reste généralement stable pendant moins de temps que celui des voyelles.

En anglais, il y a trois différentes sortes de *nasales*. La frontière entre une nasale et une voyelle est typiquement marquée par un changement abrupt de l'intensité et de la forme du spectre de fréquence. Ceci est dû à la cavité nasale. Les changements spectraux intervenant au niveau de l'amplitude et de la fréquence des formants coïncident avec la fermeture et l'ouverture du conduit nasal.

Les *fricatives* sont généralement caractérisées par un spectre large bande. Le plus souvent, il y a plus d'énergie dans les hautes fréquences que dans les basses fréquences. Souvent produites à l'aide d'une source de bruit, elles peuvent être voisées ou non-voisées. En anglais, les fricatives voisées ont souvent une durée plus courte que les fricatives non voisées. Bien que lors de la production des fricatives voisées les cordes vocales normalement vibrent, ceci n'est pas toujours le cas.

Les *plosives* sont produites grâce à une occlusion complète du conduit vocal, suivie d'un relâchement de celui-ci. Acoustiquement, elles sont caractérisées par une période de silence suivie d'une barre d'explosion et d'un spectre de bruit correspondant au relâchement du conduit vocal. Les plosives non-voisées ont souvent un spectre de bruit plus long que celui des plosives voisées. Deux principaux types d'allophones, appelés *plosives glottales*, existent. Le premier type de plosive glottale se trouve souvent inséré au début d'un mot commençant par une voyelle, et le second correspond à une variante du phonème /t/, comme dans certaines prononciations du mot "cotton".

Une *affriquée* est une plosive suivie d'une fricative. Les deux affriquées indiquées dans la table 4 ont un statut spécial en phonologie anglaise. Ce sont les seules affriquées qui peuvent se trouver au début ou à la fin d'un mot. La durée de la partie fricative est typiquement moitié de celle de la fricative correspondante.

Pour plus de détails sur la description et les caractéristiques acoustiques et phonétiques des sons de l'anglais ainsi que sur la production de la parole, le lecteur pourra se reporter aux livres de Fant, Lehiste, et O'shaughnessy [Fan73, Leh69, O'S87].

A.1.5 Spectrogrammes

Le spectrogramme est une représentation acoustique très utilisée du signal de parole. Il permet de visualiser le signal de parole (amplitude/temps) comme une forme tridimensionnelle (amplitude/fréquence/temps). La fréquence est représentée comme une fonction du temps alors que l'amplitude est donnée par le niveau de gris de chaque point. La figure 1 montre un spectrogramme de la phrase "we owe you a yoyo" prononcée par un locuteur masculin.

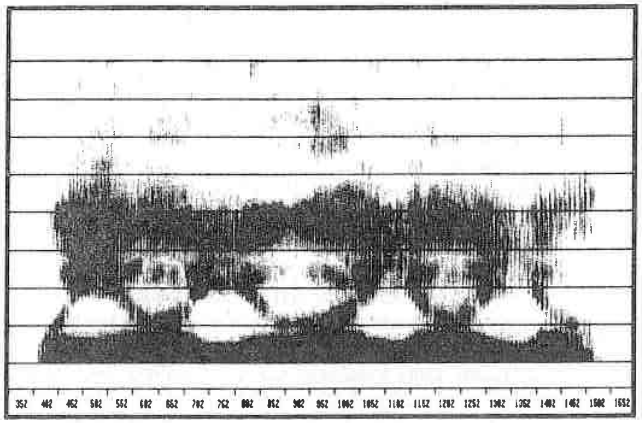


Figure 1 Spectrogramme de la phrase "we owe you a yoyo" prononcée par un locuteur masculin.

Les pics du spectre de fréquence (formants) apparaissent comme des bandes horizontales sombres. Les sons voisés sont représentés par des traits verticaux. Ceci est dû à l'augmentation d'amplitude du signal chaque fois que les cordes vocales se ferment. Le spectrogramme tient compte uniquement de l'information amplitude, laissant en particulier de côté l'information phase supposée peu importante dans beaucoup d'applications. Les spectrogrammes sont souvent utilisés pour examiner la position et le mouvement des formants, mais aussi pour fournir des informations sur les caratéristiques acoustiques et phonétiques des sons (durée, période, concentration d'énergie, transitions, etc).

Avec le progrès des technologies liées à l'informatique, les spectrogrammes numériques remplacent progressivement les spectrogrammes analogiques. Ils permettent, notamment, d'obtenir une meilleure dynamique (40–50 dB). De plus, grâce à la souplesse obtenue avec les ordinateurs, il est ainsi possible "d'éditer" le spectrogramme

et, ainsi, d'appliquer des fonctions permettant, par exemple, de visualiser le spectre d'une "frame" (pouvant être définie comme un vecteur de représentation de la parole), de calculer l'énergie dans certaines bandes de fréquence, etc. Les spectrogrammes numériques sont obtenus par le calcul d'une transformée de Fourier discrète (*DFT*) sur le signal échantillonné, multiplié au préalable par une fenêtre de taille fixe *w(n)* (généralement une fenêtre de Hamming). A mesure que la fenêtre se déplace dans le signal, de nouvelles "tranches" du spectrogramme sont calculées. En modifiant la taille de la fenêtre *w(n)*, il est possible d'obtenir des spectrogrammes à large bande (e.g. figure 1) ou des spectrogrammes à bande étroite. Les spectrogrammes à large bande ont une meilleure résolution temporelle que les spectrogrammes à bande étroite. Grâce à cette propriété, ils sont aussi beaucoup plus utilisés. Kuhn [Kuh84] proposa de représenter l'information contenue dans les spectrogrammes conventionnels (à niveaux de gris) à l'aide de la couleur, mais une telle représentation n'est pas encore très utilisée.

La question qui se pose est de savoir si le spectrogramme est une représentation permettant de retenir suffisamment d'informations pertinentes du signal de parole. Beaucoup de chercheurs, travaillant sur les modèles d'oreille, ont déjà proposé d'autres représentations (e.g. [Sen86, HL86]). Le spectrogramme conventionnel effectue une analyse avec une bande passante constante (e..g. 300 Hz pour un spectrogramme à large bande) sur une échelle de fréquence linéaire. Ce n'est certainement pas le type d'analyse suggérée par les recherches menées en psychoacoustique [ZT79]. D'après les résultats obtenus grâce à des tests perceptuels, une échelle Bark (ou Mel) semble mieux adaptée. De façon plus générale, une représentation intégrant des propriétés du système auditif humain semble souhaitable. Dans leur étude sur l'identification du locuteur à l'aide de sptectogrammes, Bolt et al. [Bol70] critiquèrent l'utilisation des spectrogrammes conventionnels. Enfin, Lonchamp, dans sa thèse d'Etat, regroupe les critiques classiques associées aux spectrogrammes et discute le rôle des formants dans la perception de la parole [Lon88]. La question reste posée. Pour des raisons historiques, et étant donné l'absence d'une meilleure solution et les connaissances déjà accumulées sur cette représentation acoustique, le spectrogramme est toujours très utilisé.

A.1.6 Quelques indices caractérisant les sons

L'acoustique et la phonétique considèrent le signal de parole comme la sortie du mécanisme de production et relient celui-ci au message linguistique. Les relations entre les phonèmes et leur représentation acoustique sont à la base de beaucoup d'applications (codage, synthèse et reconnaissance de la parole). Etant donné un signal de parole numérisé, plusieurs types de paramètres ou indices peuvent être obtenus. Certains paramètres, comme le taux de passages par zéro ou la fréquence fondamentale (F0), peuvent être calculés directement à partir du signal numérisé. Cependant, l'expérience a montré que les indices fréquenciels ont souvent amené une meilleure compréhension des relations entre les propriétés articulatoires et les propriétés acoustiques des sons de la parole (e.g. formants et résonances du conduit vocal). Un certain nombre d'indices peuvent être utilisés pour caractériser les sons. Parmi les plus populaires, citons :

1. le taux de passages par zéro,
2. la fréquence fondamentale ou taux de voisement,
3. les indices se rapportant à l'énergie (énergie dans différentes bandes de fréquence, etc),
4. les indices visant à fournir une approximation grossière du spectre fréquenciel (analyse par prédiction linéaire d'ordre réduit [Mak73, JW88]),
5. les formants ainsi que leur mouvement,
6. la durée,
7. barre d'explosion et bruit de friction associé,
8. barre de voisement,
9. seuil fricatif.

Pour une définition plus précise de ces paramètres, le lecteur pourra se référer à l'article de Schwartz et Zue [SZ76].

Les voyelles peuvent être détectées grâce à la présence d'une énergie importante en basse et moyenne fréquence. Elles ont un spectre fréquentiel relativement stable et sont caractérisées principalement par les valeurs des formants. Afin d'identifier les voyelles, il est important de prendre en compte les phénomènes de coarticulation (influence des phonèmes voisins) [Kam75].

Au niveau des consonnes, les indices acoustiques permettant de les distinguer varient au niveau des différentes consonnes et au niveau des transitions formantiques avec les voyelles adjacentes. Ainsi, pour l'identification, les deux types d'information doivent être pris en considération. Les indices utilisés dépendent du problème traité (reconnaissance, segmentation, étiquetage, etc). La table 4 résume[4] certaines caractéristiques des consonnes, telles qu'elles peuvent être observées sur un spectrogramme.

[4] Comme les affriquées résultent de la combinaison d'une plosive et d'une fricative, elles n'apparaissent pas dans cette table.

Plosives non-voisées	Barre d'explosion suivie d'un bruit de friction après une période de silence; les formants peuvent être visible pendant le bruit de friction.
Plosives voisées	Energie pas toujours visible en basse fréquence pendant la fermeture du conduit vocal; au relâchement du conduit vocal, apparition rapide de la structure formantique.
Fricatives non-voisées	Spectre de bruit haute fréquence.
Fricatives voisées	Structure formantique souvent réduite en intensité avec abaissement du premier formant; spectre de bruit haute fréquence.
Nasales	Fréquence faible du premier formant (environ 250 Hz) souvent intense comparé aux autres formants; fréquente apparition d'une discontinuité dans la structure acoustique entre la nasale et la voyelle adjacente; changement brusque de l'intensité et du spectre de fréquence.
Semi-consonnes	Structure formantique ayant un mouvement similaire à celui de la voyelle correspondante; plus faible au niveau intensité.
l	Structure formantique similaire à celle d'une voyelle avec une absence (ou une importante réduction) des plus hauts formants; énergie souvent présente aux environ de 1200 Hz.
r	Structure formantique changeante avec abaissement du troisième formant (ainsi que des plus hauts formants).
h	Voisé ou non-voisé; pas de structure formantique claire; entre deux voyelles s'adapte aux structures formantiques des voyelles adjacentes; souvent peu d'énergie dans la région du premier formant.

Table 5 Caractéristiques des consonnes telles qu'elles peuvent être observées sur un spectrogramme (d'après Ladefoged [Lad85])

Comme le mentionne Ladefoged [Lad85], l'interprétation des spectrogrammes n'est pas évidente. Aussi, les caractéristiques se trouvant dans la table ci-dessus doivent

être considérées comme des indications, plus que comme des données se trouvant invariablement dans tout spectrogramme des sons présentés.

Dans ce chapitre certaines caractéristiques des sons de l'anglais américain (et du français) ont été présentées. Nous avons vu que ces sons peuvent être décrits à l'aide d'une représentation acoustique appelée spectrogramme. Toutefois, si l'on observe le signal de parole la caractéristique la plus évidente est sa variabilité d'où le caractère instable des réalisations phonétiques. Le même phonème peut, en effet, être soumis à des influences très fortes et très diverses conduisant à de nombreuses variantes. C'est ce qui fait la difficulté de la reconnaissance automatique de la parole.

Chapitre A.2 L'OREILLE ET LES MECANISMES D'AUDITION

Nous sommes protégés par des filtres qui atténuent tous les signaux qui viennent de l'extérieur.

Ramón Sender

A.2.1 Le système auditif

A.2.1.1 Structure de l'oreille

Comme le montre la figure 2, l'oreille est formée de trois parties : l'oreille *externe*, l'oreille *moyenne*, et l'oreille *interne*.

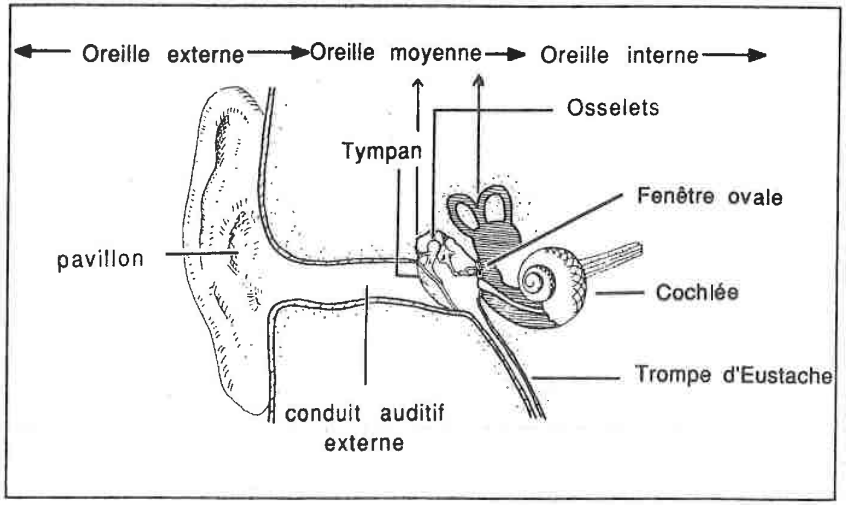


Figure 2 Coupe de l'oreille humaine (d'après Zwicker [ZF81]).

L'oreille externe achemine les variations de pression dues à la parole vers le *tympan* où l'oreille moyenne, par l'intermédiaire des osselets, transforme ces variations de pression en vibrations mécaniques. Ces vibrations mécaniques, transmises à l'oreille interne, sont ensuite converties en potentiel électrique dans les neurones auditifs qui mènent au cerveau.

A.2.1.2 L'oreille externe

La partie visible de l'oreille externe permet à l'oreille d'être plus sensible aux sons venant de face. Elle a aussi un rôle de protection. Les sons passent ensuite dans une tube uniforme pour atteindre la membrane tympanique. Comme tous les tubes, il possède des fréquences de résonnance dont la première se situe vers 3 kHz. Cette résonnance amplifie les hautes fréquences et aide probablement la perception de sons comme les fricatives.

A.2.1.3 L'oreille moyenne

La principale fonction de l'oreille moyenne est d'assurer le transfert des vibrations d'un milieu aérien vers un milieu aqueux qui est celui de la *cochlée* . L'oreille moyenne fait aussi office d'adapteur d'impédance afin d'améliorer la transmission des sons et protège l'oreille interne, plus délicate, contre les sons intenses. D'un point de vue fréquentiel, elle agit comme un filtre passe-bas ayant une atténuation d'environ 15 dB/octave après 1 kHz.

A.2.1.4 L'oreille interne

L'oreille interne contient la cochlée qui est probablement la partie la plus importante de l'oreille au point de vue perception de la parole. Comprendre le fonctionnement de la cochlée est fondamental pour modéliser les mécanismes d'audition. La cochlée transforme les vibrations mécaniques en excitations électriques en sortie des fibres nerveuses. Une coupe de la cochlée est présentée à la figure 3.

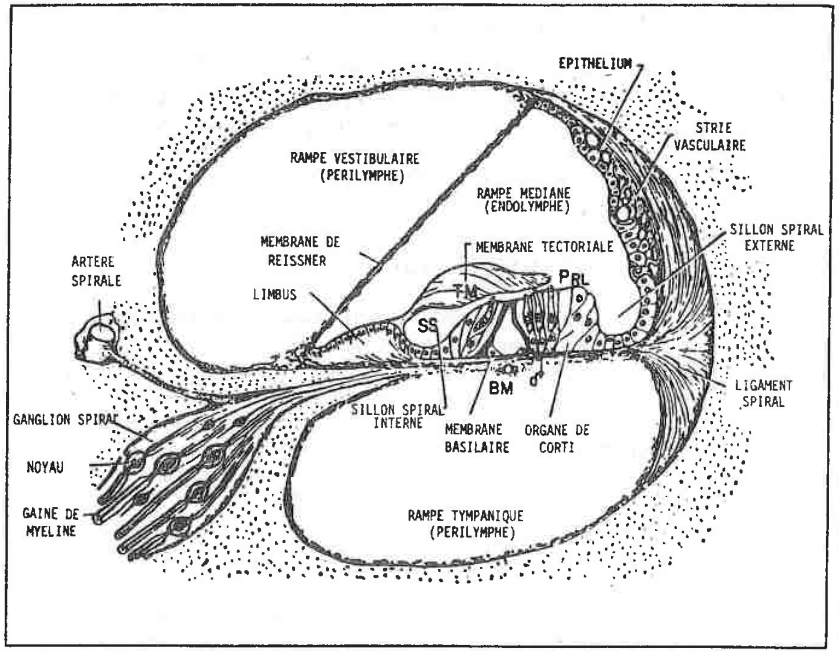


Figure 3 Coupe de la cochlée (d'après Dolmazon [Dol80]).

Les ondes sonores arrivant au niveau de l'oreille se propagent dans le canal auditif et provoquent des vibrations au niveau du tympan. Ces vibrations sont transmises à la *fenêtre ovale* qui relie l'oreille moyenne et l'oreille interne. Le liquide contenu dans la cochlée est alors mis en mouvement par les vibrations de la fenêtre ovale qui provoque à son tour des vibrations de la *membrane basilaire*. Celles-ci sont prises en compte par les *cellules ciliées*, qui reposent sur la membrane basilaire, afin de générer au niveau des fibres nerveuses des impulsions électriques qui acheminent l'information acoustique au cerveau. Zwicker and Feldtkeller [ZF81] montrèrent que la résolution fréquentielle est directement liée à la résolution spatiale des fibres nerveuses situées sur la membrane basilaire. Cette organisation fréquentielle est connue sous le nom de *tonotopie*. Chaque neurone ou fibre nerveuse du nerf auditif répond de façon optimale à seulement une bande de fréquence étroite.

Chaque fibre nerveuse possède une *courbe d'accord* appelée aussi courbe d'isocadence. Cette courbe représente l'intensité d'un son pur, qui évoque une cadence moyenne de décharge (activité au niveau des fibres nerveuses), donnée en fonction de la fréquence. La figure 4 présente les courbes d'accord de fibres nerveuses du chat pour huit bandes de fréquence.

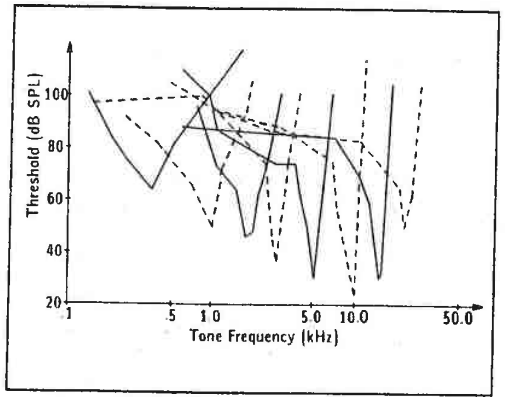


Figure 4 Courbes d'accord de fibres nerveuses du chat pour huit bandes de fréquence (d'après Evans [Eva75]).

Dans la cochlée, il y a deux types de cellules ciliées : les *cellules internes et les cellules externes*. De précédentes études ont montré [Dal72] que les cellules externes étaient principalement des détecteurs de déplacement alors que les cellules internes étaient avant tout des détecteurs de vitesse, et que le stimulus lié au déplacement était plus important que celui lié à la vitesse. Cette propriété a été exploitée au cours des études réalisées (voir section E.2.7).

A.2.2 Perception des sons

A.2.2.1 Bandes critiques

L'oreille humaine possède la remarquable faculté d'intégrer certaines zones de fréquences en bandes appelées *bandes critiques*. D'un point de vue physiologique, un filtre à bande critique peut être considéré comme un filtre passe-bande dont la réponse en fréquence correspond grossièrement à une courbe d'accord d'une fibre nerveuse. D'un point de vue psychoacoustique, une bande critique définit une bande de fréquence pour laquelle la perception d'un stimulus à bande étroite change soudainement lorsque la modification en fréquence du stimulus l'entraîne hors de la bande critique. Lorsque deux sons simultanés sont envoyés à l'entrée d'un filtre à bande critique, le son qui a la plus grande énergie dans la bande critique est perceptuellement dominant et *masque* l'autre son.

Une mesure perceptuelle, appelée *Bark*, sert de lien entre la fréquence acoustique (Hertz) et la résolution fréquentielle de l'oreille, pour laquelle un Bark couvre une largeur de bande critique. Une expression analytique de cette correspondance entre une fréquence f et un taux de bande critique z est donnée par [ZT80] :

$$z = 13\,arctan\left(0.76\frac{f}{1000}\right) + 3.5\,arctan\left(\frac{f}{7500}\right)^2. \quad (1)$$

Pour une mesure similaire, appelée *Mel*, la correspondance est linéaire en fréquence jusqu'à 1 kHz et logarithmique pour les fréquences supérieures à 1 kHz. L'expression analytique définissant cette mesure est donnée par :

$$y = 2595\,log_{10}\left(1 + \frac{f}{700}\right). \quad (2)$$

Cette échelle Mel a été utilisée dans beaucoup d'applications. Citons, par exemple, le codeur à prédiction linéaire de Makhoul et Cosell [MC76], le vocodeur à composante principale de Zahorian et Rothenberg [ZR81] et de façon plus générale la reconnaissance automatique de la parole (e.g. [DM80]).

A.2.2.2 Sonie

La *sonie* d'un son pur dépend de son intensité et de sa fréquence. Pour un son quelconque, elle est souvent définie comme étant l'intensité d'un son pur à 1 kHz qui serait perçu avec la même force sonore que le son en question. La sonie est mesurée en *phones*, qui sont similaires aux dB pour un son pur proche de 1 kHz. La force sonore des sons purs a été étudiée par de nombreux auteurs (e.g. Fletcher et Munson [FM33]). En comparant des sons purs à différentes fréquences et amplitudes, des *courbes d'isosonie* ont été déterminées. Ces courbes sont présentées à la figure 5 où, en abcisse on trouve la fréquence, en ordonnée le niveau du son pur et en paramètre le niveau d'isosonie qui est constant le long de chaque courbe.

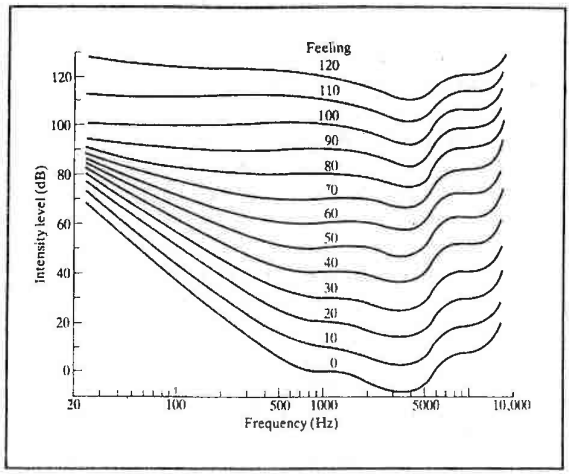


Figure 5 Courbes d'isosonie. Le nombre au dessus de chaque courbe indique le niveau d'isosonie exprimé en phones (d'après Fletcher et Munson [FM33]).

Comme notre perception de la sonie ne correspond pas directement à des mesures d'intensité, il était nécessaire de trouver des méthodes permettant de dériver la sonie à partir de l'intensité. Si un sujet entend deux sons purs à une fréquence donnée, il est capable de comparer les deux sons purs en terme de force sonore. Les résultats des tests effectués ont montré que la sonie (mesure perceptuelle) était reliée à l'intensité (mesure physique) par une loi de puissance, appelée loi de Stevens [Ste57], donnée par :

$$J = kI^{0.3}. \quad (3)$$

Dans cette formule, I représente l'intensité et J la sonie. Une autre mesure de la sonie est le *sone* qui est tel que, un doublement de la sonie équivaut à une augmentation d'intensité de 10 dB.

A.2.2.3 Saturation, adaptation, masquage, suppression, et inhibition latérale

Le système auditif est souvent représenté par une série de mécanismes (ou d'étapes) *linéaires* et *non-linéaires*. Dans cette section, quelques-uns des mécanismes non-linéaires, souvent introduits dans les modèles auditifs réalisés par ordinateur, sont définis brièvement. La connaissance de ces mécanismes résulte d'études menées en psychoacoustique et physiologie.

L'intensité et la durée d'un stimulus sonore influencent les réponses des fibres nerveuses du système auditif. Lorsque la durée du stimulus est fixée, le taux de décharge

augmente régulièrement avec le niveau de pression acoustique (ou *SPL* en abréviation de "sound pressure level") jusqu'à un seuil de saturation [Kia68, RHAB71]. Ceci est appelé : phénomène de *saturation.*

La sonie d'un son dépend de son intensité mais aussi de la présence éventuelle d'autres sons au même moment. Les sons se masquent entre eux. Pour évaluer les effects du masquage (d'un point de vue psychoacoustique), il faut mesurer l'augmentation d'intensité qu'il est nécessaire d'appliquer à un son pour être entendu en présence d'un son qui le masque. Cette procédure est similaire à celle utilisée pour déterminer les courbes d'isosonie (voir figure 5). Le concept de bande critique, énoncé précédemment, a été déduit de tests utilisant l'effet de masque. Ce phénomène de masquage est connue sous le nom de *masquage simultané.*

Une autre vue de l'effet de masque, populaire parmi les neurophysiologistes, tient compte du fait que le son masquant appauvrit l'activité qu'un autre son aurait s'il avait été présenté tout seul. Javel et al. [Jav83] ont observé que la réponse des fibres nerveuses à un son pur émis à la fréquence caractéristique (CF) de ces fibres pouvait être appauvrie par un autre son pur, même si celui-ci ne produisait aucune excitation des fibres nerveuses en question. Ceci est appelé : *suppression à deux tons.* Ce phénomène donne plus d'importance aux composantes intenses du stimulus, au niveau des réponses des fibres nerveuses.

Lorsque le niveau de pression acoustique est fixe, le taux de décharge décroît régulièrement lorsque la durée du stimulus augmente. En fait le taux de décharge s'adapte et approche une asymptote correspondant à un taux de décharge stable [KWTC65]. C'est le phénomène d'*adaptation* appelé aussi *adaptation à court terme.* L'adaptation joue probablement un rôle au niveau de la perception des changements rapides en fréquence et en amplitude ainsi que dans la prise en compte du contexte. L'*adaptation à long terme* doit être distinguée de l'adaptation à court terme. Dans ce cas, les intervalles de temps considérés sont de l'ordre de quelques minutes.

Une propriété importante, probablement liée à l'adaptation [Del84], est le *masquage postérieur.* Son effet est de diminuer les réponses à un son donné à cause d'un son précédent, généralement plus intense [HD79]. Comme l'adaptation à long terme, le masquage postérieur se distingue de l'adaptation à court terme par les intervalles de temps considérés. Le masquage antérieur [Ell62], qui caractérise le masquage d'un son par un son qui suit, a aussi été avancé mais c'est un mécanisme beaucoup plus difficile à expliquer.

L'*inhibition latérale* est généralement décrite comme étant la suppression de l'activité de fibres nerveuses localisées sur la membrane basilaire provoquée par l'activité de fibres adjacentes le long de la membrane. Il a été montré que ce phénomène jouait un grand rôle dans la perception de la parole [Hou72]. Il contribue probablement aussi à la grande sélectivité en fréquence du système auditif.

Toutes ces transformations sont des propiétés essentielles du système auditif. La compréhension de ces mécanismes est fondamentale pour réaliser des modèles du système auditif à l'aide d'ordinateurs.

Adieu, dit le renard. Voici mon secret.
Il est très simple : on ne voit bien
qu'avec le coeur. L'essentiel est
invisible pour les yeux.

Extrait de "Le petit Prince" Saint Exupéry

PARTIE B
TECHNIQUES D'ANALYSE DE LA PAROLE

Après quelques rappels sur l'analyse par prédiction linéaire, l'utilisation de cette technique en modélisation spectrale est présentée. Ensuite, les principaux travaux effectués récemment qui utilisent des mécanismes du système auditif pour estimer le spectre de fréquence sont revus. Enfin, quelques applications de ces modèles en reconnaissance de la parole sont mentionnées.

Chapitre B.1 L'ANALYSE PAR PREDICTION LINEAIRE

Que me reste-t-il de ma vie? Que me reste-t-il.
Que cela est étrange, il ne me reste que ce que
j'ai donné aux autres.

Vahan Tekeyan

B.1.1 Introduction

Dans beaucoup d'applications liées au traitement de la parole et notamment le codage, la technique d'analyse par *prédiction linéaire* [IS68, AH71, Mak73] s'est avérée très utile. Cette technique modélise les mécanismes de production de la parole. Elle est fondée sur l'idée qu'un signal qui véhicule un message n'est jamais complètement aléatoire. Il y a une *corrélation* entre les échantillons successifs. Le codage par prédiction linéaire (en anglais "linear predictive coding" ou *LPC*) utilise cette corrélation pour réduire les données manipulées tout en préservant l'information contenue dans le signal. C'est actuellement la technique la plus utilisée pour le codage de la parole à faible débit. La popularité de la technique *LPC* est due au fait qu'elle fournit une représentation compacte et précise de l'enveloppe spectrale tout en étant relativement simple à calculer.

B.1.2 Le modèle LPC

La technique *LPC* fournit un système d'analyse-synthèse du signal de parole [MG76, Mak75a]. Le modèle de synthèse utilise une source excitatrice, $U(z)$, en entrée d'un filtre, $H(z)$, qui modèle une enveloppe spectrale et fournit en sortie le signal synthétisé $\hat{S}(z)$. Etant donné un échantillon de parole, $s(n)$, supposé stationnaire pendant une fenêtre de N échantillons, l'échantillon de synthèse $\hat{s}(n)$ peut être modélisé par une combinaison linéaire des p précédentes sorties et $q+1$ précédentes entrées du synthétiseur *LPC* :

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \hat{s}(n-k) + G \sum_{l=0}^{q} b_l u(n-l), \tag{4}$$

où G est un facteur de gain du signal de parole présenté en entrée. Le filtre $H(z)$ peut alors être défini par :

$$H(z) = \frac{\hat{S}(z)}{U(z)} = G \times \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{5}$$

Ceci conduit au schéma suivant pour le modèle de synthèse,

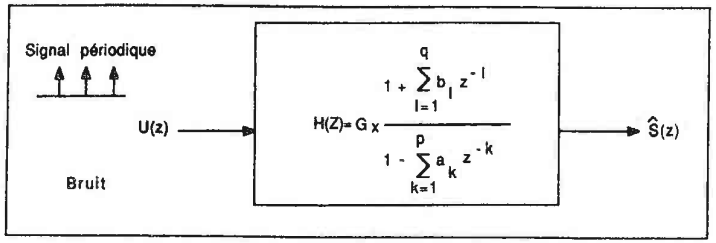


Figure 6 Le modèle de synthèse.

La plupart des travaux utilisant la technique *LPC* assument un modèle tout pôle, connu aussi sous le nom : *modèle autoregressif (AR)*, et caractérisé par $q=0$. Le modèle caractérisé par $p=0$ est appelé *modèle à moyenne flottante* (en anglais "moving average" ou *MA*), car la sortie du modèle est obtenue par une moyenne pondérée des q précédentes entrées. Le modèle le plus général, mais moins utilisé que le modèle *AR* à cause de sa complexité, est le modèle utilisant à la fois les pôles et les zéros du filtre $H(z)$. Ce modèle est connu sous le nom de *modèle autoregressif à moyenne flottante* (en anglais "autoregressive moving average" ou *ARMA*). Dans le travail réalisé, seul le modèle *AR* a été utilisé. Lorsque l' échantillon de parole $s(n)$ est filtré par l'inverse de $H(z)$, appelé aussi filtre de prédiction,

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}, \tag{6}$$

alors l'erreur de prédiction ou signal résiduel, $e(n)$, est définie par :

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k). \tag{7}$$

Si E_m est l'énergie du signal résiduel alors,

$$E_m = \sum_{n=-\infty}^{\infty} e_m^2(n) = \sum_{n=-\infty}^{\infty} [s_m(n) - \hat{s}_m(n)]^2. \tag{8}$$

Les valeurs a_k qui minimisent E_m sont obtenues lorsque $\frac{\partial E_m}{\partial a_k} = 0$ pour $k=1,2,3,...,p$. Ceci donne p équations linéaires,

$$\sum_{n=-\infty}^{\infty} s_m(n-i)\, s_m(n) = \sum_{n=-\infty}^{\infty} s_m(n-i)\, s_m(n-k), \quad for\ i = 1,2,3,...,p. \tag{9}$$

Si $R_m(i,k) = \sum_{n=-\infty}^{\infty} s_m(n-i)\, s_m(n-k)$, on obtient :

$$\sum_{k=1}^{p} a_k R_m(i,k) = R_m(i,0) \quad for\ i = 1,2,...,p. \tag{10}$$

Pour calculer les coefficients de prédiction a_k, deux méthodes sont couramment utilisées a) la méthode d'*autocorrélation*, introduite tout d'abord par Itakura [IS68] et b) la méthode de *covariance* par Atal [AS68, AH71]. La différence essentielle entre les deux méthodes est la nécessité d'utiliser une fenêtre d'analyse pour la méthode d'autocorrélation. Notons aussi que pour la méthode d'autocorrélation la stabilité du filtre *LPC* est garantie, ce qui n'est pas le cas pour la méthode de covariance. Au niveau calcul, la complexité des deux méthodes n'est pas très importante (la méthode d'autocorrélation nécessite, si N est la taille de la fenêtre, $N \times p$ (corrélation) + p^2 (solution de la matrice) multiplications). De plus, le calcul est effectué dans le domaine temporel.

B.1.3 Utilisation de la technique LPC en modélisation spectrale

B.1.3.1 Ordre du modèle LPC

Le choix de l'ordre du modèle *LPC* est un compromis entre la précision avec laquelle on veut estimer l'enveloppe spectrale, les ressources qui sont allouées au calcul (temps et mémoire), et l'application visée. En général, en reconnaissance de la parole, un ordre du modèle compris entre huit et quatorze est utilisé afin de modéliser les premiers pics (trois à cinq) du spectre de fréquence. Cependant, il est quelquefois souhaitable de faire une modélisation grossière du spectre de fréquence d'un son particulier. Ceci peut être utile en segmentation automatique, classification, ou paramétrisation de la parole avant extraction d'indices. Un modèle à deux pôles, proposé par Makhoul [Mak73], a été utilisé par Schwartz et Makhoul en segmentation automatique et classification [SM75] et par Sambur et Rabiner [SR75] en reconnaissance multilocuteur des dix chiffres.

Inversement, dans le cas de la parole bruitée, de précédentes études [Tie80] ont montré que, afin de modéliser à la fois la parole et le bruit, l'ordre du modèle *LPC* doit être augmenté par rapport au cas de la parole non-bruitée.

B.1.3.2 Analyse LPC sélective

L'analyse *LPC* standard traite toute la gamme de fréquence de façon comparable même si il est maintenant bien accepté que l'oreille résout mieux les basses fréquences, qui sont plus importantes que les hautes fréquences pour l'intelligibilité de la parole. L'analyse par *prédiction linéaire sélective* [Mak73] a été utilisée dans différents domaines du traitement de la parole, comme la reconnaissance [SM75] ou la compression de la parole [Mak74]. Elle consiste à appliquer l'analyse *LPC* sur une région préalablement sélectionnée du spectre de fréquence plutôt que sur tout le spectre. Ceci permet de

modéliser seulement les régions du spectre qui sont importantes pour l'application visée. Un autre but peut être de modéliser des bandes de fréquences avec des modèles d'ordre différent afin, par exemple, de diminuer le nombre de pôles à interpréter.

B.1.3.3 Paramètres types utilisés en reconnaissance

L'analyse *LPC* fournit p (ordre du modèle) coefficients réels a_k (k=*1,2,3,...,p et* a_0=*1*) représentant une estimation optimale à l'aide de p pôles du spectre de fréquence d'une fenêtre du signal de parole. Beaucoup de représentations paramétriques, dérivées de la représentation formée des coefficients de prédiction linéaire, ont été utilisées. Les représentations choisies, quelques-unes plus utiles que d'autres ou plus facilement interprétables physiquement, ont variées suivant les applications.

Les coefficients d'*autocorrélation*, introduits dans la section B.1.2, sont définis à partir des coefficients de prédiction a_k par la formule suivante :

$$R_a(i) = \sum_{k=0}^{p-i} a_k a_{k+i} \quad 1 \le i \le p. \tag{11}$$

Les coefficients de réflection ou coefficients *PARCOR*, k_i, peuvent être calculés à partir des coefficients de prédiction grâce à la formule récursive suivante :

$$\begin{aligned} k_i &= a_i^{(i)}, \\ a_j^{(i-1)} &= \frac{a_j^{(i)} - a_j^{(i)} a_{i-j}^{(i)}}{1 - k_i^2}, \quad 1 \le j \le i-1, \end{aligned} \tag{12}$$

où i prend les valeurs suivantes : p, p-*1*,..., *1* et initialement $a_j^{(p)} = a_j, \quad 1 \le j \le p$.

A partir des coefficients *PARCOR* , les coefficients *LAR* (en anglais "log-area-ratio") sont définis par :

$$g_i = log\left[\frac{1-k_i}{1+k_i}\right], \quad 1 \le i \le p. \tag{13}$$

Les coefficients *PARCOR* et les coefficients *LAR* ont été utilisés pour faire de la quantification. Il a été montré que les coefficients *LAR* étaient particulièrement appropriés à ce type d'application [Mak75a, MG76].

Une autre alternative aux coefficients *LPC* est disponible par l'intermédiaire des coefficients cepstraux. Ceux-ci peuvent être exprimés à l'aide du filtre de prédiction $A(z)$ par le formule suivante :

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} logA\left(e^{j\omega}\right) e^{jn\omega} dw. \tag{14}$$

Ces coefficients peuvent aussi être calculés itérativement à partir des coefficients de prédiction a_k par [MG76] :

$$-kc_k - ka_k = \sum_{n=1}^{k-1} (k-n)\, c_{k-n} a_n \quad for\ k > 0. \tag{15}$$

Enfin, citons les coefficients *LSP* (en abréviation de "line spectrum pair"), eux aussi dérivés de la représentation *LPC*. Ces coefficients ont été utilisés avec succès dans des applications de codage de la parole à bas débit [Wak81], mais aussi récemment en reconnaissance [Pal88]. Le calcul des coefficients *LSP* implique la recherche des p zéros de $A(z)$ sur le cercle unité par l'intermédiaire de deux transformations en z, $P(z)$ et $Q(z)$,

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A\left(z^{-1}\right), \\ Q(z) &= A(z) - z^{-(p+1)} A\left(z^{-1}\right). \end{aligned} \tag{16}$$

Les coefficients *LSP* correspondent aux fréquences des zéros qui se trouvent sur le cercle unité.

En reconnaissance automatique de la parole, plusieurs types de coefficients ont été utilisés. Des études récentes et plus anciennes [Ata74, Jun87] ont montré que les coefficients cepstraux et plus généralement les coefficients cepstraux pondérés (voir section C.3.2.4) permettaient d'obtenir les meilleurs scores de reconnaissance.

Chapitre B.2 LES MODELES AUDITIFS

Comme on ne connaît les hommes que par les paroles, ils faut les croire jusqu'à ce que les actions les détruisent.

Marquise de Sévigné

B.2.1 Introduction

Comme le mentionne Karjalainen dans [Kar87] "*Computational modeling of the auditory periphery has become an integral part of hearing and speech research*". Le système d'audition de l'être humain est le meilleur système pour reconnaître la parole. Aussi, un grand nombre d'études ont été entreprises pour essayer de le simuler à l'aide de modèles fonctionnels. Pour l'instant, notre compréhension des phénomènes qui se passent au niveau de l'oreille est très réduite. Cependant, le développement des ordinateurs et des logiciels de simulation facilite le test des modèles auditifs et les rend attractifs. Le concept de modèle auditif est généralement utilisé pour désigner une modélisation par ordinateur du système auditif périphérique humain. L'idée visant à réaliser de tels modèles n'est pas nouvelle. Depuis de nombreuses années, des représentations spectrales utilisant des bancs de filtres critiques ont été utilisées. Grâce à une connaissance croissante sur les réponses des fibres nerveuses au signal de parole, les modèles auditifs utilisant des concepts physiologiques ont récemment reçu beaucoup d'intérêt.

Les fonctions physiologiques de la membrane basilaire (quelques-unes ont été mentionnées précédemment) ainsi que celles liées à la cochlée sont considérées être les premières fonctions à simuler par les modèles auditifs. Le système auditif périphérique est capable d'intégrer des informations qui correspondent à des évènements acoustiques complexes (informations spectrales et temporelles). Il existe, au niveau des fibres nerveuses, des interactions entre les indices acoustiques présents dans le signal de parole [Del82]. Ces possibilités d'intégration, qui sont aussi présentes à des niveaux plus élevés du système auditif périphérique [RCM77], semblent être fonction de la tâche effectuée et du contexte acoustique. D'après Massaro [Mas87], les différentes sources d'informations sont évaluées chacune indépendamment, et l'intégration est faite de façon à ce que la source d'informations la moins ambigue ait le plus d'impact au niveau perceptuel. Au niveau analyse de la parole, cette intégration d'informations temporelles et spectrales par les modèles auditifs comporte un certain nombre d'avantages par rapport aux méthodes traditionnelles. En particulier, les modèles auditifs permettent :

1. une meilleure localisation temporelle des indices importants,
2. une meilleure détection des indices acoustiques en environnement dégradé,

3. éventuellement une réduction de la variabilité des informations intégrées, appelées aussi indices auditifs.

Les modèles auditifs comporte en général une première étape de traitement qui inclut un banc de filtres. C'est une différence importante avec les modèles dérivés de la technique *LPC*. En effet, les modèles dérivés de la technique *LPC* utilisent une largeur de bande d'analyse constante, alors que les bancs de filtres utilisent une largeur de bande d'analyse proportionnelle à la fréquence (lorsque le banc de filtres n'est pas simulé à l'aide d'une *FFT*). Ainsi, en utilisant un banc de filtres, il est possible d'obtenir une bonne résolution spectrale aux basses fréquences et une bonne résolution temporelle aux hautes fréquences (comme l'oreille humaine). Une bonne résolution spectrale aux basses fréquences permet de modéliser facilement les premiers formants, alors qu'une bonne résolution temporelle aux hautes fréquences permet de capturer les évènements rapides comme les barres d'explosion.

Comparés aux techniques traditionnelles, les modèles auditifs améliorent la résolution temporelle et permettent ainsi de prendre en compte les transitions rapides entre phonèmes. Cependant, afin de traiter correctement les informations fournies par les modèles auditifs, les systèmes de reconnaissance de parole doivent être adaptés [Del86]. Les performances humaines dépendent du système auditif périphérique mais, plus généralement, de tout le système auditif. Malheureusement, notre connaissance du système d'audition dans son ensemble est très réduite. Aussi, les modèles auditifs ont, dans le passé, surtout été utilisés pour l'extraction d'indices acoustiques plutôt que dans des systèmes de reconnaisance de parole (e.g. [Del84, Chi82, Cae85]).

Le domaine de la psychoacoustique fournit d'autres alternatives pour explorer le système auditif. Des tests perceptuels sont effectués sur des humains, et les résultats fournissent des renseignements sur le fontionnement du système auditif complet. Des concepts psychoacoustiques comme l'intonation, la sonie ou les bandes critiques ont été très utilisés dans les systèmes de traitement de la parole. Dans la prochaine section, après une brève description de quelques modèles auditifs fondés sur des concepts physiologiques ou psychoacoustiques, quelques applications de ces modèles à la reconnaissance de la parole sont citées.

B.2.2 Les modèles physiologiques et psychoacoustiques

B.2.2.1 Les modèles physiologiques

Les modèles physiologiques peuvent être regroupés en deux catégories [Del84] :

1. les modèles qui visent à fournir une représentation spectrale au niveau des fibres nerveuses,
2. les modèles qui visent à reproduire le traitement des caractéristiques dynamiques de la parole.

De notre point de vue, les caractéristiques dynamiques de la parole sont essentiellement reproduites à l'aide des mécanismes d'adaptation à court terme et de masquage postérieur. Les deux types de modèles ont été utilisés en reconnaissance de la parole. D'après nos connaissances actuelles sur le système auditif périphérique, le modèle optimal, en terme de possibilités à simuler des caractéristiques perceptuelles, varie en fonction de l'environnement acoustique [Gre88c].

Les modèles auditifs décrits dans [Lyo83, Sha86, Sha88, YS79, BCEG84, Ghi88] sont, si l'on se réfère à la précédente classification, du premier type. Lyon et Shamma [Lyo83, Sha86] introduirent une formulation linéaire des mécanismes de la membrane basilaire, une modélisation de la transformation des vibrations mécaniques de la membrane basilaire en déplacements des cellules ciliées ainsi qu'une modélisation simple, non-linéaire, de la transduction en potentiels électriques. Dans un nouveau modèle [Sha88], Shamma proposa l'utilisation de réseaux d'inhibition latérale (en anglais "lateral inhibitory networks" ou *LIN*) afin de reproduire le mécanisme d'inhibition latérale. Young et Sachs [YS79] présentèrent une mesure utilisant les propriétés de synchronisation des fibres nerveuses pour représenter les voyelles. Cette mesure, appelée indice moyen de synchronisation (en anglais "average localized synchronized rate" ou *ALSR*), est calculée en moyennant l'amplitude spectrale d'histogrammes, obtenus à une fréquence donnée F, sur un groupe de fibres nerveuses dont la fréquence caractéristique est proche de la fréquence F. Blomberg et al. [BCEG84] proposèrent un modèle dérivé de celui de Young et Sachs, où la *fréquence dominante* de réponse pour chaque fibre le long de la membrane basilaire est extraite d'une représentation obtenue à l'aide d'un banc de filtres critiques. Cependant, cette représentation ainsi que celle proposée par Ghitza [Ghi88], qui utilise un ensemble d'histogrammes d'intervalles de temps (en anglais "ensemble interval histogram" ou *EIH*) comme représentation spectrale, ne respecte plus le principe de *tonotopie* énoncé en section A.2.1.4. Citons enfin les modèles de Dolmazon [Dol82] et d'Alinat [Ali73] qui simulent les vibrations mécaniques de la membrane basilaire ainsi que certains mécanismes de la cochlée.

Des modèles auditifs visant à traiter les caractéristiques dynamiques de la parole ont aussi été proposés [Del82, Del84, Coh85, Cae79, Sen86]. Delgutte [Del82, Del84], Caelen [Cae79], et Cohen [Coh85] inclurent dans leur système une modélisation de l'adaptation à court terme. Ces modèles sont particulièrement sensibles aux changements rapides d'intensité qui interviennent dans le spectre de fréquence. Le modèle auditif

développé par Cohen a été utilisé dans un système de reconnaissance de parole à base de modèles de Markov cachés. Seneff [Sen86] proposa un système modélisant à la fois l'adaptation à court terme et le masquage postérieur. Ajoutant à ce modèle un détecteur de synchronisme ainsi qu'un détecteur d'enveloppe, Seneff l'appliqua, respectivement, à la detection de la fréquence fondamentale et à de la classification grossière. Une revue des modèles auditifs utilisant des concepts physiologiques est faite dans [Gre88b]. Pour de plus amples informations, le lecteur pourra s'y reporter.

B.2.2.2 Les modèles psychoacoustiques

Les concepts de sonie et de bandes critiques [ZS65] sont devenus plus ou moins des concepts standards souvent introduits dans les modèles psychoacoustiques. Des exemples de tels modèles sont décrits dans [ZT79, BCEG84, KK84, Bla87, HHW85b]. Quelques unes des non-linéarités utilisées dans les modèles physiologiques et décrites précédemment (voir section A.2.2.3), peuvent aussi être expliquées d'un point de vue psychoacoustique. Les mécanismes d'adaptation [Bla87], de masquage [Bla87, BCEG84], et de saturation [HHW85b, Bla87, BCEG84] ont aussi été inclus dans des modèles psychoacoustiques. Citons enfin l'étude de Schwartz [Sch81b] essayant de corréler des données psychoacoustiques avec les histogrammes de décharges des fibres nerveuses et le modèle de Klatt [Kla82] qui est un système mixte modélisant à la fois des concepts physiologiques et psychoacoustiques.

B.2.2.3 Application à la reconnaissance de la parole

Les applications réalisées à l'aide de modèles auditifs ont été essentiellement l'extraction d'indices [Sha88, Bla85], le développement de spectrogrammes auditifs [CG82, Kla82], et la reconnaissance de la parole [Ghi88, HL86, BCEG84, Her87]. Lorsque ces modèles furent appliqués à la reconnaissance de la parole en tant que modèles d'analyse acoustique, ils n'ont pas permis d'améliorer systématiquement les scores de reconnaissance par rapport à des méthodes plus traditionnelles comme les techniques *LPC* ou bancs de filtres. En particulier, Blomberg et al. rapportèrent que, dans certains cas, les performances obtenues étaient même moins bonnes [BCEG84]. Une explication possible est que ces modèles ne reproduisent qu'une petite partie du système auditif humain. Une grande partie de l'algorithme de reconnaissance n'est pas en accord avec le processus humain. Cependant, certaines études ont montré que des résultats encourageants pouvaient aussi être obtenus avec des modèles auditifs. Hermansky [Her87] montra qu'un ordre réduit du modèle d'analyse perceptivement fondée, *PLP*, donnait de meilleurs scores de reconnaissance que des modèles utilisant la technique *LP* en reconnaissance multilocuteur (multi-références). Cohen [Coh85] obtint, grâce à un modèle auditif, une diminution de 40% du taux d'erreur de reconnaissance, par rapport à un système utilisant un banc de filtres, pour de la reconnaissance monolocuteur de phrases formées de mots isolés. Enfin, Ghitza et Hunt [Ghi88, HL86] montrèrent que les modèles auditifs pouvaient être particulièrement intéressants pour traiter de la parole dégradée par du bruit.

Dans la brume de l'Hiver
j'apprenais enfin qu'il y
avait en moi un invincible
été.

Albert Camus

PARTIE C
RECONNAISSANCE AUTOMATIQUE DE MOTS ISOLES

En reconnaissance de parole par ordinateur, trois types de problèmes, classés par ordre de difficultés croissantes, peuvent être distingués : la reconnaissance *de mots isolés*, la reconnaissance *de mots enchaînés*, et la reconnaissance *de la parole continue*. Ce qui est appelé de la parole continue, c'est la conversation naturelle pratiquement sans contraintes. La parole constituée de mots enchaînés est un compromis entre les deux autres types de parole. Dans ce cas, le locuteur n'a pas besoin de séparer les mots prononcés par des pauses, mais il doit obéir à un certain nombre de règles lui fixant les mots à utiliser, la syntaxe, etc. Dans ce travail, nous nous sommes intéressés à la reconnaissance de mots isolés (un mot est toujours suivi d'une pause). Après une introduction sur le domaine, les mesures de distance couramment utilisées dans les algorithmes traitant de ce type de problème sont présentées. Les récents travaux visant à triompher des limites des systèmes de reconnaissance conventionnels sont ensuite examinés. Enfin, les problèmes posés par la reconnaissance automatique de la parole en présence de bruit sont soulevés et les travaux menés dans ce domaine sont présentés.

Chapitre C.1 PRELIMINAIRES

Qu'y a-t-il donc en un nom ?
Ce que nous nommons rose,
sous un autre nom, sentirait
aussi bon.

William Shakespeare

C.1.1 Généralités

Une des premières restrictions des systèmes de reconnaissance de mots isolés, c'est la taille du vocabulaire. Les performances des systèmes actuels sont fortement influencées par la taille et la difficulté des vocabulaires utilisés [PP89].

De nos jours, dans la plupart des systèmes, le signal de parole est représenté par une séquence de vecteurs (appelés aussi "frames") calculés toutes les 10 à 20 ms. Considérant chaque mot comme une unité indivisible (reconnaissance globale), le système compare les vecteurs de paramètres du mot prononcé aux vecteurs de paramètres (stockés en mémoire) de chaque mot du vocabulaire. Le mot ressemblant le plus au mot prononcé est alors désigné comme étant le mot reconnu. De tels systèmes opèrent généralement sur des vocabulaires de taille petite ou moyenne (10 à 200 mots). Cependant, pour des vocabulaires de grande taille, le temps de calcul est trop important pour des ordinateurs utilisant une seule unité centrale. De plus, la procédure d'apprentissage devient difficile car tout le vocabulaire doit être prononcé. Ainsi, afin de traiter des vocabulaires de taille importante, des unités de référence plus petites sont utilisées : phonèmes, syllabes, demi-syllabes [ASS85, DS77, RRWK83]. Toutefois, le passage à des formes de référence plus petites entraîne une modification des algorithmes utilisés et par voie de conséquence une diminution des performances obtenues [RRWK83]. Dans ce travail, nous nous sommes intéressés à la reconnaissance de mots isolés pour des *vocabulaires de taille petite ou moyenne*.

Les systèmes actuels peuvent être regroupés en deux catégories : monolocuteur et multilocuteur. Les systèmes monolocuteur, pour s'adapter à un nouveau locuteur, nécessitent l'enregistrement au préalable de tous les mots du vocabulaire, énoncés par ce locuteur. Les systèmes de reconnaissance multilocuteur ne nécessitent pas un tel apprentissage et sont donc très utiles lorsque le nombre de locuteurs testant le système est important. Dans ce cas, un apprentissage est réalisé au préalable en utilisant une population supposée représentatrice de l'ensemble des locuteurs de test. Certains systèmes, grâce à des procédures d'apprentissage des caractéristiques du locuteur de test, s'adaptent automatiquement au fur et à mesure que le locuteur utilise le système. Dans

ce travail, nous nous sommes intéressés à la *reconnaissance multilocuteur* de mots isolés, et nous n'avons utilisé aucune procédure d'adaptation automatique.

Une façon simple mais inefficace de tenir compte de la variabilité de la parole est de stocker systématiquement en mémoire toutes les formes de référence des différents locuteurs constituant la population d'apprentissage. Généralement, le nombre des formes de référence est réduit à l'aide d'algorithmes de classification (en anglais "clustering"). Les deux techniques les plus couramment utilisées en reconnaissance multilocuteur sont la technique des *K-means* [LRRW79] et la technique *UWA* (abréviation de "unsupervised without averaging") [RW79]. Ce sont des méthodes non-hiérarchiques (il n'y a pas de hiérarchie entre classes) qui nécessitent toutefois une connaissance minimale sur les données à classer. Afin de réduire la quantité d'informations maniplulées, une technique de quantification vectorielle [BGGM80] est quelquefois utilisée. Cette technique utilise deux étapes : la détermination d'un "codebook" qui est un ensemble fini de vecteurs prototypes représentatifs des données d'apprentissage, et la caractérisation de chaque vecteur par un numéro. La quantification vectorielle est à même de réduire l'espace nécessaire au stockage des données. C'est pourquoi cette technique de codage est utilisée de plus en plus souvent comme première étape d'un système de reconnaissance [RLS83, AHW87].

Les performances des systèmes de reconnaissance de parole sont difficiles à évaluer [Mar82, BBM88]. Certains paramètres comme les équipements d'enregistrement, la taille et la difficulté du vocabulaire, la difficulté de la tâche (monolocuteur ou multilocuteur), sont des facteurs importants à prendre en considération. Pour mesurer la performance des systèmes, le taux de reconnaissance est couramment utilisé. Toutefois, Moore [Moo77] proposa une mesure qui prend en compte la complexité du vocabulaire en tenant compte des confusions réalisées par l'être humain.

Pour de plus amples informations sur la reconnaissance automatique de mots isolés, le lecteur pourra se reporter à plusieurs articles généraux [Red76, Whi76, RL81, Vai85, Mar87] et livres [Bri86, O'S87] cités en référence.

C.1.2 Les algorithmes de reconnaissance traditionnels

C.1.2.1 La programmation dynamique

Un système de reconnaissance de mots isolés pour un petit vocabulaire utilise généralement le mot comme unité de reconnaissance. Cependant, la variation d'élocution conduit à l'obtention de formes vocales qui ont des longueurs et des rythmes différents pour un même mot, ces distorsions n'étant pas linéaires. Afin de résoudre le problème d'alignement entre la forme test et la forme de référence, l'algorithme de comparaison, appelé algorithme de *programmation dynamique*, applique une fonction de recalage temporel. Cet algorithme [Vin68, SC71] utilise une mesure de distance qui indique le taux de dissemblance entre les deux formes comparées. La programmation dynamique est une méthode très utilisée, qui donne de bons résultats (e.g. [Boy87]) mais qui est coûteuse en temps de calcul. De plus, chaque vecteur de paramètres est pondéré de la même façon lors du calcul du taux de dissemblance, ce qui ne permet pas de traiter particulièrement certaines parties importantes d'un mot. Pour l'influence de certains paramètres (contrainte de pente, pondération locale, etc) de l'algorithme de programmation dynamique sur les scores de reconnaissance, le lecteur pourra se reporter à l'article de Myers [MRR80].

L'augmentation de puissance des microprocesseurs et plus récemment l'arrivée des circuits intégrés de traitement de signal (Intel 2920, Nec 7720, TMS 320, AMI S2811, etc) permit de réaliser des systèmes de reconnaissance avec quelques circuits tout en palliant aux problèmes des temps de calcul rencontrés. En particulier, l'algorithme de programmation dynamique fut intégré dans un circuit VLSI donnant lieu à des réalisations de systèmes de reconnaissance de mots isolés (1000 mots) [Kav84] ou de mots enchaînés (quelques centaines de mots : *GSM d'AT&T*, *MUPCD* du *LIMSI*) par circuit.

C.1.2.2 Les modèles de Markov cachés

Une alternative à la programmation dynamique pour la reconnaissance automatique de la parole (*RAP*) est obtenue grâce à un modèle stochastique connue sous le nom de *modèle de Markov caché* (en anglais "hidden Markov model" ou *HMM*). Il a été montré que, pour beaucoup d'applications, ce modèle fournit des performances comparables à l'algorithme de programmation dynamique avec, toutefois, des performances supérieures en temps de calcul. Il a été utilisé par beaucoup de chercheurs travaillant en traitement de la parole (e.g. [Bak75, Jel76, Mar85, AHW87]). Un des points important des modèles de Markov cachés est que le nombre d'états, dont le rôle est de modéliser les événements acoustiques de la parole, n'a pas besoin d'être le même que le nombre de "frames". Ainsi, plusieurs "frames" sont associées à un même état. Des transitions entre états, auquelles sont attribuées des probabilités, décrivent l'évolution temporelle de la forme modélisée. Les modèles de Markov cachés utilisent aussi une forme de programmation dynamique pour déterminer la séquence des états $i_1, i_2, ..., i_N$. Cependant, comme le nombre d'états à examiner est généralement beaucoup moins important que le nombre de "frames", l'avantage le plus important des modèles de Markov cachés, comparés à l'algorithme de programmation dynamique, est le temps de calcul en phase de reconnaissance. Toutefois,

ces modèles ont d'importants problèmes d'apprentissage car ils nécessitent beaucoup de données afin d'extraire des paramètres robustes. Enfin, même si les modèles de Markov cachés fournissent de bons résultats pour beaucoup d'applications, il est très difficile d'inclure des connaissances (e.g. acoustiques) dans ces modèles. C'est aussi le cas pour la méthode de programmation dynamique. Ceci implique que, pour des tâches complexes comme la reconnaissance de la parole continue, d'autres méthodes doivent être envisagées.

Pour d'autres informations sur les modèles de Markov cachés, le lecteur pourra se reporter aux articles de Levinson et al. et de Rabiner et al. [LRS83, RLS83].

C.1.2.3 Utilisation de connaissances sur la parole

La reconnaissance à base d'indices est une alternative à la comparaison de formes. Ce type de reconnaissance utilise des indices extraits automatiquement sur le signal de parole. Ces indices sont des paramètres acoustiques permettant d'identifier des événements phonétiques. Les recherches dans ce domaine ont été motivées par des expériences de lecture de spectrogrammes [ZC79, CZ80, CRZR80, SEM86, Car86]. Ces expériences ont montré qu'il était possible de déterminer des indices acoustiques constituant des informations indépendantes du locuteur pour les différents segments phonétiques. Ceci amena, en particulier, le développement de systèmes de reconnaissance de mots isolés comme *FEATURE* [Col83] et *MULTIFON* [Fon84] et de systèmes de décodage acoustico-phonétique de la parole continue comme *APHODEX* [Foh86] réalisé au C.R.I.N et *SERAC* [Gil84] du C.N.E.T. Dans le même esprit, Morishima et al. [MHM86] utilisèrent des heuristiques et des connaissances permettant d'inférer certains types de consonnes afin de faciliter la prise de décision.

Le problème général rencontré par les systèmes manipulant des connaissances exprimées à l'aide d'indices est de trouver comment combiner ces indices afin de fournir des décisions cohérentes. Plus généralement, les problèmes qui se posent sont liés à la représentation des connaissances et à la stratégie adoptée pour les utiliser. Dans ce domaine, beaucoup de travail reste encore à faire.

Une des tendances actuelles est d'inclure des connaissances dans les systèmes conventionnels utilisant des techniques comme la programmation dynamique ou les modèles de Markov cachés. Le but visé est d'améliorer les performances ou les temps de calcul. Par exemple, les transitions entre phonèmes ont été utilisées pour donner plus d'importance aux taux de dissemblance d'un algorithme de programmation dynamique lorsque les changements spectraux sont importants [Jun87]. Les mesures de distance tendent à prendre en compte des propriétés ayant trait à la perception de la parole [Kla82]. Broad [Bro86] utilisa une modèle de coarticulation pour tenir compte du fait que, pour une vitesse d'élocution rapide, certains paramètres (comme les trajectoires des formants) ont varié par rapport à une vitesse d'élocution lente. Enfin, certains travaux récents sur les modèles de Markov cachés peuvent être considérés comme une combinaison d'approches statistiques et phonétiques [RJ86, Sch86].

C.1.2.4 Connexionnisme

Les *réseaux de neurones* (connus aussi sous le nom *modèles connexionnistes*) ont récemment reçu beaucoup d'attention dans des domaines comme l'intelligence artificielle ou le traitement de la parole. En traitement de la parole, un des buts est d'acquérir des connaissances dans la transformation signal-symbole par apprentissage automatique. Beaucoup de chercheurs ont été principalement intéressés par les propriétés d'apprentissage automatique attribuées aux réseaux de neurones. En traitement de parole, ces modèles ont été appliqués tout d'abord à des problèmes de classification puis ensuite à la reconnaissance.

Un réseau de neurones est un ensemble d'unités de traitement, appelées neurones, et de connexions. A chaque noeud est associé un nombre réel qui est son activation. A chaque connexion est associée un nombre réel qui est son poids. Ce poids représente la force de la connexion entre deux neurones. Les *perceptrons à niveaux multiples* (en anglais "multilayer perceptrons" ou *MLP*) constituent une classe de machines connexionnistes souvent utilisée en traitement de la parole. Un *MLP* réalise tout simplement une application d'une forme à n dimensions vers m symboles de sortie, à travers plusieurs niveaux intermédiaires. Dans un tel réseau, le poids des connexions peut être obtenu par apprentissage automatique en utilisant l'algorithme de rétro-propagation [RHW86] des erreurs. L'algorithme d'apprentissage, qui est itératif, consiste à minimiser une erreur quadratique entre le vecteur de sortie et celui désiré,

$$E_p = \frac{1}{2} \sum_j (t_{pj} - o_{pj})^2 . \tag{17}$$

où t_{pj} représente la valeur cible pour le modèle p et l'unité j, et o_{pj} représente la valeur correspondante en sortie. Tous les poids des connexions sont mis à jour à chaque passage, dans le but de diminuer cette erreur. Un neurone est généralement représenté par une fonction continue non-linéaire de la combinaison linéaire des sorties des neurones du niveau inférieur. Cette fonction est souvent de la forme :

$$y = \frac{1}{1 + e^{-x}} . \tag{18}$$

Les perceptrons à niveaux multiples ont été appliqués avec succès au traitement de la parole dans diverses applications comme la classification de voyelles [HL87, Phi87] ou l'identification de phonèmes [Wai87, BW88]. Citons aussi les travaux de Béroule [Bér85] sur le développement d'un modèle de mémoire d'inspiration psychologique et de Kohonen [Koh88] sur la réalisation d'une machine à écrire à entrée vocale basée sur la reconnaissance par phonèmes. Enfin, des études récentes, appliquées à la reconnaissance de parole, ont montré comment des connaissances acoustiques et phonétiques pouvaient être introduites dans des systèmes connexionnistes [LZ88]. Les réseaux de neurones constituent une alternative aux méthodes plus conventionnelles comme les modèles statistiques ou à base de connaissances. Lorsque beaucoup d'informations doivent être

combinées, comme dans les systèmes à base de règles de production, les réseaux de neurones constituent une approche intéressante à cause de leurs possibilités de généralisation. Comme dans les modèles de Markov cachés, l'étape de reconnaissance est très rapide. Les valeurs associées aux neurones de sortie sont calculées en une seule passe. La procédure d'apprentissage peut être très longue [HL87] mais ne nécessite pas de distribution statistique sous-jacente comme dans le cas des modèles de Markov cachés. Etant donné leurs facultés de généralisation, ces modèles facilitent le traitement de stimuli dégradés comme la parole en présence de bruit [BW87, BW88, TW88, Leb88]. Cependant, il a été montré que, lorsque la topologie du réseau n'est pas bien adaptée à l'application considérée, les performances diminuent [Kos87]. Enfin, à l'inverse des modèles de Markov cachés, un problème lié aux réseaux de neurones est leur difficulté à tenir compte du caractère séquentiel de la parole. Une introduction aux réseaux de neurones et à leurs applications pourra être trouvée dans [Lip87, Lip88].

Chapitre C.2 LE SYSTEME ET LES SOLUTIONS VISES

C.2.1 Introduction

Les sections précédentes ont été consacrées à une revue des travaux effectués en reconnaissance de la parole, particulièrement en reconnaissance de mots isolés, ainsi qu'à l'introduction des connaissances nécessaires à la compréhension du travail présenté. Avant de passer aux études réalisées, les problèmes étudiés sont rappelés et les solutions ensisagées sont explicitées en essayant de clarifier l'incidence de chacune des études sur le système final de reconnaissance automatique de la parole tel que nous le concevons.

C.2.2 Réalisation d'un système robuste de reconnaissance de mots isolés

Les chapitres précédents ont mis en évidence les principales limitations des systèmes de reconnaissance automatique multilocuteur de mots isolés actuels, à savoir :

1. la difficulté de prendre en compte les variations intra- et interlocuteur,
2. la dificulté de traiter des vocabulaires difficiles,
3. le manque de robustesse au bruit.

Dans le travail réalisé, ces différents problèmes ont été étudiés afin d'aboutir à un système de reconnaissance multilocuteur de mots isolés performant pouvant opérer dans des conditions réelles et pas seulement dans des conditions de laboratoire. Les problèmes liés à la robustesse des systèmes de reconnaissance automatique de la parole sont très difficiles à résoudre et aucune solution convenable n'est actuellement disponible. L'essentiel de notre travail a utilisé comme base un algorithme de programmation dynamique pour effectuer la comparaison des mots de test et de référence. Toutefois, un tel système ne tient pas compte des caractéristiques particulières du signal de parole. Aussi nous avons cherché à introduire dans ce système des connaissances liées à différents domaines comme la psychoacoustique, la physiologie et la phonétique. Ces connaissances permettent de simuler certains phénomènes propres à l'être humain et tiennent compte des caractéristiques uniques du signal de parole qui sont différentes de tout autre signal. Dans notre esprit, l'amélioration des systèmes de reconnaissance automatique actuels passe par l'étude du système humain dans son ensemble et l'application des résultats de ces études grâce à la réalisation de modèles fonctionnels. Cependant, nos connaissances en la matière sont limitées. C'est la raison pour laquelle nous nous sommes orientés vers des études de modèles hybrides permettant d'intégrer quelques-unes de nos connaissances tout en n'oubliant pas notre ignorance. Ceci nous a amené aux études suivantes :

1. étude de plusieurs modèles d'analyse acoustique, en particulier des modèles auditifs, et mesures de distance afin de déterminer ceux qui permettent de prendre en compte au mieux la variabilité intra- et interlocuteur et par voie de conséquence d'obtenir les meilleures performances,
2. optimisation du meilleur modèle trouvé grâce 1) à des tests effectués en reconnaissance monolocuteur et interlocuteur permettant de raffiner l'adéquation entre le modèle d'analyse acoustique et la mesure de distance et 2) à l'introduction de connaissances physiologiques,
3. étude du comportement de plusieurs modèles d'analyse acoustique et mesures de distance en présence de bruit additif,
4. développement d'un nouveau modèle d'analyse acoustique fondé sur des concepts physiologiques et plus robuste en présence de bruit,
5. étude de l'incidence de l'effet Lombard sur plusieurs modèles d'analyse acoustique,
6. étude d'un système permettant d'effectuer la discrimination entre des mots acoustiquement similaires grâce à l'utilisation de connaissances phonétiques.

Dans les prochains chapitres (à partir de la partie D) ces études sont développées et les résultats obtenus sont rapportés. De plus, quelques outils réalisés afin de faciliter les buts visés sont aussi détaillés.

Chapitre C.3 SIMILITUDE ET MESURES DE DISTANCE

L'amitié sans confiance,
c'est une fleur sans parfum.

Laure Conan

C.3.1 Introduction

Tout système de *RAP* utilisant des formes de référence emploie un module qui compare des segments de parole en terme de similitude ou dissimilitude. Ce module est divisé en deux parties : le *module d'analyse*, qui est chargé d'extraire un ensemble de paramètres à partir du segment de parole, et le *module qui contient la mesure de distance*. Ces deux modules interagissent. Ils ne doivent donc pas être étudiés séparément.

Quelques uns des premiers travaux menés en identification ou vérification du locuteur [Ata74, Pfe74] utilisaient une distance Euclidienne calculée sur les coefficients de prédiction linéaire ou les coefficients *PARCOR*. Cependant, la distance Euclidienne appliquée à de tel paramètres n'a pas de signification physique dans le domaine des fréquences. Il est important de tenir compte de la signification physique de la mesure de distance utilisée.

Dans leur discussion théorique sur les mesures de distance appliquées au traitement de la parole [GM76], Gray et Markel proposèrent les quatre conditions nécessaires que doit satisfaire une mesure de distance pour la parole :

1. $d(x,y)=d(y,x)$ symétrie,
2. $d(x,y)>0$ pour x différent de y et $d(x,x)=0$ (définie positive),
3. $d(x,y)$ doit pouvoir être interprétée physiquement dans le domaine des fréquences,
4. le calcul de $d(x,y)$ doit être rapide.

Une mesure de distance $d(x,y)$ est appelée *métrique* si elle satisfait les trois conditions [DH73] : commutativité, transitivité, et l'inégalité triangulaire. Toutes les mesures de distance utilisées en *RAP* ne sont pas des métriques. Citons, en particulier, la distance de *Mahalanobis* [Mah36] qui est définie par :

$$d(x,y) = (x-y)^T W^{-1} (x-y), \tag{19}$$

où W^{-1} est une matrice définie positive qui permet différentes pondérations. Celles-ci dépendent de la capacité des vecteurs de caractéristiques à identifier les segments

de parole dans l'espace des caractéristiques. Cette mesure de distance n'est pas une métrique car une matrice différente doit être calculée pour chaque mot. Toutefois, en pratique, une matrice fixe est utilisée. Si W (et W^{-1}) est la matrice indentité, alors $d(x,y)$ est la *distance Euclidienne*. Dans le cas général, W est la matrice d'autocovariance du vecteur de référence. Le principal inconvénient de cette mesure de distance provient de la difficulté à estimer correctement W^{-1} à partir de données limitées.

C.3.2 Les mesures de distance spectrales

C.3.2.1 La mesure de distorsion d'Itakura-Saito

La proposition de cette *mesure de distorsion* [IS70, Ita75] a joué un rôle de détonateur au niveau de la reconnaissance de parole dans le domaine des fréquences *LPC*. Cette mesure de distorsion, calculée entre deux formes représentées par un vecteur *LPC* de test a' et un ensemble de L vecteurs *LPC* de référence a_i ($i=1, 2, ..., L$), est définie par :

$$d_{IS_i} = d_{IS}\left(a_i, a'\right) = \frac{\sigma_i^2}{\sigma'^2} \frac{a_i^T V' a_i}{a'^T V' a'} + log\left(\frac{\sigma'^2}{\sigma_i^2}\right) - 1, \tag{20}$$

où V' est la matrice d'autocorrélation de la forme test, et σ' et σ_i sont, respectivement, les gains *LPC* des formes de test et de référence. Cette mesure est clairement asymétrique. C'est pourquoi nous avons utilisé le terme mesure de distorsion et non pas le terme mesure de distance (en accord avec la définition de mesure de distance proposée par Gray et Markel).

C.3.2.2 Les mesures de distorsion : "log likelihood ratio" et "likelihood ratio"

La mesure de distorsion "log likelihood ratio" (*LLR*) [Ita75] part du principe que l'amplitude ne doit pas être utilisée en reconnaissance de la parole car un même mot peut être prononcé plus ou moins fort. La mesure *LLR*, qui est une variante de la mesure de distorsion d'Itakura-Saito où le gain a été normalisé, est définie par :

$$d_{LLR_i} = d_{LLR}\left(a_i, a'\right) = log\left[\frac{a_i^T V' a_i}{a'^T V a'}\right]. \tag{21}$$

Une autre possibilité est de positionner la valeur des gains, σ and σ', de telle façon que les formes de test et de référence soient comparées uniquement sur la base de leur spectre de fréquence (i.e. $\sigma=\sigma'$). Ceci entraîne la définition de la mesure de distorsion suivante :

$$d_{LR_i} = d_{LR}\left(a_i, a'\right) = \frac{a_i^T V' a_i}{a'^T V a'} - 1. \tag{22}$$

Cette mesure est appelée mesure de distorsion "likelihood ratio" (*LR*). Afin d'améliorer la non-symétrie de cette mesure de distorsion, Gray et Markel [GM76] ont proposé la mesure de distance *COSH* dans le cadre d'une application visant à évaluer la qualité d'un signal de parole synthétisé.

Dans les prochaines sections, pour des raisons de simplicité, nous utiliserons souvent les termes distorsion et distance à la place de mesure de distorsion et mesure de distance.

C.3.2.3 La distance cepstrale Euclidienne

Les coefficients cepstraux peuvent être calculés directement à partir des coefficients de prédiction. Ils correspondent aux coefficients de la série de Fourier du logarithme du spectre de fréquence. Atal [Ata74] et Junqua [Jun87] ont montré, respectivement pour l'identification (et la vérification) du locuteur et la *RAP*, que des mesures de distance utilisant des coefficients cepstraux donnaient les meilleurs résultats. Dans les travaux présentés, pour désigner la distance Euclidienne sur des coefficients cepstraux non-pondérés, nous avons utilisé le terme *distance cepstrale Euclidienne*. Cette distance est définie par la formule suivante :

$$d_{cep} = \sum_n (c_t(n) - c_r(n))^2 , \qquad (23)$$

où $C_t(n)$ et $C_r(n)$ sont les *n-ième* coefficients cepstraux d'un vecteur, respectivement, de la forme test et de la forme de référence.

C.3.2.4 La distance cepstrale pondérée

La distance *cepstrale pondérée* est une variante de la distance cepstrale Euclidienne. Elle est définit par :

$$d_{wcep} = \sum_n (w(n)(c_t(n) - c_r(n))^2 , \qquad (24)$$

où $w(n)$ est la pondération appliquée aux coefficients cepstraux. Dans le cas où $w(n)$ est une fonction linéaire, la pondération peut être considérée comme faisant partie de la représentation paramétrique de la parole ou de la mesure de distance. Dans la suite, les deux possibilités seront utilisées.

Pondérer les coefficients cepstraux par $w(n)$ est équivalent à travailler dans le domaine *"group delay spectrum"* (obtenue à partir de la phase de la transformée de Fourier du signal [YSR84]) au lieu de travailler dans le domaine du cepstre *LPC* [IU87]. De précédentes études ont montré que la distance cepstrale pondérée était très utile en vérification du locuteur [Fur81] et en *RAP* [Pal82]. Lorsque $w(n)=n$, les coefficients cepstraux pondérés de (24) (i.e. $n \times c(n)$) sont appelés les coefficients *"root-power sums"* [Sch81a]. Dans les prochaines sections, la distance Euclidienne appliquée aux coefficients "root-power sums" sera utilisée sous le nom distance *RPS*. Ses bonnes performances en *RAP* ont déjà été rapportées [Pal82, HW86]. Lorsque la même pondération, $w(n)=n$, est

appliquée à un autre ensemble de coefficients que les coefficients cepstraux (*PARCOR*, etc), la distance Euclidienne utilisant ces coefficients pondérés est appelée *distance pondérée par l'index*. Cependant, dans ce cas la distance perd sa signification physique.

Une autre pondération des coefficients cepstraux est connue sous le nom de *"bandpass liftering (BP)"* [JRW86]. Elle a été très utilisée sur des coefficients dérivés de la technique *LPC*. Lorsque cette fonction de pondération est appliquée au modèle d'analyse d'ordre quatorze (qui sera utilisée par la suite avec ce type de pondération) elle peut être exprimée par :

$$w(n) = 1 + 10.5 sin\left(\frac{\pi n}{21}\right) \quad for\ n = 1\ to\ 21. \qquad (25)$$

Il a été montré qu'une distance cepstrale pondérée utilisant cette fonction de pondération permettait de diminuer l'erreur de reconnaissance obtenue avec une distance cepstrale Euclidienne [JRW86].

Itakura and Umezaki [IU87] proposèrent pour $w(n)$ une combinaison de "lifter" Gaussien et de "lifter" exponentiel définie par la formule suivante :

$$w(n) = n^s exp\left(\frac{-n^2}{2\tau^2}\right) \quad (s \geq 0), \qquad (26)$$

où s et τ sont des paramètres qui peuvent être ajustés. Typiquement, afin de lisser le spectre *LPC*, la distance cepstrale pondérée qui utilise cette fonction de pondération doit utiliser un nombre important de coefficients cepstraux (voir [IU87]). Une telle distance, évaluée en *RAP* avec la technique LPC [IU87], permit d'obtenir les meilleurs scores de reconnaissance pour $s=1$ et $\tau=5$.

Tohkura [Toh87] utilisa un cas particulier de la distance de Mahalanobis où la matrice de covariance V' etait formée uniquement de sa partie diagonale (ce qui suppose que les autres termes de la matrice sont négligeables). Cette distance, appliquée aux coefficients cepstraux, devient une distance cepstrale pondérée où la pondération $w(n)$ est l'inverse du *n-ième* élément de la matrice de covariance V'. Grâce à cette distance, chaque coefficient cepstral $c(n)$ est égalisé au niveau variance par la pondération $w(n)$. Cette mesure, qui dépend de la quantité de données utilisées lors de l'apprentissage pour calculer la matrice V', s'est avérée mieux se comporter dans le cadre d'applications d'identification du locuteur et de *RAP* que des distances plus traditionnelles comme les distances cepstrale Euclidienne ou "likelihood ratio" [SR86, Toh87].

C.3.3 Les mesures de distance et la perception de la parole

C.3.3.1 Introduction

Il est souhaitable, lorsque c'est possible, de relier une distance mathématique à une certaine forme de phénomènes perceptuels. Gray et Markel [GM76] discutèrent les relations entre les mesures de distances et les mouvements des formants, sur la base d'études perceptuelles rapportées par Flanagan [Fla55b, Fla55a, Fla72]. En 1981, Sugiyama [SS81] proposa des mesures pondérant les pics du spectre de fréquence en s'appuyant sur le fait que le système auditif humain était plus sensible aux pics du spectre à court terme qu'aux vallées. En 1982, Klatt [Kla82], sur la base d'études utilisant des distances phonétiques mesurées par des humains, proposa des distances sensibles à la pente du spectre de fréquence.

Les caractéristiques du système auditif humain ne sont pas les mêmes sur l'ensemble des fréquences audibles. L'oreille a une plus grande sensibilité dans les basses fréquences. Utilisant cette connaissance, des mesures spectrales ont été proposées. Ces mesures sont connues sous les noms de *mesure spectrale à pondération en fréquence* [CM82, SS87] et *mesure spectrale à déformation en fréquence* [NSRK85, Nod88]. Il a été montré que ces mesures spectrales étaient particulièrement utiles dans des environnements bruités [SS87, Nod88].

Une mesure de distance doit être performante en environnement bruité. De plus, il est souhaitable d'améliorer sa similitude avec le jugement humain. Ainsi, les deux conditions suivantes doivent être ajoutées à la définition de mesure de distance donnée par Markel et Gray,

1. les caractéristiques du système auditif humain doivent être prises en considération,
2. les conditions environnantes ne doivent pas être négligées.

Dans les prochaines sections, des mesures spectrales qui tiennent compte de caractéristiques du système auditif humain sont décrites. Leur application à des environnements bruités sera décrite dans un prochain chapitre.

C.3.3.2 Les mesures spectrales à pondération et à déformation en fréquence

Une mesure de distorsion Itakura-Saito à pondération en fréquence a tout d'abord été proposée par Chu et Messerrschmit [CM82] pour un vocodeur *LPC*. Ils pondérèrent alternativement un ou plusieurs pôles au cours de l'analyse spectrale du signal de parole. Pour de la parole non bruitée, ils obtinrent des résultats supérieurs à ceux obtenus sans aucune pondération. Pour la *RAP*, Soong et Sondhi [SS87] proposèrent une mesure de distorsion Itakura-Saito à pondération en fréquence de façon adaptative et étudièrent ses performances en environnement bruité. Cette mesure de distorsion est définie par :

$$d_{WI} = log \int_{-\pi}^{\pi} F(w) \frac{|B(w)|^2}{|A(w)|^2} \frac{dw}{2\pi}, \tag{27}$$

où w est la fréquence en radians, $B(w)$ et $A(w)$ sont, respectivement, le spectre de référence et le spectre de test et $F(w)$ une fonction de pondération agissant comme un facteur de modification de la largeur de bande du spectre test. Les performances de cette mesure spectrale se sont avérées être similaires à celles obtenues sans aucune pondération pour un rapport signal sur bruit (*SB*) élevé. En revanche, pour un faible *SB*, elle donna de bien meilleures performances. Ceci est dû à deux importantes caractéristiques de cette mesure spectrale :

1. une pondération spectrale non uniforme,
2. un ajustement adaptatif du facteur de pondération.

Les mesures de distorsion à déformation en fréquence ont été étudiées par Nocerino et al., et Noda dans [NSRK85, Nod88]. Nocerino proposa un algorithme efficace pour déformer l'échelle des fréquences en une échelle Bark, alors que Noda proposa une mesure spectrale à déformation en fréquence effectuant une expansion de l'axe des fréquences en fonction du *SB*, ceci à chaque fréquence du spectre. Si Nocerino n'observa aucune dégradation des performances en *RAP* dans un environnement non-bruité, Noda trouva que sa nouvelle mesure spectrale améliorait les performances, par rapport à des mesures traditionnelles, en vérification du locuteur et en présence de bruit.

C.3.3.3 Les mesures sensibles à la pente de fréquence

La *mesure de distorsion sensible à la pente de fréquence* proposée par Klatt [Kla82] a été testée avec succès en *RAP* [NSRK85]. Toutefois une évaluation de cette mesure par comparaison à des mesures plus mathématiques dans le cadre de sons synthétiques et de l'identification de voyelles ne lui fut pas favorable [YT86]. Cette mesure de distorsion donne plus d'importance aux pics du spectre de fréquence qui sont perceptivement significatifs. Alors que cette mesure avait été appliquée originellement à des spectres issus de filtres à bandes critiques, Hanson et Wakita [HW86] l'appliquèrent à des spectres issus de modèles tout pôle. Ceci donne la formule suivante pour deux spectres issus de modèles tout pôle $\frac{1}{A_T}(w)$ (test) et $\frac{1}{A_R}(w)$ (référence) :

$$d_{SS} = \frac{1}{\pi} \int_0^{\pi} \left\{ \frac{\partial}{\partial w} log \left| \frac{1}{A_T(w)} \right|^2 - \frac{\partial}{\partial w} log \left| \frac{1}{A_R(w)} \right|^2 \right\}^2 \partial w. \tag{28}$$

Grâce à l'utilisation de propriétés bien connues des coefficients cepstraux, il a été montré [Sch81a] que cette formule pouvait être calculée en utilisant la distance *RPS* (définie à la section C.3.2.4), qui est un cas particulier de distance cepstrale pondérée. La distance *RPS* s'est avérée donner des performances régulièrement meilleures que la distance cepstrale Euclidienne, en particulier pour des environnements bruités [HW86].

C.3.4 Divisibilité : règles de décision et traitement statistique

Dans une approche traditionnelle de reconnaissance de formes, un algorithme de classification utilise un certain nombre de caractéristiques afin de déterminer une ou plusieurs classes. Habituellement, le but est de déterminer la classe la plus vraisemblable. En *RAP*, à cause de la large distribution des paramètres acoustiques utilisés, une classification ou reconnaissance précise des sons est très difficile. Ainsi, afin de faciliter le classement des mots (ou phrases) acoustiquement similaires, des scores sont souvent associés aux différentes hypothèses. Une des façons de modéliser les variations est d'utiliser un *modèle probabiliste* [Jel76, Sch86]. Cependant, en *RAP*, il y a eu beaucoup d'efforts entrepris pour essayer de reconnaître des mots ou des phonèmes à partir de connaissances fines sur les caractéristiques des phonèmes appartenant à différents contextes. Le plus souvent de tels systèmes n'utilisent pas des méthodes probabilistes mais plutôt un ensemble d'*heuristiques*. Ces heuristiques, qui amènent le système à trouver la prochaine "meilleure" hypothèse, varient d'un ensemble de conditions à un autre. Une autre approche consiste à utiliser un algorithme de classification fondé sur une mesure de distance. Comme les mesures de distance ont été présentées dans les sections précédentes, nous nous intéresserons dans cette section aux méthodes probabilistes et heuristiques.

Si les données statistiques sont suffisantes, un modèle probabiliste des variations des unités acoustiques utilisées est très utile pour distinguer ces unités entre elles. Les modèles probabilistes permettent facilement l'attribution de scores aux hypothèses concurrentes. Des rapports de *maximum de vraisemblance* permettent d'associer des scores aux hypothèses et le théorème de Bayes [AR77] est souvent utilisé pour prédire les hypothèses probables. Pour obtenir des performances intéressantes, les méthodes probabilistes nécessitent une phase d'apprentissage importante. Les modèles de Markov cachés en conjonction avec l'algorithme de Baum [BE67] fournissent un mécanisme aidant à la spécification du niveau de structure désiré et, à partir de ces contraintes, trouvent un modèle pour les données d'apprentissage utilisées. Dans les systèmes qui utilisent plusieurs formes de référence par mot du vocabulaire, la règle des plus proches voisins peut être utilisée [GLM84].

Une approche heuristique est souvent sélectionnée lorsque seulement une connaissance limitée des relations entre les différentes unités à classifier est disponible. Cette approche suppose que les relations entre les caractéristiques utilisées lors de la classification sont connues. Le programme fait des choix suivant un arbre de décision hiérarchique, et éventuellement désigne la forme sélectionnée. Cette méthode, qui donne des performances intéressantes (au niveau des résultats obtenus), a un principal défaut : les connaissances doivent être représentées sous la forme de programmes ou de règles qui utilisent un ensemble complexe de conditions et de seuils. Chaque changement dans le système peut être difficile. Des connaissances heuristiques peuvent aussi aider à maximiser la vraisemblance d'un groupe d'hypothèses qui seront utilisées pour déterminer la solution correcte. L'utilisation de connaissances heuristiques implique souvent le développement de stratégies complexes de contrôle et de recherche dans l'espace des caractéristiques ou des hypothèses.

Chapitre C.4 UTILISATION DE METHODES DISCRIMINANTES EN RAP

Le bien, s'il avait une cause, cesserait d'être le bien, tout comme s'il avait un effet.

Léon Tolstoï

C.4.1 Difficultés de l'approche traditionnelle de reconnaissance des formes

Les problèmes rencontrés dans les systèmes de reconnaissance de parole actuels peuvent être résumés par les points suivants :

1. variabilité due au locuteur. Le signal de parole contient des informations phonétiques et des informations dépendantes du locuteur qui ne sont pas faciles à séparer [HJ88],
2. variabilité de la parole (amplitude, vitesse d'élocution, prononciation). Un locuteur ne prononce un mot jamais de la même façon,
3. ambiguité. Il n'y a pas une correspondance biunivoque entre les variables acoustiques et phonémiques. Les humains pour résoudre le problème utilisent leur connaissance du langage et du contexte,
4. bruit de fond et plus généralement conditions d'enregistrement,
5. détection des frontières de mot. Les performances des systèmes de reconnaissance de parole dépendent beaucoup de la précision avec laquelle les frontières de début et fin de mot sont déterminées.
6. intégration des connaissances. Différentes sources d'informations doivent être combinées mais notre connaissance des processus d'intégration n'est pas très importante.

La plupart des points cités sont très liés au fait que le signal de parole est le plus souvent manipulé comme un signal quelconque. Les systèmes de reconnaissance automatique actuels ont beaucoup de mal à tenir compte des caractéristiques *uniques* du signal de parole. Ils utilisent souvent des modèles "d'ignorance" [MS85], comme l'algorithme de programmation dynamique, qui tiennent compte de tous les indices acoustiques même ceux pouvant conduire à des erreurs. L'algorithme de programmation dynamique donne la même importance à tous les détails acoustiques même si les transitions sont, perceptuellement, plus importantes que les parties stables du signal [FA86]. La plupart des systèmes de reconnaissance ne tiennent aucun compte d'informations linguistiques, auditives ou

perceptuelles. Les algorithmes de traitement du signal généralement utilisés sont très sensibles aux variations dues au locuteur, à l'environnement, la vitesse d'élocution, etc.

Des études destinées à inclure des connaissances sur la parole dans les systèmes traditionnels ou à réaliser des systèmes à base de connaissances sur la parole ont vu le jour (voir section C.1.2.3). Dans les systèmes à base de connaissances, l'extraction automatique d'indices est une partie importante du système. L'hypothèse souvent sous-jacente à de tels systèmes est qu'il existe des invariants dans la parole. Ceci revient à dire qu'à un segment de parole donné est associé un ensemble déterminé de caractéristiques acoustiques. Une telle hypothèse simplifie le raisonnement même si elle est quelquefois à la base des erreurs commises. En fait la parole est hautement variable [Kla85] et définir des indices robustes d'un son donné est une tâche très difficile.

C.4.2 Les méthodes discriminantes

La procédure permettant d'améliorer les performances obtenues pour des vocabulaires contenant des mots similaires acoustiquement est appelée *discrimination*. Les méthodes de discrimination orientent la reconnaissance sur les parties des mots qui sont différentes acoustiquement. Par définition, les informations discriminantes sont présentes uniquement dans une partie du mot. Avec des méthodes traditionnelles, des différences mineures mais de longue durée peuvent l'emporter sur des différences majeures et ainsi contribuer à de mauvaises décisions.

Rabiner et Wilpon, grâce à un système à deux passes, améliorèrent les scores de reconnaissance dans le cas d'un vocabulaire difficile contenant des mots acoustiquement similaires [RW81]. Tout d'abord le mot de test est assigné à une sous-classe du vocabulaire contenant des mots facilement confondus, grâce à un algorithme de programmation dynamique. Ensuite, le mot sélectionné est comparé aux mots de la même sous-classe à l'aide d'une fonction de pondération statistique qui donne plus d'importance à la partie du mot facilitant la discrimination. Une variante de cette méthode a été proposée par Casacuberta et al. [CV88] qui n'ont utilisé aucune hypothèse sur la distribution statistique sous-jacente des données manipulées. Ils utilisèrent une généralisation de la fonction linéaire discriminante généralisée [DH73], dans laquelle la dimension de l'espace de représentation dépend de la classe de mots manipulés. Grâce à cette méthode ils obtinrent une diminution de 50% de l'erreur de reconnaissance (en reconnaissance monolocuteur) obtenue avec un algorithme de programmation dynamique.

Lamel and Zue [LZ82] proposèrent une méthode donnant plus d'importance aux transitions consonne-voyelle pour le vocabulaire alphanumérique. Leur système permit de diminuer l'erreur de reconnaissance de 3.5% (en reconnaissance monolocuteur), par rapport à une procédure de reconnaissance standard, sur le vocabulaire *E-SET* (constitué des mots similaires contenant la voyelle /i/). De plus, en séparant le vocabulaire en sous-classes grâce au contour de l'énergie calculée pour les basses fréquences, ils obtinrent une diminution de 30% du temps de calcul.

Une autre possibilité est d'effectuer la discrimination sur la base d'indices phonétiques. Cole et al. [CSL85] utilisèrent un arbre hiérarchique de décision pour

reconnaître le mot prononcé. Un avantage d'une telle méthode est que la connaissance, exprimée sous la forme d'indices, peut être améliorée compte tenu des confusions du système. Cependant, l'arbre de décision est très dépendant de la liste de mots étudiés. Une telle méthode a permis de diminuer de 16.7% l'erreur de reconnaissance (en reconnaissance multilocuteur) sur le vocabulaire *E-SET*.

Bradshaw et al. [BCL82] proposèrent un système utilisant à la fois des formes de référence et des indices extraits automatiquement. Chaque indice extrait fut considéré comme une information supplémentaire ajoutée à chaque "frame" temporelle représentant le signal de parole. Une pondération adéquate fut ensuite appliquée aux différentes informations spectrales et temporelles avant d'utiliser un algorithme de programmation dynamique. Par comparison à d'autres systèmes de reconnaisance, Bradshaw et al. montrèrent que leur nouvelle méthode permettait d'obtenir les meilleurs résultats (10% d'erreur en reconnaissance monolocuteur sur le vocabulaire *E-SET*).

Les méthodes discriminantes permettent d'améliorer les performances des systèmes de reconnaissance pour des vocabulaires difficiles. Une alternative à ces méthodes est d'essayer d'améliorer la qualité des formes de référence. Dans ce travail, nous nous sommes intéressés uniquement aux méthodes discriminantes.

Chapitre C.5 RAP EN ENVIRONNEMENT BRUITE

Les gens ignorants croient que c'est le bruit que font les chats en se battant qui est insupportable. Ce n'est pas vrai ... c'est parce qu'ils font trop de fautes de grammaire.

Mark Twain

C.5.1 Introduction

Le problème de la *RAP* en environnement bruité a récemment intéressé beaucoup de chercheurs. La principale raison est que les systèmes actuels donnent de bons résultats en environnement de laboratoire mais ont des performances qui se dégradent rapidement dans des environnements bruités. Afin de prendre en compte ce problème, trois types d'approches ont été considérées [EWR87] :

1. amélioration des systèmes existants qui ont donné de bonnes performances en laboratoire,
2. conception de nouveaux systèmes, plus robustes en présence de bruit,
3. apprentissage dans les mêmes conditions que celles rencontrées lors des tests.

La troisième approche est quelquefois difficile à mettre en oeuvre, à cause des difficultés à reproduire pour l'apprentissage les conditions de tests. Quant aux deux premières approches, elles ont donné lieu à des développements sur :

1. les *mesures de distance spectrales* [HW86, IU87, MJ88a, MI86, SS87] qui utilisèrent souvent la technique de prédiction linéaire pour obtenir une représentation paramétrique de la parole,
2. la *compensation* et la *suppression* du bruit [CM81, EWR87, EMJ88, Kay80, NMW83, PB81],
3. le *développement de nouvelles méthodes d'analyse* [HL86, Ghi87] plus robustes en présence de bruit,
4. l'estimation spectrale de modèles *AR* et *ARMA* appliquée à la parole bruitée [Cad82, MJ88b].

Ces différentes approches ont généralement été testées en présence de bruit ajouté (bruit blanc ou bruit blanc filtré). Cependant, lorsque la parole est produite dans un environnement bruité, des changements interviennent au niveau de la structure phonétique de la parole à cause de l'effort vocal effectué. Aussi, ajouter du bruit pour tester la robustesse des systèmes est déjà une approximation grossière des conditions réelles.

Dans le cadre d'applications de reconnaissance de parole, les sources de bruit qui sont intéressantes sont les suivantes [Red76] :

1. bruit de fond dans un environnement de bureau (ventilateurs, néons, machines à écrire, claviers d'ordinateurs, conversations de fond). Ce type de bruit n'est pas de nature additive et est souvent caractérisé par des concentrations d'énergie dans certaines parties du spectre de fréquence. Il ne peut, généralement, pas être représenté par du bruit blanc,
2. bruit sur des lignes téléphoniques (clics, distorsion, écho, translation de fréquence),
3. bruit dû à la circulation dans une machine mobile (automobile par exemple). Ce type de bruit peut en général être simulé par du bruit ajouté,
4. bruit de moteur (automobile et avion) qui peut être simulé par du bruit périodique (des pics discrets à bande spectrale étroite).

Toutes ces sources de bruit sont difficiles à traiter. Le bruit blanc filtré, souvent utilisé pour tester les systèmes de reconnaissance, est seulement un cas particulier des environnments bruités auquels ces systèmes peuvent être soumis.

Quelques études visant à identifier les erreurs concernant la perception de distinctions phonétiques en présence de bruit ont déjà été menées. Les résultats obtenus ont montré que certaines distinctions étaient plus robustes que d'autres. Par exemple, dans le cas des consonnes, la nasalité et le voisement, qui sont liés à des indices basse fréquence, sont les plus robustes [MN55, WB73]. Pour les voyelles, les distinctions liées au premier formant sont plus robustes que celles liées au deuxième formant et à la durée [Pic57]. Néanmoins, ces résultats dépendent des caractéristiques spectrales du bruit considéré.

C.5.2 Détection des frontières de mot

La détermination des frontières de mot n'est pas très difficile si le *SB* est grand (supérieur à 60 dB). Malheureusement, la plupart des systèmes de reconnaissance se trouvant dans des environnements réels doivent fonctionner avec des *SB* beaucoup plus petits : typiquement 30 ou 40 dB et atteignant quelquefois 10 dB. Avec de telles conditions, les fricatives et les nasales de faible énergie et les sons voisés de faible amplitude, en début ou en fin de mot, sont difficiles à détecter. La détection des frontières de mot est une cause importante des dégradations de performance des systèmes de reconnaissance en présence de bruit.

Afin d'atténuer les problèmes, des microphones à casque se plaçant très près de la bouche et diminuant le bruit sont utilisés. Des mesures du bruit de fond, pendant des intervalles où il n'y a pas de parole, suivies d'une soustraction de celui-ci au signal de parole peuvent aussi être envisagées.

Très peu de chercheurs se sont intéressés au problème. Rabiner et Sambur ont tout d'abord utilisé l'énergie globale pour détecter des frontières approximatives de mots, puis ont raffiné leur détection en explorant des intervalles autour des premières frontières trouvées [RS75]. Martin proposa l'utilisation d'une technique de comparaison de formes

pour distinguer des sons provoqués par des bruits de respiration ou des souffles [Mar75]. Enfin, Wilpon et Rabiner présentèrent un algorithme utilisant des modèles de Markov cachés pour détecter les frontières de mots [WR87]. Dans la même étude, ils montrèrent aussi qu'une autre façon d'améliorer les performances des systèmes de reconnaissance en présence de bruit pouvait être l'élimination du module de détection explicite des frontières de mots.

La détermination précise des frontières du signal de parole est un des problèmes les plus critiques en reconnaissance de mots isolés mais aucune solution n'est à l'heure actuelle convenable, en particulier pour des environnements fortement bruités.

C.5.3 Application des modèles d'analyse acoustique à la parole bruitée

C.5.3.1 Analyse spectrale en milieu bruité

Il a été montré que l'analyse par prédiction linéaire, qui est probablement le modèle d'analyse acoustique le plus utilisé en reconnaissance de parole, était très sensible au bruit [SJ76, Lim78, Tie80]. Celui-ci dénature le spectre de fréquence et le prédicteur essaye de coller au spectre dénaturé plutôt qu'à la parole sous-jacente. Aussi, il est souhaitable de développer des techniques qui tiennent compte des effets du bruit.

Plusieurs approches ont été proposées pour améliorer la robustesse de la technique *LPC*. Tierney, dans son étude concernant la technique *LPC* en présence de bruit [Tie80], montra que, pour effectuer une meilleure modélisation spectrale, l'ordre du modèle *LPC* devait être suffisamment grand afin de modéliser à la fois la parole et le bruit. Cependant, cette méthode n'est certainement pas aussi efficace que l'élimination du bruit avant ou pendant l'analyse spectrale. Ephraim et al. estimèrent le modèle *LPC* du signal de parole non-bruité à partir d'une modélisation de la parole bruitée [EWR87]. Ayant remplacé le modèle *LPC* par leur nouveau modèle, ils obtinrent une amélioration des scores de reconnaissance équivalente à une augmentation du *SB* d'environ 10 dB. Beaucoup de travaux ont aussi été réalisés afin d'effectuer une modélisation spectrale robuste à l'aide des modèles *AR* (e.g. [MJ88b]) et *ARMA* (e.g. [Cad82]). Néammoins, ces travaux font souvent l'hypothèse que le bruit peut être modélisé par un spectre ajouté au spectre de puissance du signal, ce qui est dans beaucoup de cas une simplification trop importante.

D'autres types d'analyse spectrale ont aussi été considérés. Il est maintenant bien accepté que les systèmes à banc de filtres sont moins sensibles au bruit que les systèmes utilisant l'analyse *LPC*. En reconnaissance de parole, il a été montré que des modèles auditifs pouvaient être plus robustes que les techniques banc de filtres ou *LPC* [HL86, HL88, Ghi86, Ghi87]. Ghitza [Ghi87] montra qu'une technique à synchronisation temporelle était plus robuste qu'une technique classique d'évaluation de la puissance du spectre de fréquence. Hunt et Lefebvre, grâce à un modèle de la cochlée [HL86] utilisant les travaux de Seneff [Sen84], montrèrent que leur modèle donnait de meilleurs résultats, en présence de bruit, qu'une technique de banc de filtres utilisant une représentation paramétrique cepstre-Mel.

Tout système de reconnaissance de parole comprend un modèle de représentation paramétrique de la parole et une mesure de similitude (voir section C.3.1). Les deux modules interagissent et sont affectés par le bruit. Ainsi, une autre possibilité pour améliorer les performances des systèmes de reconnaissance en milieu bruité consiste à développer de nouvelles mesures de distance plus robustes en présence de bruit.

C.5.3.2 Les mesures de distance en milieu bruité

Les systèmes de reconnaissance de parole utilisant les mesures spectrales traditionnelles avec l'analyse *LPC* sont très sensibles au bruit. Les mesures utilisées sont si sensibles qu'une distorsion de l'enveloppe spectrale dégrade les performances. Le fait, pour une mesure de distance, d'être très sensible à l'enveloppe spectrale est utile pour l'identification du locuteur. Cette propriété est néammoins regrettable en reconnaissance, où la mesure doit seulement tenir compte du degré de similarité ou de dissimilarité des phonèmes ou mots comparés. Ainsi, des mesures spectrales appliquées à la reconnaissance de parole en environnement bruité ont besoin d'être développées.

Après une étude expérimentale de plusieurs mesures spectrales pour de la parole bruitée, Matsumoto et Imai [MI86] ont rapporté que les mesures à pondération en fréquence amélioraient la robustesse et que la mesure "log likelihood ratio" donnait les moins bons résultats pour de faibles *SB*. Ceci fut confirmé par Soong et Sondhi [SS87] qui proposèrent une nouvelle mesure à pondération en fréquence (décrite à la section C.3.3.2) fournissant de meilleurs résultats que la mesure de distorsion d'Itakura-Saito pour un faible *SB*. Dans le cadre de la vérification du locuteur en environnement bruité, Noda démontra [Nod88] les avantages d'une mesure de distance à déformation en fréquence par rapport à des mesures plus traditionnelles opérant sur une échelle de fréquence linéaire. Parallèlement, il a été observé que les mesures de distance sensibles à la pente du spectre de fréquence donnaient aussi de bons résultats en environnement bruité. Hanson et Wakita [HW86] montrèrent que la distance *RPS* était considérablement plus performante que la distance cepstrale Euclidienne. La plupart des mesures exploitent le fait que, en présence de bruit, les pics du spectre de fréquence sont applatis (particulièrement les pics d'ordre élevé) et que l'information importante est contenue dans ces pics. Par voie de conséquence, les mesures proposées sont sensibles aux pics spectraux et pondèrent davantage les parties à fort *SB* que les parties à faible *SB*.

Comme l'indiquèrent Mansour et Juang [MJ88a], il n'y a pas de raison évidente pour conserver la propriété de symétrie d'une mesure de distance si on connait auparavant que le signal de test et de référence seront perturbés différemment par le bruit. En accord avec cette remarque et après une étude sur l'effet du bruit blanc ajouté sur les vecteurs de cepstre, ils proposèrent une famille de mesures de distorsion plus robustes en présence de bruit. En particulier, ils montrèrent que du bruit blanc ajouté réduit la norme des vecteurs de cepstre et affecte de façon minimale la déviation angulaire entre deux vecteurs de cepstre. Ils proposèrent un algorithme d'égalisation optimale adaptative par "frame" qui est caractérisé par l'utilisation de la mesure de distorsion suivante :

$$d = |C_t|^{\alpha} (1 - cos\beta), \tag{29}$$

où $cos\beta$ est défini par :

$$cos\beta = \frac{\mathbf{C_t} \bullet \mathbf{C_r}}{|\mathbf{C_t}|\,|\mathbf{C_r}|}, \tag{30}$$

et C_t, C_r, $|C_t|$, $|C_r|$ sont, respectivement, les vecteurs de cepstre de test et de référence suivis de leur norme. Cette mesure de distorsion générale, appelée *mesure de distorsion cepstrale projetée*, donna les meilleurs résultats pour α=1. Un gain de 15 dB fut obtenu [MJ88a] par rapport à la distance cepstrale pondérée de Juang et al. [JRW86].

C.5.4 La parole produite dans du bruit

Les sections précédentes ont montré qu'un certain nombre de travaux ont été réalisés afin d'améliorer les performances des systèmes de reconnaissance en milieu bruité. Néammoins, afin de réaliser des systèmes robustes, il est important de comprendre les différences acoustiques et phonétiques entre de la parole normale et de la parole produite dans du bruit. Pour étudier ces variations, un environnement bruité est simulé en injectant du bruit par l'intermédiaire d'un casque à des locuteurs auquels il est demandé de parler dans ces conditions. Il a été montré que les locuteurs augmentent leur effort vocal en présence d'un bruit de fond important. Ceci est appelé l'*effet Lombard* [Lom11, LT71].

Les études liées à ce phénomène ont été en partie motivées par les efforts récents visant à intégrer des systèmes de reconnaissance dans des cockpits d'avions militaires. Dans ce cas, le système de reconnaissance doit obtenir de bonnes performances avec un pilote sous tension et un bruit ambiant important. Néanmoins, l'apprentissage est effectué habituellement lorsque le pilote est calme et le milieu ambiant non-bruité. Le bruit dans le cockpit affecte le signal acoustique de deux façons : en s'ajoutant au signal de parole à l'entrée du microphone, et en influençant la production de la parole.

Plusieurs études ont montré que la prononciation varie en milieu bruité [PBNY85, SJA88, BMM88, HY88]. Pisoni et al. montrèrent que, dans le cas de mots isolés, le bruit affecte l'intensité, la fréquence fondamentale et le spectre de fréquence [PBNY85]. Bond et al. étendirent cette étude à la parole continue [BMM88] et rapportèrent que,

1. les changements de durée, intervenant dans les segments ou les mots, sont inconsistants. Ainsi, il est difficile de les attribuer au bruit,
2. le fréquence fondamentale, le seuil fricatif, et l'énergie totale aumentent,
3. la fréquence du premier formant augmente et celle du troisième formant diminue.

Ces résultats sont en accord avec ceux de Pisoni at al. [PBNY85]. Stanton et al. trouvèrent des différences dans les indices acoustiques des phonèmes pour de la parole normale et de la parole prononcée à haute voix ou caractérisée par l'effet Lombard [SJA88]. Ils rapportèrent que, pour les voyelles et quelques-unes des semi-voyelles, il y a une migration de l'énergie vers les moyennes fréquences au détriment de l'énergie basse et haute fréquence. Pour les fricatives non-voisées et les plosives, la migration d'énergie est caractérisée par un glissement vers les hautes fréquences. Pour presque toutes les semi-voyelles, la pente spectrale basse fréquence (0–3 kHz) augmente alors que la pente spectrale haute fréquence (0–8 kHz) diminue. Enfin, la fréquence du premier formant a tendance à augmenter.

La figure 7 montre des spectrogrammes du mot "seven" produit par un locuteur masculin dans un environnement normal et en présence de bruit injecté par l'intermédiaire d'un casque (85 dB SPL).

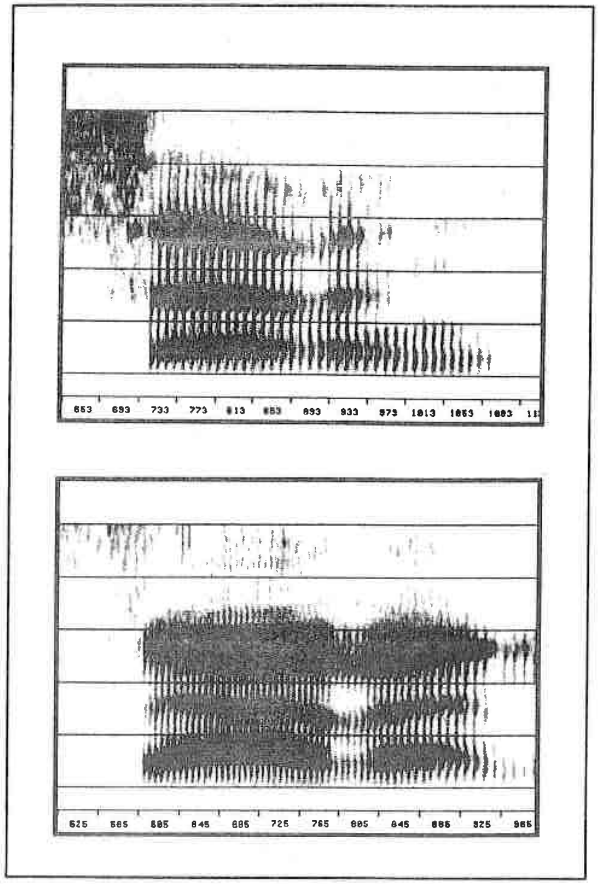


Figure 7 Spectrogrammes du mot "seven" produit par un locuteur masculin dans un environnement normal (spectrogramme du haut) et en présence de bruit injecté par l'intermédiaire d'un casque (spectrogramme du bas).

La valeur et le mouvement des formants sont quelque peu différents. En environnement bruité, la répartition d'énergie a changé et la fréquence du fondamental a augmenté. Enfin, la fin du mot est plus accentuée.

La motivation première de ces études acoustiques est d'explorer des alternatives aux systèmes actuels qui pourraient améliorer les performances pour de la parole produite sous tension et en présence de bruit. Rajasekaran et al. [RDP86], utilisant une modification de la technique *LPC* standard [RD85] et un microphone suppresseur de

bruit, rapportèrent que, en *RAP*, la variabilité due à la production de la parole dans du bruit dégradait davantage les performances que le bruit ambiant ajouté au signal de parole. Hattori et Yoshida [HY88] proposèrent une méthode de normalisation des voyelles, pour de la parole produite dans du bruit, qui améliora de 20% les scores de reconnaissance (pour des mots isolés) par rapport à une méthode traditionnelle utilisant la programmation dynamique. Cette dernière étude suggère que les performances des systèmes de reconnaissance devraient pouvoir être améliorées grâce à l'introduction de connaissances sur les changements acoustiques dus à l'effet Lombard. Actuellement c'est une voie de recherche qui reste à explorer.

Si vous voulez voir les vallées
grimpez au sommet de la montagne
Si vous désirez voir le sommet
élevez vous dans les nuages
mais si vous cherchez à comprendre
les nuages fermez vos yeux et pensez

Kahil Gibran

PARTIE D
OUTILS ET MODELES POUR LA RAP

Les différents outils et modèles utilisés dans le cadre de ce travail sont décrits. Après la présentation d'un modèle d'analyse acoustique perceptivement fondé (PLP [HHW85b]), le système hybride *ORION*, qui a été développé dans cette étude, est détaillé. Ce système est un outil de travail permettant de faciliter le test automatique d'algorithmes de reconnaissance de parole. Enfin, deux modules particuliers d'*ORION* sont exposés : *STAR*, un logiciel d'analyse et de traitement de la parole, et *SAIPH*, un système de segmentation automatique.

Chapitre D.1 PLP : UN MODELE D'ANALYSE ACOUSTIQUE PERCEPTIVEMENT FONDE

Pour voler à la vitesse de la pensée
vers tout lieu existant il te faut
commencer par être convaincu que
tu es déjà arrivé à destination ...

Extrait de "Jonathan Levingston the Seagull" Richard Bach

D.1.1 Introduction au modèle PLP

L'analyse par prédiction linéaire perceptivement fondée [HHW85b] (en anglais "perceptually-based linear prediction" ou *PLP*) est une nouvelle méthode d'analyse qui modélise un spectre auditif par un modèle tout pôle d'ordre réduit en utilisant la technique d'autocorrélation de la prédiction linéaire. Comme l'indique la figure 8, *PLP* diffère de l'analyse standard *LPC* par une intégration en bandes critiques du spectre de puissance [Fle40], une pré-accentuation par des courbes d'isosonie [RD56], et une conversion d'intensité en sonie [Ste57]. Ces différentes étapes, qui simulent des concepts psychoacoustiques bien établis, sont suivies d'une modélisation par la fonction d'un modèle tout pôle d'ordre réduit qui fournit une représentation compacte de la forme du spectre auditif en terme de pôles. L'avantage de cette fonction est de mettre l'accent, lors de la modélisation, sur les pics du spectre de fréquence.

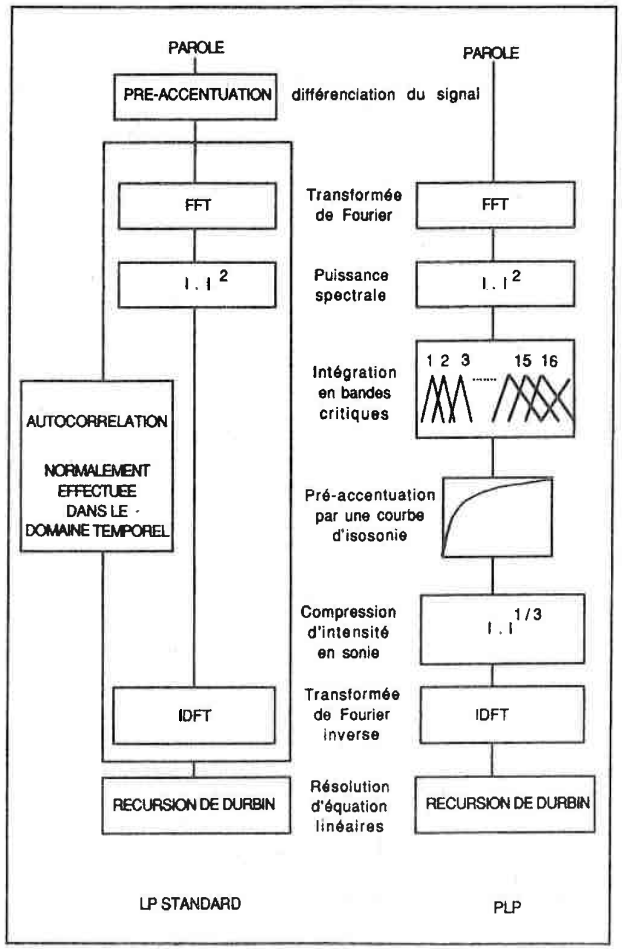


Figure 8 Diagramme fonctionnel de la technique d'analyse PLP.

La compatibilité des paramètres obtenus avec cette méthode d'analyse et la technique de prédiction linéaire est très utile au niveau des applications pratiques. Dans les prochaines sections, les différentes étapes de la méthode sont décrites.

D.1.2 Obtention du spectre auditif

Afin d'obtenir le spectre auditif un banc de 17 filtres, intégrant en bandes critiques, est utilisé. Leur fréquence centrale est uniformément espacée dans le domaine Bark défini par :

$$z = 6log\left[\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right], \tag{31}$$

où f est la fréquence en Hertz et z varie dans la bande $0 \leq f \leq 5\ kHz \quad (0 \leq z \leq 16.9\ Bark)$. La fréquence centrale du *k-ième* filtre à bande critique est z_k=*0.9994k*. Un filtre à bande critique est simulé en sommant le spectre à court terme, obtenu à l'aide d'une *FFT* d'un segment de parole multiplié par une fenêtre de Hamming, qui est ensuite pondéré. La fonction de pondération utilisée est la suivante :

$$C_k(\omega) = \begin{cases} 10^{1.0(z-z_k+0.5)} & for\ z \leq z_k - 0.5 \\ 1 & for\ z_k - 0.5 < z < z_k + 0.5 \\ 10^{-2.5(z-z_k-0.5)} & for\ z \geq z_k + 0.5 \end{cases} \tag{32}$$

Comme la valeur de sortie d'un filtre dont la fréquence centrale serait 0 Bark n'est pas bien définie (bien que nécessaire), cette valeur est prise égale à celle du filtre dont la fréquence centrale est 1 Bark [HHW85b]. Les filtres sont asymétriques : pente de 10 dB/Bark vers les basses fréquences, et pente de –25 dB/Bark vers les hautes fréquences. Ils sont similaires à ceux définis par les coubes de masquage.

La courbe d'isosonie utilisée peut être approchée par la fonction suivante :

$$E(\omega) = 1.151\sqrt{\frac{(\omega^2 + 144 \times 10^4)\,\omega^2}{(\omega^2 + 16 \times 10^4)\,(\omega^2 + 961 \times 10^4)}}\ . \tag{33}$$

où ω est la fréquence en radians. Cette courbe est une approximation de la réponse auditive pour des niveaux moyens d'intensité (+10 dB/oct de 0 – 0.4 kHz, plat de 0.4 – 1.2 kHz, +6 dB/oct de 1.2 – 3.1 kHz, plat de 3.1 – 5 kHz). F_k, qui représente la sortie du *k-ième* filtre pondérée par la fonction d'isosonie, est donnée par :

$$F_k = E(\omega_k) \int_0^{\pi} C_k(\omega)\, P(\omega)\, d\omega. \tag{34}$$

Enfin, la conversion d'intensité en sonie transforme F_k en (une compression à l'aide de la racine cubique est utilisée) :

$$Q(\omega_k) = [F_k]^{\frac{1}{3}}. \tag{35}$$

Ces différentes étapes génèrent une représentation discrète du spectre auditif (18 valeurs) qui est ensuite traitée par la fonction d'un modèle tout pôle.

D.1.3 Approximation du spectre auditif par un modèle tout pôle.

Toute fonction non-négative peut être approchée par le spectre d'un modèle tout pôle à l'aide de la technique d'autocorrélation de la prédiction linéaire. Ceci est réalisé grâce à la mise en correspondance du spectre auditif avec son équivalent dans le domaine d'autocorrélation en utilisant une *DFT* inverse. Ensuite, la résolution des équations de Yule-Walker [Mak75b] fournit un ensemble de coefficients du filtre tout pôle. Le rôle du modèle tout pôle est de réduire la dimension du spectre auditif et d'augmenter la résolution spectrale. Celle-ci est augmentée car la position des pics dans le modèle tout pôle n'est pas restreinte à la fréquence centrale des filtres à bandes critiques. De plus, un modèle d'ordre réduit permet d'éliminer les pics parasites du spectre de fréquence.

D.1.4 Applications de l'analyse PLP

De précédentes études ont montrées [HHW85b] qu'un modèle d'ordre réduit (ordre cinq) de l'analyse *PLP* est en accord avec le concept de Carlson et Fant *F1, F2'* [CFG75, Bla83] et le concept de Chistovich d'intégration auditive de 3.5 Barks [CSL78]. Ceci montre que, en intégrant des connaissances sur le fonctionnement du système auditif avec des techniques standards de modélisation spectrale, des résultats déduits de tests psychoacoustiques peuvent être obtenus. A cause du faible nombre de paramètres de sortie et de sa compatibilité avec l'analyse par prédiction linéaire, cette analyse perceptuelle est très intéressante pour des applications pratiques de reconnaissance de parole.

Comme l'indique la figure 9, où le spectre de puissance et le spectre auditif sont comparés pour le mot "nine" prononcé par un locuteur féminin, le spectre dérivé de l'analyse *PLP* est plus lisse, en temps et en fréquence, que celui obtenu à l'aide de la technique *LP*. Cette propriété, de l'analyse *PLP*, est une caractéristique souhaitée pour un modèle d'analyse acoustique. Hermansky [Her87] montra, dans le cadre d'une petite base de données, qu'un modèle d'ordre cinq de l'analyse *PLP* donnait des performances supérieures à celles fournies par un modèle dérivé de l'analyse *LP*, dans le cadre de la reconnaissance de mots isolés.

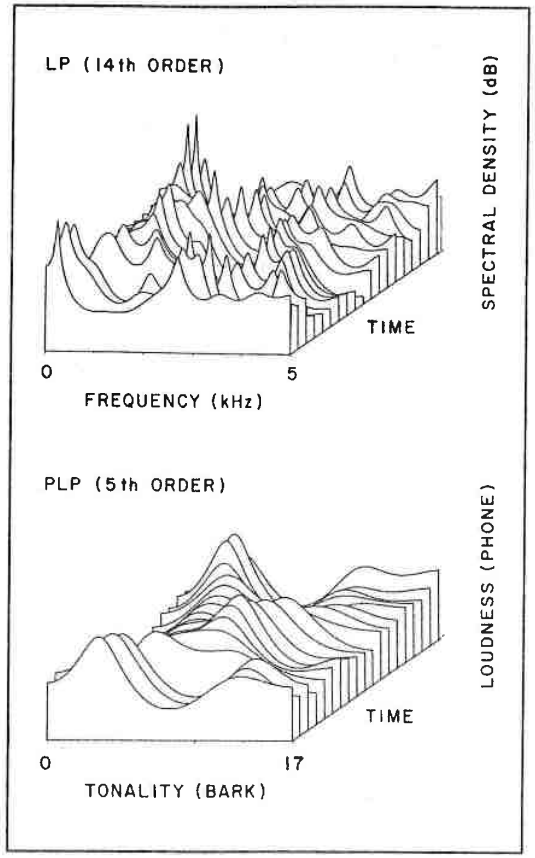


Figure 9 Spectre de puissance et spectre auditif du mot "nine".

Au niveau temps de calcul, les techniques d'analyse *LP* et *PLP* sont à peu près équivalentes (approximativement 3500 multiplications par "frame", voir Table 6). Cependant, le faible nombre de coefficients rendus par l'analyse *PLP* permet de gagner en espace mémoire et en temps de calcul dans les traitements qui suivent. Tout au long de ce travail l'analyse *PLP* a été beaucoup utilisée. En particulier, certains avantages et inconvénients ont été mis à jour. Ceux-ci seront explicités au cours de ce document et résumés dans la conclusion.

LP 14		PLP 5	Normale	Simplifiée
pré-accentuation	200	fenêtre	200	200
fenêtre	200	FFT	2100	2100
autocorrélation	2800	intégration en bandes critiques	2200	450 (intégration à bande limitée)
durbin	200	racine cubique	150 (table de correspondance + interpolation)	150 (table de correspondance + interpolation)
cepstre	200	DFT inverse	30	30
		durbin	30	30
		cepstre	30	30
TOTAL	3600	TOTAL	4700	3000

Table 6 Coût approximatif, en nombre de multiplications, de deux analyses couramment utilisées : LP d'ordre quatorze et PLP d'ordre cinq.

Chapitre D.2 ORION : UN SYSTEME (ET UN OUTIL) POUR LA RECONNAISSANCE AUTOMATIQUE DE MOTS ISOLES

A- Qu'est ce que tu dis de cette nuit ?
B- J'ai jamais vu la même.
A- Moi non plus, dit l'homme. ORION ressemble à une fleur de carotte.
B- Pardon ? demanda Jourdan.
A- ORION est deux fois plus grand que d'habitude. C'est ça, là haut. Arrête-toi. Là haut. Là. Tu vois ?
B- Non.
A- J'ai jamais su rien désigner, dit l'homme. C'est curieux, ça. On m'en a toujours fait le reproche. On m'a dit : <<On ne voit jamais ce que vous voulez dire.>>

Extrait de "Que ma joie demeure" Jean Giono

D.2.1 Introduction

Deux grandes approches ont été proposées pour construire des systèmes de reconnaissance automatique de la parole. La première est fondée sur une représentation paramétrique de la parole qui utilise des mesures faites sur le signal de parole. Cette phase est en général suivie d'un algorithme de reconnaissance des formes (du type de la programmation dynamique) afin d'effectuer la classification. Pour déterminer la représentation paramétrique, les modèles auditifs ont récemment reçu beaucoup d'attention. Bien que notre connaissance des phénomènes liés à la perception de la parole soit très limitée, il a été montré qu'un système qui modélise des propriétés du système auditif humain pouvait générer une meilleure représentation de la parole que les systèmes traditionnels [Her87, Jun87, Lyo82]. La deuxième approche utilise un langage de représentation pour décrire les unités acoustiques [Car86, ZL86]. Une telle approche utilise intensivement les techniques d'intelligence artificielle. Dans ce cas, le principal problème est l'acquisition et la représentation de la connaissance.

La reconnaissance fondée sur l'extraction d'indices acoustiques a été proposée comme une alternative à la reconnaissance de formes [Col83]. Cependant, comme notre connaissance des mécanismes liés à la parole est très limitée, les systèmes de reconnaissance

ne doivent pas "oublier" de prendre en compte notre *ignorance* [MS85]. Un exemple de modèle "d'ignorance" est fourni par l'algorithme de programmation dynamique qui modèle notre ignorance à propos des variations de la parole dans l'espace temps. L'introduction de connaissances sur la parole dans les systèmes de *RAP* est un bon moyen d'améliorer les performances des systèmes actuels. Toutefois, ces connaissances doivent être introduites avec précaution. Dans le même ordre d'idée, il a été montré que des systèmes hybrides, utilisant des modèles mathématiques et des connaissances sur la parole, permettaient d'améliorer les scores de reconnaissance et aidaient à triompher des limitations associées aux systèmes traditionnels [BCL82].

L'approche qui est proposée est fondée sur l'utilisation de plusieurs sources de connaissances dans un système hybride à deux passes, appelé *ORION*, qui utilise des connaissances phonétiques pendant la deuxième passe. Les connaissances phonétiques facilitent la discrimination des mots appartenant à des classes du vocabulaire constituées de mots acoustiquement similaires. *ORION* utilise un modèle d'analyse acoustique qui simule des propriétés du système auditif humain et accentue l'importance donnée aux transitions. De plus, il bénéficie de modèles perceptuels à la fois dans la partie analyse acoustique et dans l'extraction d'indices.

Le développement de ce système est né de deux idées. La première est venue de notre incapacité, avec les systèmes de comparaison de formes, à améliorer les scores de reconnaissance pour des vocabulaires difficiles. Même en donnant plus d'importance aux transitions, pour un vocabulaire comme le "*E-SET*={B, C, D, E. G, P, T, V, Z, FEED}", notre meilleur système (utilisant un algorithme de programmation dynamique) obtint 68% de mots reconnus (en reconnaissance multilocuteur avec neuf références par mot). Ceci dirigea nos recherches vers le développement d'un système pouvant traiter de façon particulière les vocabulaires difficiles sans perdre les avantages d'un algorithme de comparaison de formes pour les autres mots du vocabulaire. La deuxième idée est liée à l'évolution rapide des systèmes de reconnaissance. De nouvelles méthodes d'analyse, mesures de distance, indices, sont souvent testés et comparés. Aussi, un outil de recherche, permettant une grande flexibilité sans perdre les avantages d'un système de reconnaissance intégré, nous apparut nécessaire. Ces considérations conduirent au développement du système *ORION* qui est décrit dans la section suivante.

D.2.2 Vue générale du système

La figure 10 montre un diagramme du système réalisé.

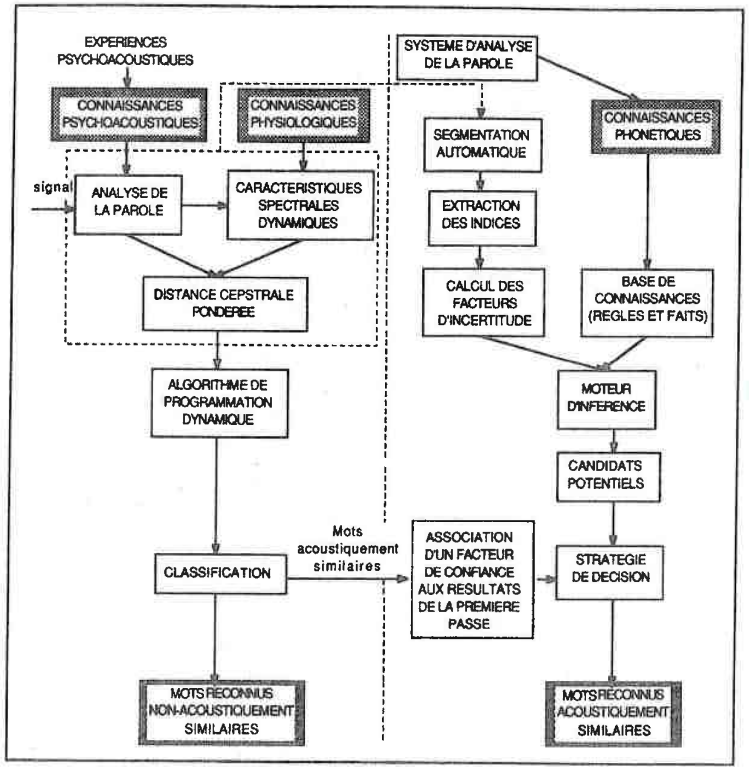


Figure 10 Diagramme fonctionnel du système hybride ORION.

Pendant la première passe, un algorithme de programmation dynamique utilise le modèle d'analyse acoustique *PLP* pour générer des coefficients cepstraux du spectre auditif. Ceux-ci sont ensuite combinés avec d'autres coefficients, modélisant des caractéristiques spectrales dynamiques du signal d'entrée, pour former un vecteur représentant un segment de parole. L'ensemble des vecteurs, représentant un mot, est ensuite traité par un algorithme de programmation dynamique qui utilise une distance cepstrale pondérée. La première passe fournit le mot reconnu pour les mots du vocabulaire qui n'appartiennent pas à une classe de mots confondus facilement. Les classes de mots confondus facilement (ou difficiles à reconnaître) ont été identifiées à l'aide de matrices

de confusion fournies par des tests préliminaires. Une deuxième passe est invoquée si les premiers candidats fournis par l'algorithme de programmation dynamique appartiennent à une classe de mots du vocabulaire étudié définie comme difficile à reconnaître. Dans cette étude, les efforts ne se sont pas portés sur l'algorithme de classification. Cependant, des tests préliminaires ont montré que, pour le vocabulaire étudié, la méthode choisie suffisait à résoudre le problème posé par la classification. Des indices temporels et fréquentiels sont alors extraits automatiquement avant d'être fournis à un moteur d'inférence. Grâce à une base de connaissances construite au préalable, le moteur d'inférence génère des candidats pondérés par un facteur de confiance. Une stratégie de décision, prenant en compte les résultats donnés par le moteur d'inférence et ceux générés par l'algorithme de programmation dynamique, est alors chargée de fournir la décision finale. Les candidats générés par l'algorithme de programmation dynamique sont aussi pondérés par un facteur de confiance obtenu à l'aide de matrices de confusion.

Avant d'extraire les indices temporels et fréquenciels, un algorithme de segmentation automatique est appliqué afin de déterminer les différents phonèmes du mot analysé. Ceci permet d'extraire la plupart des indices dans la partie du mot qui permet de le discriminer avec les autres mots de la même classe.

Le système a été développé de façon modulaire afin de permettre une souplesse importante pour tester de nouveaux algorithmes. Notre connaissance (phonétique, du système auditif, etc) n'est pas statique, pas plus que ne l'est notre compréhension des techniques à introduire dans les systèmes de reconnaissance. Aussi, ces systèmes doivent être aisément modifiables pour pouvoir intégrer facilement les changements liés à l'évolution de nos connaissances. Le système développé est organisé autour d'un ensemble de modules. Chaque module est bien séparé des autres et peut être facilement modifié ou remplacé. Ceci permet de se concentrer sur une partie du système sans avoir à redévelopper un nouveau système chaque fois qu'une modification doit être faite. De plus, lors des tests, le système complet ou une partie du système seulement peut être évalué. Un tel système est un outil de recherche qui permet une importante souplesse. Il s'est avéré très utile dans les études réalisées.

Chapitre D.3 STAR : UN LOGICIEL D'ANALYSE ET DE TRAITEMENT DE LA PAROLE

Les gens ont des étoiles qui ne sont pas les mêmes.
Pour les uns, qui voyagent, les étoiles sont des guides.
Pour d'autres elles ne sont rien que de petites lumières.
Pour d'autres, qui sont savants, elles sont des problèmes.
Mais toutes ces étoiles là se taisent. Toi, tu auras des
étoiles comme personne n'en a...

Extrait de "Le petit prince" Antoine de Saint-Exupéry

Afin de faciliter l'extraction d'indices acoustiques et une évaluation interactive de certains algorithmes, un logiciel d'analyse et de traitement de la parole a été développé. Son but est de faciliter "l'édition" interactive de spectrogrammes numériques mais aussi d'autres représentations de la parole (spectrogrammes auditifs, spectrogrammes obtenus à partir de l'analyse *LP*, spectrogrammes obtenus à partir de l'analyse *PLP*, etc). Plusieurs représentations ont été sélectionnées afin d'étudier leur utilité à fournir des indices distinctifs.

Quelques unes des fonctions de base disponibles sont les suivantes :

1. sélection d'une partie du spectrogramme (suivant l'axe temps ou l'axe fréquence),
2. visualisation du spectre d'une "frame",
3. visualisation de l'énergie dans une bande de fréquence (prédéterminée ou déterminée manuellement à l'aide de curseurs),
4. extraction manuelle des mouvements des formants,
5. acquisition et restitution de la parole,
6. affichage de mesures en des points sélectionnés interactivement (valeur des formants, etc),
7. utilisation des couleurs,
8. etc.

Ces fonctions peuvent être appliquées à chacune des représentations possibles. Le logiciel a été conçu très ouvert afin de faciliter l'addition de nouvelles fonctions. L'interface utilisateur se présente sous la forme de menus sélectionnés à l'aide de la souris. Enfin, plusieurs fenêtres (utilisant éventuellement des représentations différentes) sont disponibles en même temps.

En complément des fonctions déjà citées, des traitements de plus haut niveau sont accessibles,

1. segmentation automatique du signal de parole,
2. visualisation des marques de segmentation déterminées automatiquement ou manuellement,
3. extraction d'indices temporels ou fréquenciels,
4. classification grossière des segments déterminés automatiquement à l'aide de l'algorithme de segmentation.

Au niveau de chacune des représentations, la correspondance entre les numéros de "frame" du spectrogramme (ou pseudo-spectrogramme) et du signal temporel est disponible. Ceci permet la prise en compte d'outils existants (comme *ILS*) qui peuvent être exécutés dans d'autres fenêtres. Ainsi, de nouvelles fonctions comme : visualisation du signal temporel, détection du fondamental, etc, peuvent aussi être utilisées. Le but visé était de développer un logiciel facile d'emploi venant en complément des logiciels de traitement de signal disponibles sur le marché. *Ce n'était pas de réaliser un autre logiciel de traitement de signal.*

Ce logiciel a été développé dans le même esprit que le système *ORION*, c'est-à-dire comme un outil de recherche facilement utilisable et extensible. Lors de sa phase de conception, l'accent a été mis sur sa souplesse plus que sur le nombre de fonctions disponibles immédiatement. Aussi, ce système a été constitué de façon incrémentale en ajoutant des fonctions lorsque celles-ci s'avéraient nécessaires. Comme l'utilisation d'un logiciel de traitement de signal est différent d'un chercheur à l'autre les fonctions manipulées ne sont pas les mêmes. Par conséquent, l'extensibilité est une caractéristique importante de ce type de logiciel. Ce système a été développé en language C sur une station de travail SUN qui utilisait des ressources distantes (accessible par réseau local) comme un processeur vectoriel et des périphériques d'acquisition et de restitution de la parole.

Chapitre D.4 SAIPH : UN SYSTEME DE SEGMENTATION AUTOMATIQUE

Oh oui quelle chance avons nous toi et moi d'habiter l'intemporel : nous qui flânant sommes descendus des odorantes montagnes d'éternel à présent pour folâtrer parmi des mystères tels que naître et mourir un jour (ou même un peu moins peut-être).

E. E. Cummings

D.4.1 Introduction

Lorsque la reconnaissance de parole traite de vocabulaires difficiles, il est important que le système de reconnaissance s'intéresse particulièrement aux parties des mots qui sont importantes pour la discrimination. Un système fondé sur l'extraction d'indices a souvent besoin de connaître les frontières de ces parties discriminantes. Ceci nous amena à développer un système de segmentation automatique utilisé dans la deuxième passe de *ORION*. Ce système, appelé *SAIPH*, est fondé sur l'utilisation de la technique d'analyse *PLP* et de caractéristiques spectrales dynamiques du signal de parole. Le spectre de fréquence obtenu à l'aide de l'analyse *PLP* a été représenté par M coefficients cepstraux (M est l'ordre du modèle tout pôle). L'analyse PLP fournit un spectre plus lissé que celui rendu par l'analyse *LP* [HHW85b, Her87]. Ceci est une caractéristique souhaitable pour utiliser des paramètres spectraux dynamiques. Afin d'obtenir un spectre de fréquence assez détaillé, mais aussi suffisamment lissé, un ordre du modèle égal à huit a été sélectionné.

Ce système a été utilisé dans *ORION* mais aussi en tant que composante d'un système d'étiquetage automatique de la parole (voir partie F).

D.4.2 Une mesure de transition

Afin de segmenter un mot en unités élémentaires, une mesure de transition, pouvant préserver les changements intervenant dans le signal de parole, doit être définie. Une façon de représenter les transitions est de modéliser les caractéristiques dynamiques du spectre de fréquence. *SAIPH* utilise des coefficients de régression [Fur86a] dérivés des coefficients cepstraux fournis par l'analyse *PLP*. Une mesure de transition (pour la "frame" k) a été définie par :

$$TM(k) = \sum_{i=1}^{M} \frac{(i \times r_i^k)^2}{M} \qquad (36)$$

où M est l'ordre du modèle d'analyse et r_i le i-ième coefficient de régression.

Dans l'équation (36), chaque coefficient de régression est multiplié par son index avant de l'élever au carré. Comme le montre la figure 11, où l'inverse de l'écart type des coefficients de régression est représenté, cette multiplication est une bonne approximation de la normalisation des variations de chaque coefficient de régression.

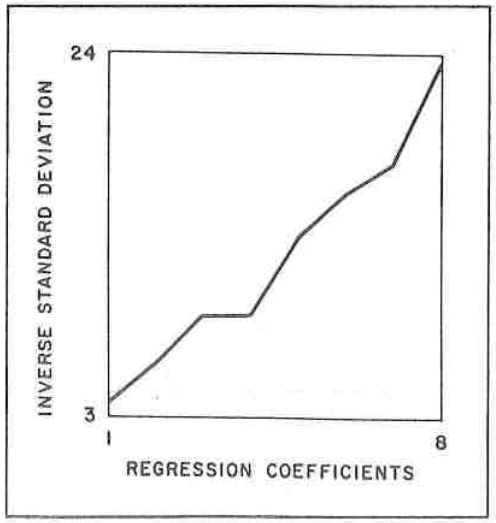


Figure 11 Inverse de l'écart type des coefficients de regression obtenus à partir de l'analyse PLP pour un vocabulaire alphanumérique.

L'équation (36) peut aussi être interprétée comme le calcul d'une mesure de transition sur les coefficients cepstraux pondérés *RPS* [Pal82]. Comme cela a été indiqué dans la

section C.3.2.4, ces coefficients ont été utilisés avec succès en *RAP* [HW86, Her87, Pal82]. Une de leurs caractéristiques est d'être très sensible aux brusques variations de pente survenant autour des pics du spectre de fréquence.

La fenêtre optimale pour calculer les coefficients de régression a été trouvée expérimentalement égale à 70 ms (avec une taille de "frame" de 10 ms). Enfin, le calcul de cette mesure de transition a été suivi d'un lissage par une moyenne flottante pour éliminer les pics parasites.

D.4.3 Segmentation automatique

Tout d'abord un algorithme, permettant de générer au plus deux candidats pour les débuts et fins de de mots, est appliqué. Cet algorithme utilise l'énergie du signal temporel et le taux de passages par zéro comme principaux paramètres. Puis, des marques de segmentation sont déterminées en utilisant la mesure de transition, le logarithme de l'energie normalisée du signal temporel et les fontières de mots. Durant cette étape, seulement les marques de segmentation correspondant aux pics dont l'amplitude est supérieure à un seuil défini expérimentalement (*SL*) et aux plateaux de la courbe de transition sont conservées. Un plateau est défini comme une portion de la courbe de transition qui reste dans un certain intervalle d'amplitude. Les plateaux sont utiles pour identifier les transitions lentes occasionnées par les liquides et les semi-consonnes (r, l, w, y). Les pics dont l'amplitude est inférieure à *SL* représentent des transitions incertaines qui pourront être éventuellement considérées lors de traitements complémentaires. La valeur du seuil a été choisie afin d'encourager la sur-segmentation et non la sous-segmentation.

Un ensemble de règles de production (connaissances heuristiques) est ensuite appliqué pour combiner les différents paramètres retenus. Les règles sont exprimées comme l'indique l'exemple suivant :

si ((LOC[0] -ST1) >th1)
et ((VAL[LOC[0]] >th2) ou (LRMS[LOC[0]] >th3))
et (LRMS[ST1] >th4)
alors ajouter une marque de segmentation à la position ST1

où *LOC[0]* est la première marque de segmentation trouvée à l'aide de la mesure de transition, *VAL[LOC[0]]* est la valeur donnée par la mesure de transition pour la première marque de segmentation, *LRMS* est le logarithme de l'énergie du signal temporel, *ST1* est le début de mot trouvé par l'algorithme qui détecte les frontières de mots et *th1*, *th2*, *th3*, *th4* sont quatre seuils définis expérimentalement.

Dans *SAIPH*, environ 10 règles sont utilisées. Le principal objectif de ces règles est de rajouter des marques de segmentation quelquefois mal détectées par la mesure de transition (e.g. périodes de silence avant les plosives), de raffiner les frontières de mots, qui sont calculées approximativement à l'aide de l'énergie et du taux de passages par zéro,

et d'éliminer les clics et le bruit à la fin des mots. La figure 12 montre les paramètres utilisés lors de l'étape de base du processus de segmentation. L'algorithme qui vient d'être décrit est appelé segmentation de base car des traitements supplémentaires peuvent être appliqués afin d'améliorer les résultats de la segmentation (voir section F.1.4.2).

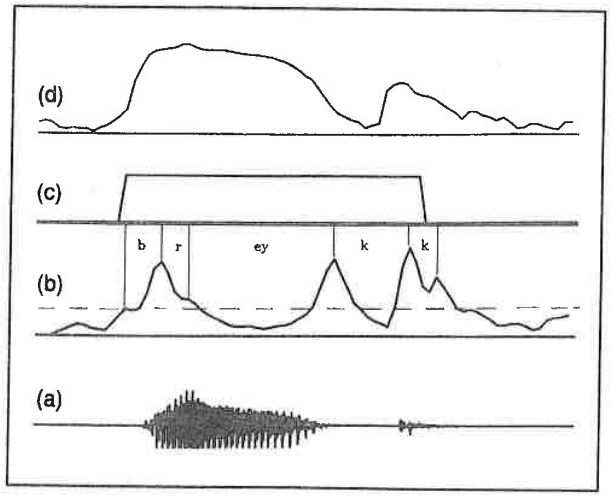


Figure 12 Signal temporel (a) et paramètres utilisés dans l'algorithme de segmentation (mesure de transition (b), frontières de mots (c) et énergie (d)).

D.4.4 Evaluation de la segmentation automatique

Les erreurs de segmentation sont difficiles à définir avec précision car elles dépendent du choix de la segmentation de référence à laquelle les résultats sont comparés. Dans l'évaluation présentée, les résultats ont été comparés à ceux obtenus par une segmentation manuelle réalisée au préalable. L'évaluation a été faite en terme de segments corrects, segments manquants (sous-segmentation), segments parasites (sur-segmentation) et écarts par rapport aux segments de référence. Chaque marque de segmentation obtenue manuellement ou automatiquement a été affichée sur le spectrogramme du mot correspondant et a été vérifiée. Dans cette évaluation, la segmentation manuelle fut aussi contrôlée.

La segmentation automatique fut évaluée sur une base de données de 104 mots, appelée "keyboard" (voir section E.1.1). Un locuteur masculin et un locuteur féminin furent considérés. Ces données de test représentent plus de 1000 segments et environ 5 minutes de parole. Les résultats, moyennés pour les deux locuteurs, sont présentés à la table 7.

segments corrects (I)	segments avec un écart > 20 ms	sur-segmentation > seuil SL (II)
93.1%	4.1% (de I)	21.8%

sur-segmentation > seuil SL dans les plosives, les fricatives, et les fins de mots	sur-segmentation < seuil SL
64.7% (de II)	43.3%

segments manquants (III)	segments manquants pour l, r, w, les plosives et les nasales	segments manquants pour lesquels aucune transition n'a été détectée
6.9%	83.0% (de III)	2.1%

erreurs de la segmentation manuelle (IV)	segments non-présents qui ont été insérés par la segmentation manuelle
6.1%	31.3% (de IV)

Table 7 Résultats moyennés (pour un locuteur masculin et un locuteur féminin) de l'évaluation de la segmentation automatique (les chiffres Romains près des résultats indiquent les colonnes, identifiées par les mêmes chiffres, auquelles les pourcentages correspondent). Lorsqu'aucun chiffre Romain n'est indiqué près du pourcentage cela signifie que celui-ci se réfère au nombre total de segments possibles.

La sur-segmentation, qui est obtenue lorsque la valeur de la mesure de transition est supérieure au seuil *SL*, représente non seulement des marques de segmentation parasites, mais aussi des marques correctes ayant une mesure de transition faible. Environ 93% des segments sont identifiés correctement. Enfin, il est intéressant de noter que 6% des segments obtenus avec la segmentation manuelle ne sont pas corrects.

Afin de valider le choix de l'analyse *PLP*, *SAIPH* fut comparé à un système similaire utilisant l'analyse *LP* (avec toujours un modèle d'ordre huit) à la place de l'analyse *PLP*. Les résultats obtenus sont présentés à la table 8.

Résultats obtenus avec l'analyse PLP		
segments corrects	sur-segmentation > seuil SL	sur-segmentation < seuil SL
93.8%	16.8%	37.1%

Résultats obtenus avec l'analyse LP		
segments corrects	sur-segmentation > seuil SL	sur-segmentation < seuil SL
92.3%	29.6%	54.3%

Table 8 Comparaison des résultats de segmentation obtenus avec l'analyse PLP puis avec l'analyse LP pour un locuteur masculin (avec une taille de "frame" de 10 ms). Ces pourcentages se réfèrent au nombre total de segments possibles.

Le pourcentage de segments corrects est similaire à celui obtenu avec l'analyse *PLP* mais la sur-segmentation est beaucoup plus importante.

Enfin, le système fut testé en utilisant une taille de "frame" différente (6 ms à la place de 10 ms). La table 9 montre les résultats obtenus.

Résultats obtenus avec l'analyse PLP		
segments corrects	sur-segmentation > seuil SL	sur-segmentation < seuil SL
88.2%	48.5%	107.6%

Table 9 Résultats de segmentation obtenus avec une taille de "frame" de 6 ms pour un locuteur masculin. Ces pourcentages se réfèrent au nombre total de segments possibles.

La sur-segmentation est aussi beaucoup plus importante. Le nombre de marques de segmentation correctes est légèrement plus faible. Ceci est probablement dû à la valeur du seuil *SL* qui, dans ce cas, aurait dû être abaissé. Une taille de "frame" de 10 ms, choisie initiallement, est un bon compromis entre la capture des informations qui varient rapidement et trop de sur-segmentation.

En conclusion, les principaux avantages de l'algorithme proposé, comparé à d'autres systèmes de segmentation automatique présentés dans la littérature (voir section F.1.2), sont les suivants :

1. le spectre lissé rendu par l'analyse *PLP* limite les marques de segmentation parasites,
2. une nouvelle mesure de transition modèle de façon adéquate les changements spectraux,
3. les marques de segmentation sont générées à l'aide d'une mesure de transition qui ne dépend pas du locuteur.

It is a law of human life, as certain as gravity : to live fully, we must learn to use things and love people ... not love things and use people

John Powell

PARTIE E
ETUDES ET REALISATIONS

Grâce à une évaluation comparative de plusieurs modèles d'analyse acoustique, il est montré qu'un modèle utilisant la technique *PLP* donne les meilleurs résultats. Une optimisation de ce modèle est ensuite présentée. En particulier, il est montré qu'une nouvelle mesure de distance qui pondère chaque coefficient cepstral par une puissance de son index améliore quelques insuffisances de la distance *RPS* et que des caractéristiques spectrales dynamiques sont particulièrement utiles avec l'analyse *PLP* pour traiter des vocabulaires de mots acoustiquement similaires.

Les études menées en *RAP* dans le cadre de parole bruitée sont ensuite décrites. En particulier, une évaluation de l'analyse *PLP* en environnement bruité, le développement d'un modèle auditif (*SLP*) fondé sur des concepts physiologiques et l'étude de "lifters" cepstraux et mesures de distance en présence de bruit sont rapportés.

Enfin, il est montré, grâce à un système hybride, que l'utilisation de connaissances phonétiques pour traiter des mots acoustiquement similaires améliore les scores de reconnaissance.

Chapitre E.1 ETUDE DU MODELE D'ANALYSE ACOUSTIQUE PLP EN RAP

Just because the message may
never be received does not
mean it is not worth sending.

Segaki

E.1.1 Conditions expérimentales

Dans le cadre de ce travail (et pas seulement pour les tests présentés dans ce chapitre) les bases de données suivantes ont été utilisées :

base de données D1 : 104 mots habituellement employés pour définir les touches d'un clavier. Les mots ont été produits dans un environnement non-bruité. 10 locuteurs (6 hommes et 4 femmes) et 3 répétitions ont été considérés. Cette base de données est appelée "keyboard".

base de données D2 : un sous-ensemble de la base de données *D1* constitué du vocabulaire alphanumérique.

base de données D3 : un autre sous-ensemble de la base de données *D1* constitué de classes de mots acoustiquement similaires {B, D, G, P, T, V, Z, C, E, FEED, F, S, X, LINE, NINE, ONE}.

base de données D4 : vocabulaire des chiffres produits en environnement non-bruité par des locuteurs de dialectes différents. 96 locuteurs et une répétition (48 hommes et 48 femmes) ont été considérés.

base de données D5 : vocabulaire de 49 mots (alphabet et chiffres plus des mots de contrôle) enregistrés en présence de bruit injecté par l'intermédiaire d'un casque (bruit blanc à 85 dB SPL). Deux répétitions produites par 10 locuteurs (5 hommes et 5 femmes) ont été utilisées.

Tous les mots ont été enregistrés isolément (à une fréquence d'acquisition de 10 kHz) et les frontières de mots ont été déterminées manuellement. Aucune technique de classification ou procédure spéciale de sélection des mots de référence n'a été utilisée. Pour les tests effectués en présence de bruit additif, le *SB* a été défini comme le rapport entre la puissance du spectre de fréquence du signal de parole (moyennée sur tout le mot) sur la puissance du spectre de fréquence du bruit.

E.1.2 Préliminaires

La technique d'analyse *PLP*, décrite dans la section D.1.1, est constituée des traitements suivants :

1. intégration en bandes critiques du spectre de parole à court terme,
2. pré-accentuation par une courbe d'isosonie,
3. conversion d'intensité en sonie,
4. application d'un modèle tout pôle sur l'enveloppe du spectre auditif.

Dans toutes les études présentées, 17 filtres à bandes critiques ont été utilisés sur une gamme de fréquence couvrant 17 Barks. Ces filtres ont été simulés en intégrant le spectre de fréquence produit par une *FFT* du signal de parole multiplié par une fenêtre de Hamming. La taille des "frames" de parole a été fixée à 10 ms. Le spectre obtenu à l'aide de l'analyse *PLP* a été approché par M coefficients cepstraux déduits d'un modèle tout pôle d'ordre M. Dans tous les tests, un algorithme de programmation dynamique, utilisant une contrainte locale symétrique et une contrainte de pente d'ordre 2 [MRR80], a permis d'effectuer les comparaisons entre les mots de test et les mots de référence.

Le but de ce travail est l'amélioration des systèmes de reconnaissance multilocuteur de mots isolés. En particulier, nous nous sommes intéressés à l'adéquation entre le modèle d'analyse acoustique et la mesure de distance. Généralement, en reconnaissance multilocuteur, une technique de classification est utilisée pour sélectionner les mots de référence. Dans les études présentées, une approche différente fut suivie. Des tests de reconnaissance monolocuteur, interlocuteur (où les mots produits par un locuteur sont comparés à ceux produits par un locuteur différent), et multilocuteur (avec un petit nombre de références par mot, sélectionnées sans utiliser aucune procédure particulière) ont été réalisés. L'idée se trouvant derrière les tests interlocuteur et multilocuteur (avec un faible nombre de mots de référence) est l'amplification des différences possibles de performance entre les modèles d'analyse acoustique. En particulier, le lissage provoqué par l'utilisation d'un algorithme de classification ou par l'utilisation d'un grand nombre de références par mot a voulu être évité. Dans notre cas, l'intérêt se trouve plus au niveau de l'étude de performances comparatives qu'au niveau des scores de reconnaissance obtenus. Des travaux précédents ont montré, expérimentalement, que les propriétés d'un modèle d'analyse acoustique en reconnaissance monolocuteur et interlocuteur étaient intégrées (d'une manière complexe) en reconnaissance multilocuteur [Her87]. Nous avons donc cherché à optimiser le modèle étudié en reconnaissance monolocuteur et interlocuteur afin de déterminer les directions vers lesquelles le modèle devait évoluer pour améliorer les performances obtenues en reconnaissance multilocuteur.

E.1.3 Etude comparative de plusieurs modèles d'analyse acoustique

E.1.3.1 Le modèle d'analyse acoustique "critical-band slope metric"

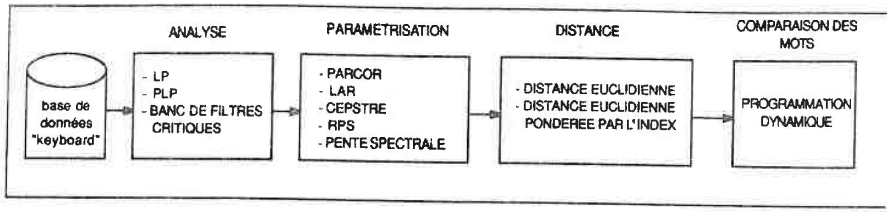
Dans l'étude suivante, un des modèles d'analyse acoustique considéré est le "critical-band slope metric" (*CB_SM*) qui est similaire à celui proposé par Klatt [Kla82]. Le banc de filtres critiques est celui utilisé lors de la première étape de l'analyse *PLP*. Une pré-accentuation par une courbe d'isosonie est réalisée pendant le filtrage. Cette étape est suivie d'une approximation de l'échelle phone de la sonie grâce à l'application de la fonction logarithme sur les sorties des filtres. Enfin, la pente spectrale à la fréquence centrale de chaque filtre *i*, *SL[i]*, est calculée en prenant la différence suivante (voir [Kla82]) :

$$SL[i] = dB[i+1] - dB[i]. \qquad (37)$$

où *dB[i]* représente la sortie en décibels du *i-ième* canal. 17 coefficients sont fournis à un algorithme de programmation dynamique qui utilise une distance Euclidienne.

E.1.3.2 Effet de l'ordre du modèle

Des études préliminaires [Her87] ont suggéré, à partir d'une étude comparative, que les meilleures performances (pour les modèles étudiés), en *RAP* multilocuteur de mots isolés, étaient obtenues à l'aide d'un modèle d'ordre réduit utilisant l'analyse *PLP*. Dans le but de valider cette hypothèse (avec une base de données plus importante) et de tester différentes représentations paramétriques de la parole avec les techniques *LP* et *PLP* (PARCOR, "log area ratio", etc), une série de tests a été menée en reconnaissance monolocuteur et multilocuteur. Dans ces tests l'ordre du modèle d'analyse a varié de cinq à quatorze. Neuf modèles d'analyse acoustique ont été comparés en reconnaissance monolocuteur et cinq en reconnaissance multilocuteur. La base de données *D1* ("keyboard") a été utilisée. Trois types d'analyse : *LP*, *PLP* et banc de filtres critiques, et deux types de distance : Euclidienne (avec le cas particulier de cepstrale Euclidienne) et Euclidienne sur des coefficients pondérés par leur index (avec le cas particulier de *RPS*) ont été considérées. La figure 13 résume les modèles étudiés.



Analyse	Paramétrisation	Distance	Notation
LP	cepstrale	Euclidienne	LP_CEPS
LP	cepstrale	Euclidienne pondérée par l'index	LP_RPS
PLP	cepstrale	Euclidienne	PLP_CEPS
PLP	cepstrale	Euclidienne pondérée par l'index	PLP_RPS
PLP	PARCOR	Euclidienne	PLP_PAR
PLP	PARCOR	Euclidienne pondérée par l'index	PLP_PAR.IW
PLP	"log-area ratio"	Euclidienne	PLP_LAR
PLP	"log-area ratio"	Euclidienne pondérée par l'index	PLP_LAR.IW
filtrage en bandes critiques	pente spectrale	Euclidienne	CB_SM

Figure 13 Evaluation monolocuteur

Le modèle *CB_SM* est celui décrit dans la section E.1.3.1. Dans tous les tests, *M* coefficients (*M* représente l'ordre d'analyse) ont été considérés. Deux répétitions de chaque mot ont été utilisées, ce qui a induit plus de 2000 comparaisons pour le cas monolocuteur et plus de 15000 comparaisons pour le cas multilocuteur. Dans les tests multilocuteur, chaque essai a utilisé les vocabulaires de deux locuteurs (un homme et une femme) comme ensemble de référence et les vocabulaires des autres locuteurs comme ensemble de test. La figure 14 montre les résultats, moyennés pour tous les locuteurs, des tests monolocuteur de cinq des modèles d'analyse acoustique (tous les résultats n'ont

pas été présentés pour des raisons de clarté) qui ont fourni les meilleurs résultats. Ce sont aussi les cinq modèles qui ont été testés en reconnaissance multilocuteur.

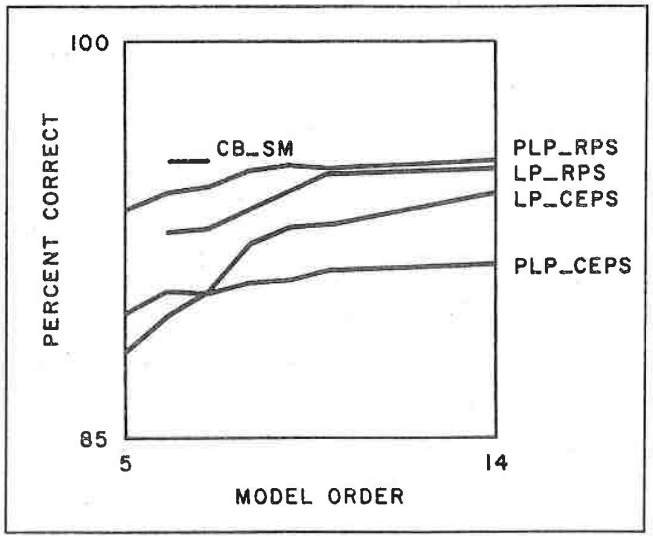


Figure 14 Comparaison de plusieurs modèles d'analyse acoustique en reconnaissance monolocuteur.

D'après ces résultats et une observation plus détaillée des scores de reconnaissance, les conclusions suivantes peuvent être avancées :

1. Les meilleurs résultats (environ 95.5 %) sont obtenus en utilisant les modèles *PLP_RPS* et *CB_SM*,
2. les scores de reconnaissance diminuent avec l'ordre du modèle,
3. la paramétrisation cepstrale donne les meilleurs résultats (en accord avec Atal [Ata74]) par comparaison aux autres types de paramétrisation considérés,
4. la distance Euclidienne sur des coefficients pondérés par leur index améliore, dans le cas des coefficients cepstraux et *PARCOR* (non représentés sur la figure 14), les scores de reconnaissance des "mauvais" locuteurs (qui sont ceux pour lesquels de mauvais scores de reconnaissance sont obtenus) de façon importante. Ceci avait déjà été suggéré par Tohkura pour les coefficients cesptraux obtenus à partir de l'analyse *LPC* [Toh85].

Au niveau temps de calcul, les techniques d'analyse *LP* et *PLP* sont comparables (environ 3500 multiplications par "frame", voir section D.1.4). Cependant, d'un point de vue pratique, un modèle *PLP_RPS* d'ordre réduit est plus intéressant.

En reconnaissance multilocuteur, il est souhaitable d'extraire uniquement les caractéristiques phonétiques alors qu'en reconnaissance monolocuteur toutes les informations (aussi bien les informations phonétiques que les informations liées au locuteur) sont utiles. Les résultats des tests multilocuteur, indiqués à la figure 15, montrent la capacité du modèle *PLP_RPS* à préserver l'information phonétique.

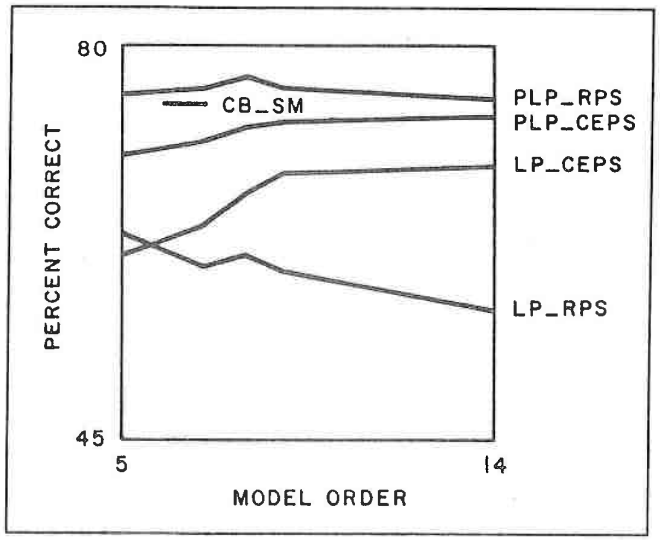


Figure 15 Comparaison de plusieurs modèles d'analyse acoustique en reconnaissance multilocuteur.

Une augmentation de l'ordre du modèle *PLP_RPS* au delà de l'ordre huit diminue les scores de reconnaissance. De manière plus générale, la combinaison d'un modèle d'ordre élevé et de la distance *RPS* diminue les performances. Dans notre étude, il a été remarqué que la distance *RPS* est particulièrement sensible aux variations de pentes survenant autour des pics du spectre de fréquence. Ainsi, un ordre du modèle trop élevé entraîne une accentuation des détails du spectre de fréquence, en particulier ceux qui fournissent des informations parasites. Le fait que l'analyse *PLP* fournisse un spectre de fréquence plus lisse que celui rendu par l'analyse *LP* semble être une des causes associées aux meilleures performances obtenues avec le modèle *PLP_RPS*.

E.1.3.3 Extension de la comparaison à des mesures de distance cepstrales pondérées présentées récemment.

Dans cette section, plusieurs modèles utilisant les techniques *LP* et *PLP* et des mesures de distance cepstrales pondérées sont considérés. Le modèle d'analyse *PLP* d'ordre cinq, qui a été trouvé optimal dans [Her87], et le modèle d'analyse *PLP* d'ordre huit, déduit de nos précédentes expériences, ont été utilisés. La figure 16 montre des spectres de fréquence de la voyelle /ae/ obtenus à partir des différents modèles étudiés. Cette voyelle a été choisie comme illustration mais les conclusions avancées ont été vérifiées sur des données plus importantes.

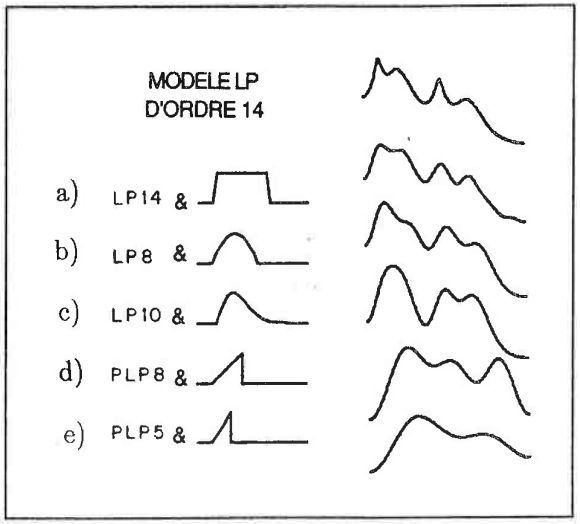


Figure 16 Spectres de fréquence d'un même segment de parole, obtenus à l'aide de plusieurs modèles d'analyse acoustique. Les "lifters" utilisés sont représentés schématiquement dans la partie gauche de la figure. Les pondérations non-nulles commencent au premier coefficient cepstral c(1) et finissent à c(M), où M représente l'ordre d'analyse. Pour plus de détails se référer au texte.

a. le "lifter" cepstral carré supprime les formants à bande étroite,
b. le "bandpass liftering" de Juang et al. accentue le premier formant et diminue les largeurs de bande [JRW86],
c. le "lifter" d'Itakura et Umezaki fusionne les deux premiers formants, alors que les autres formants ont leur largeur de bande augmentée mais ne sont pas fusionnés [IU87],
d. l'ordre huit du modèle *PLP_RPS* transforme l'axe des fréquences dans une échelle Bark, résoud les deux premiers formants et fusionne les autres formants en un pic proéminent,

e. l'ordre cinq du modèle *PLP_RPS* a son premier pic qui représente le premier formant, alors que les autres formants sont fusionnés dans un large deuxième pic.

La table 10 montre les résultats obtenus avec les modèles présentés ci-dessus en utilisant la base de données *D2* (vocabulaire alphanumérique). Dans les deux répétitions considérées, une a été utilisée comme référence et l'autre comme test. Toutes les combinaisons possibles de locuteurs ont été envisagées (en reconnaissance interlocuteur).

	Modèles d'analyse acoustique étudiés				
	a	b	c	d	e
nombre de coefficients	14	12	60	8	5
monolocuteur	90.6%	88.9%	91.1%	91.4%	90.0%
interlocuteur	52.6%	55.2%	56.4%	59.5%	62.6%

Table 10 Scores de reconnaissance obtenus pour plusieurs modèles proposés récemment.

Le modèle *PLP8_RPS* donne les meilleurs résultats en reconnaissance monolocuteur alors qu'en reconnaissance interlocuteur c'est le modèle *PLP5_RPS* qui se comporte le mieux. Compte tenu des résultats obtenus avec la technique d'analyse *PLP*, les travaux suivants ont consisté à étudier plus particulièrement cette méthode pour essayer de l'optimiser.

Chapitre E.2 OPTIMISATION DU MODELE PLP

Love has meaning only as it is
experienced in the "now"

Leo Buscaglia

E.2.1 Introduction

Afin de concevoir le modèle optimal pour la *RAP*, il est nécessaire d'étudier la technique d'analyse et la mesure de distance en même temps plutôt que de les étudier séparément. Cette tendance est visible dans beaucoup d'études récentes, particulièrement celles utilisant des paramètres cepstraux. Les coefficients cepstraux, qui ont été trouvés dans la précédente étude supérieurs aux autres paramètres de prédiction linéaire pour la reconnaissance de parole, ont été largement étudiés avec des distances cepstrales pondérées [YR79, Pal82, Toh85, HHW85a, JRW86].

Le fait de pondérer les coefficients cepstraux, influence la représentation finale de la parole utilisée par l'algorithme de comparaison des mots (voir section E.1.3.3). Yegnanarayana [YR79] montra comment la largeur de bande des pics spectraux et la pente spectrale étaient affectées dans sa comparaison entre la représentation *RPS* non-tronquée (équivalente au spectre de fréquence "group delay") et la représentation cepstrale non-tronquée (équivalente au logarithme du spectre de fréquence). Les coefficients de pondération peuvent être dérivés à l'aide d'une approche statistique [Toh85] ou à l'aide d'une approche fondée sur l'étude de la sensibilité des coefficients à certains indices spectraux. Dans le cadre de cette étude, c'est la deuxième approche qui a été étudiée.

Ce sont les irrégularités ou les pics parasites introduits par l'analyse qui affectent négativement le résultat de la comparaison dans la distance cepstrale pondérée. Ces irrégularités doivent être atténuées ou mieux encore supprimées. Un moyen simple d'accomplir ceci est de tronquer la représentation infinie des coefficients *RPS* à un nombre réduit. Tohkura, et Hanson et al. [Toh85, HHW85a] trouvèrent que la troncature de la représentation paramétrique à l'ordre *M*, ordre du modèle tout pôle, était proche de l'optimum pour un spectre de fréquence obtenu par l'analyse *LPC*. Juang et al. [JRW86] montrèrent que la troncature progressive des coefficients cepstraux d'ordre élevé par la partie décroissante du "lifter" *BP* (voir section C.3.2.4) était profitable. La partie croissante de ce "lifter" atténue les effets de la pente spectrale et accentue certains pics spectraux. Itakura et Umzaki [IU87] proposèrent la combinaison d'un "lifter" Gaussien et d'un "lifter" exponentiel (voir section C.3.2.4). La partie décroissante du "lifter" Gaussien lisse le spectre *LP*, sa partie croissante et le "lifter " exponentiel suppriment la pente spectrale et accentuent les pics spectraux. Le lissage par la partie décroissante du

"lifter" Gaussien augmente le nombre de coefficients de la réprésentation paramétrique utilisée.

Dans tous les travaux effectués, il a été observé que l'analyse *PLP* fournissait un spectre de fréquence relativement lisse et qu'ainsi un lissage progressif dans la mesure de distance ne semblait pas nécessaire [Jun87, HHW85a, HHW85b, Her87, AHW87, GM87]. Comme le spectre *PLP* est relativement lisse, la troncature des coefficients *RPS* n'affecte pas de façon significative son enveloppe.

L'étude qui suit a été consacrée à l'analyse *PLP* et aux mesures de distance cepstrales pondérées. Plus particulièrement cette étude a permis de :

1. clarifier l'impact de la sensibilité aux pics et à la pente du spectre de fréquence de la mesure de distance, en reconnaissance monolocuteur et interlocuteur,
2. montrer que l'utilisation de caractéristiques spectrales dynamiques dans la représentation paramétrique de la parole permettait d'augmenter les scores de reconnaissance,
3. améliorer le modèle *PLP_RPS*.

E.2.2 Procédure expérimentale

Afin d'optimiser la sensibilité du modèle d'analyse acoustique à certains paramètres spectraux, deux types de tests ont été menés a) tests monolocuteur b) tests interlocuteur. Puis, le modèle, déduit des expériences précédentes (a et b), a été validé en reconnaissance multilocuteur (plusieurs références par mot). Pendant la phase d'optimisation, l'intérêt ne se situait pas au niveau des performances obtenues mais plutôt au niveau de la sensibilité du système aux facteurs de pondération cepstraux. Par conséquent, l'influence de ces facteurs de pondération a été amplifiée (en évitant le lissage provoqué par l'utilisation d'un grand nombre de références ou d'un algorithme de classification) grâce aux types de tests réalisés durant la phase d'optimisation. Le modèle utilisé en reconnaissance multilocuteur a été déduit des modèles optimaux trouvés en reconnaissance monolocuteur et interlocuteur.

L'étude d'optimisation a été menée sur le modèle PLP d'ordre cinq.

E.2.3 Optimisation de la sensibilité aux pics spectraux

Il est souhaitable que la sensibilité aux pics spectraux de la mesure de distance soit faible pour des mots appartenant à une même classe et grande pour des mots appartenant à des classes différentes. Afin de progresser dans cette direction, la largeur de bande des pics spectraux a été modifiée pour contrôler la sensibilité de la distance dans le domaine cepstral et pour trouver la sensibilité optimale afin d'obtenir les meilleurs scores de reconnaissance.

Le *n-ième* coefficient cepstral est donné par :

$$c(n) = \sum_{k=1}^{M} \frac{q_k^n}{n} = \sum_{k=1}^{M} \frac{|q_k|^n e^{i\phi_k}}{n}, \tag{38}$$

où q_k est la *k-ième* racine du polynôme d'ordre M du modèle tout pôle. Si chaque coeffficient cepstral est multiplié par L^n, $L>0$, de telle façon que le coefficient cepstral modifié soit donné par :

$$c'(n) = \sum_{k=1}^{M} \frac{(L\,|q_k|)^n e^{i\phi_k}}{n}, \tag{39}$$

alors les largeurs de bande de tous les pôles sont changées par la même valeur,

$$\delta B = \frac{-1}{\pi T} ln\ L, \tag{40}$$

où B est la largeur de bande et T la période d'échantillonnage. Les largeurs de bande sont augmentées (i.e. la sensibilité aux pics spectraux est atténuée) pour $L<1$, et diminuées (i.e. la sensibilité aux pics spectraux est accentuée) pour $L>1$. Cette technique est équivalente à la multiplication des coefficients de prédiction du filtre tout pôle par une exponentielle croissante ou décroissante [Mak73],

$$a_k' = a_k e^{kln\ L}. \tag{41}$$

Un test, utilisant la base de données *D2* (vocabulaire alphanumérique), a été mené avec les deux distances : cepstrale Euclidienne et *RPS*. Les résultats obtenus sont présentés[5] à la figure 17.

[5] Dans les courbes présentées ASR est une abréviation de "automatic speech recognition".

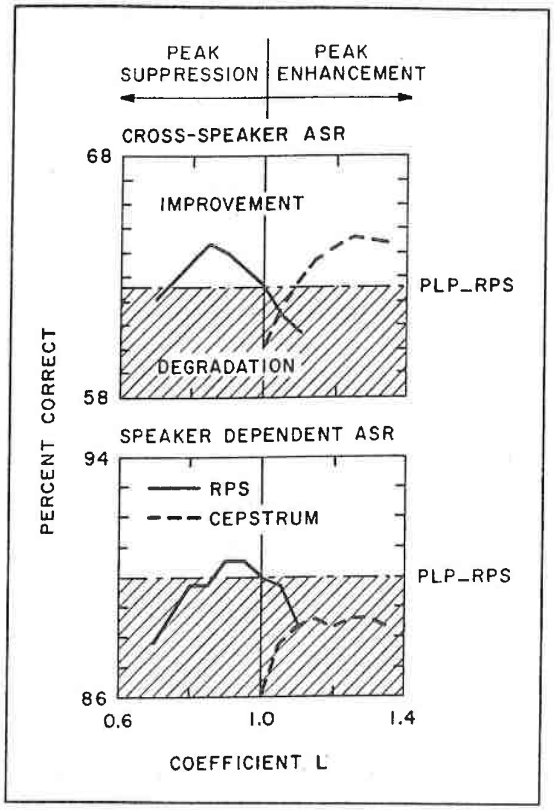


Figure 17 Effet de la suppression ou de l'accentuation des pics spectraux sur les scores de reconnaissance des modèles PLP_RPS et PLP_CEPS.

Pour la reconnaissance monolocuteur, la sensibilité aux pics spectraux de la distance *RPS* (i.e. $L=1$) est proche de l'optimum. Lorsque la distance cepstrale Euclidienne est utilisée, l'accentuation des pics spectraux ($\delta B<0$) est profitable. En reconnaissance interlocuteur, une accentuation des pics spectraux est profitable dans le cas de la distance cepstrale Euclidienne alors qu'une suppression des pics spectraux est profitable dans le cas de la distance *RPS* (par rapport au modèle *PLP_RPS*). Par conséquent, la sensibilité aux pics spectraux qui est optimale pour un modèle *PLP* (en reconnaissance interlocuteur) se trouve quelque part entre les sensibilités observées des distances cesptrale Euclidienne et *RPS*.

E.2.4 Optimisation de la sensibilité à la pente spectrale

La sensibilité à la pente spectrale est aussi un facteur important à prendre en considération dans les mesures de distance. La pente spectrale globale peut être facilement influencée par des facteurs comme les conditions d'enregistrement, les caractéristiques glottales du locuteur ou l'effort glottal produit par le locuteur. La suppression naturelle de la pente spectrale par la distance *RPS* est considérée comme un avantage de cette distance [Toh85, HW86, JRW86]. Dans cette section, la pente spectrale est supprimée ou augmentée afin de mesurer son impact sur les scores de reconnaissance.

Pour contrôler la pente spectrale, une filtrage inverse par un filtre tout zéro du premier ordre et donné par :

$$F(z) = 1 - Ca_1^{(1)} z^{-1}, \tag{42}$$

a été utilisé. Le filtrage a été réalisé par la convolution dans le domaine d'autocorrélation de la réponse impulsionnelle de (42) avec le spectre auditif. Dans cette formule, $a_1^{(1)}$ est le coefficient autoregressif du filtre *PLP* du premier ordre et C est une constante pour contrôler la pente. Le filtre tout pôle *PLP* du premier ordre est une approximation du spectre auditif par une réponse à un pôle réel qui reflète essentiellement la pente spectrale globale du spectre auditif. Plus $Ca_1^{(1)}$ est près du cercle unité dans le domaine en Z, plus la pente du filtre est raide. Si $Ca_1^{(1)}$ change de signe, la direction de la pente change aussi. La sensibilité de la mesure de distance à la pente spectrale est atténuée pour $C>0$ et augmentée pour $C<0$.

Des tests ont été effectués avec les distances cepstrale Euclidienne et *RPS* en utilisant la base de données *D2*. Les résultats obtenus sont présentés à la figure 18.

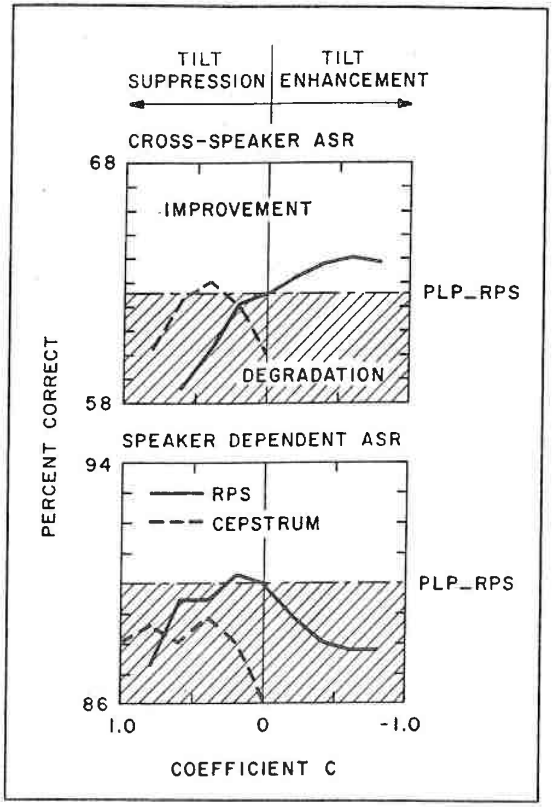


Figure 18 Effet de la suppression ou de l'accentuation de la pente spectrale globale sur les scores de reconnaissance des modèles PLP_RPS et PLP_CEPS.

La sensibilité optimale à la pente de fréquence globale est différente en reconnaissance monolocuteur et interlocuteur. En reconnaissance monolocuteur, elle est proche de l'optimum dans le cas de la distance *RPS*. En reconnaissance interlocuteur, la suppression de la pente globale du spectre de fréquence est profitable dans le cas de la distance cepstrale Euclidienne et une accentuation est profitable pour la distance *RPS* (par rapport au modèle *PLP_RPS*). Par conséquent, la sensibilité optimale à la pente spectrale globale, en reconnaissance interlocuteur, est quelque part entre les sensibilités observées des distances cepstrale Euclidienne et *RPS*.

E.2.5 Optimisation du "lifter" exponentiel

D'après les résultats précédents, en reconnaissance interlocuteur, à la fois la sensibilité aux pics spectraux et la sensibilité à la pente spectrale globale ont leur optimum quelque part entre les sensibilités des distances cepstrale Euclidienne et *RPS*. Ceci suggère que la distance optimum doit se trouver sur le continuum formé par les deux distances. La pondération cepstrale proposée par Itakura and Umezaki [IU87] et exprimée à l'aide de l'équation (26), peut être vue comme une combinaison de la multiplication des coefficients cepstraux par une puissance de n (index du coefficient) et d'un lissage Gaussien. Pour $S=0$, le lissage Gaussien est appliqué aux coefficients cepstraux et pour $S=1$ aux coefficients *RPS*. Comme le spectre de fréquence rendu par l'analyse *PLP* est relativement lisse, le lissage par la troncature des coefficients cepstraux à $n=M$ est suffisant. Par conséquent, uniquement le "lifter" exponentiel donné par :

$$E_n = n^S \quad S \geq 0. \tag{43}$$

a été considéré. Cette décision a été validée par une série importante de tests ou les "lifters" (26) et (43) ont été comparés.

Le "lifter" exponentiel constitue un continuum entre les distances cepstrale Euclidienne *(S=0)* et *RPS (S=1)*. Le modèle d'analyse acoustique devient progressivement plus sensible aux pics spectraux et moins sensible à la pente spectrale globale lorsque S varie de 0 à 1. Des tests de reconnaissance ont été effectués pour S variant dans l'intervalle $<0, 1.2>$. Comme le montre la figure 19, en reconnaissance monolocuteur l'optimum est proche de $S=1$, i.e. près de la distance *RPS*. En reconnaissance interlocuteur, l'optimum est situé entre les distances cepstrale Euclidienne et *RPS (S=0.4–0.6)*.

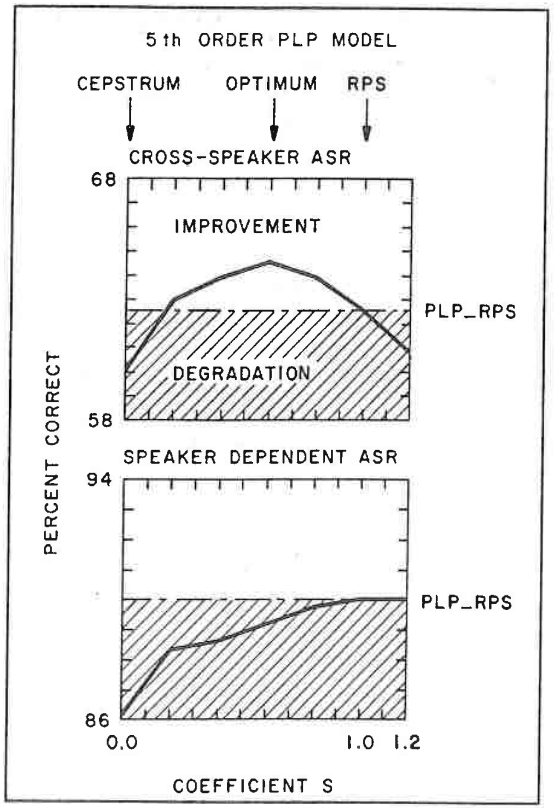


Figure 19 Effet de la pondération des coefficients cepstraux par le "lifter" exponentiel sur les scores de reconnaissance du modèle PLP d'ordre cinq. Notez la position de l'optimum.

Les résultats de ces tests confirment ceux qui ont été rapportés dans les sections E.2.3 et E.2.4 où il a été observé que l'optimum devait se trouver entre les distances cepstrale Euclidienne et *RPS*.

E.2.6 Optimisation de la résolution spectrale

Afin de trouver comment la sensibilité optimum aux pics spectraux et à la pente globale du spectre de fréquence était influencée par la résolution spectrale, les tests décrits à la section E.2.5 ont été reproduits pour différents ordres du modèle d'analyse *PLP*. La figure 20 résume les résultats obtenus.

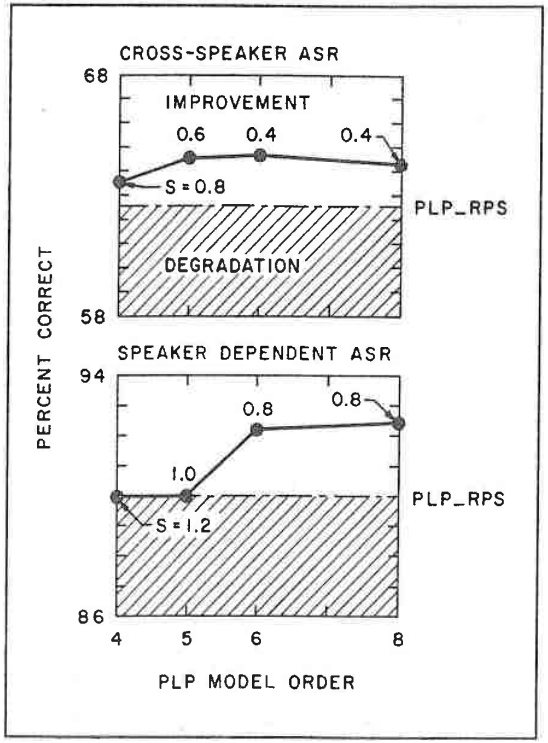


Figure 20 Maxima des scores de reconnaissance pour différents ordres du modèle PLP. L'exposant optimum du "lifter" exponentiel est indiqué sur la figure pour chaque cas.

La valeur optimale de S diminue lorsque l'ordre du modèle augmente. Ceci signifie que, plus la résolution spectrale est faible, plus la mesure de distance doit être sensible aux pics spectraux et moins sensible à la pente spectrale globale. En reconnaissance monolocuteur, les scores de reconnaissance augmentent doucement mais constamment avec l'ordre du modèle. En reconnaissance interlocuteur, un modèle d'ordre réduit, *(M=5–6)*, donne les meilleurs résultats.

E.2.7 Utilisation de caractéristiques spectrales dynamiques

Bien que la mesure de distance utilisant le "lifter" exponentiel optimisé décrit dans la section précédente améliore les scores de reconnaissance avec la base de données *D2* (vocabulaire alphanumérique), le gain est essentiellement dû à une amélioration de performance pour les mots qui ne sont pas très similaires acoustiquement. Lorsque des tests ont été effectués avec la base de données *D3*, l'utilisation du "lifter" exponentiel n'a pas modifié de façon significative les scores de reconnaissance. Par conséquent, d'autres informations sont nécessaires pour améliorer les performances des systèmes de reconnaissance lorsqu'un vocabulaire de mots acoustiquement similaires est utilisé.

Bien qu'il ait été rapporté [EB82, Fur86b] que préserver les transitions de la parole, par l'intermédiaire du modèle d'analyse, pouvait améliorer les scores de reconnaissance beaucoup d'irrégularités, comme les pics parasites ou les fluctuations dans l'estimation des largeurs de bande des pics spectraux, peuvent être accentuées par l'utilisation de caractéristiques spectrales dynamiques. Heureusement, le spectre de fréquence fourni par l'analyse *PLP* contient peu de telles irrégularités et ainsi semble être un bon candidat pour utiliser des caractéristiques spectrales dynamiques. Ces considérations ont conduit à l'étude qui suit.

Deux techniques différentes de représentation des caractéristiques spectrales dynamiques de la parole ont été envisagées :

1. une fonction calculant la différence dans le domaine temporel du spectre à court terme de la parole exprimé à l'aide d'une représentation paramétrique [EB82] et,
2. des coefficients de régression obtenus à partir d'une représentation paramétrique cepstrale de la parole [Fur86b].

Une étude pilote a montré que la représentation des caractéristiques spectrales dynamiques de la parole à l'aide de coefficients de régression était moins bruitée que la simple différence entre deux "frames". Par conséquent, une mesure de distance tenant compte de caractéristiques spectrales instantanées et dynamiques a été définie par :

$$d(k,l) = \sum_{j=1}^{M} \left(W\left(p_{k,j} - p_{l,j}\right)\right)^2 + \sum_{j=1}^{R} \left(\left((1-W)\left(r_{k,j} - r_{l,j}\right)\right)^2, \qquad (44)$$

où k est la *k-ième* "frame" du mot test, l est la *l-ième* "frame" du mot de référence, R est le nombre de coefficients de régression, W est un facteur de combinaison entre 0 et 1, $p_{i,j}$ est le *j-ième* coefficient cepstral pondéré pour la *i-ième* "frame" et $r_{i,j}$ est donné par :

$$r_{i,j} = \frac{\sum_{q=-Q}^{Q} q p_{i+q,j}}{\sum_{q=-Q}^{Q} q^2}, \qquad (45)$$

où Q est une constante telle que $2Q+1$ représente le nombre de "frames" de la fenêtre utilisée pour calculer les coefficients de régression.

Les coefficients *RPS* ont été sélectionnés pour représenter les coefficients cepstraux pondérés et une optimisation expérimentale de l'équation (44) a été effectuée avec la base de données *D3* (vocabulaire de mots acoustiquement similaires). L'optimisation a été menée en reconnaissance multilocuteur (9 références par mot). L'esprit de cette optimisation est différent de celui de l'optimisation qui vient d'être décrite (aucun test monolocuteur ou interlocuteur n'a été effectué) car cette étude a été réalisée de façon indépendante et les résultats ont été combinés par la suite avec ceux de l'optimisation précédente. De plus, comme le vocabulaire considéré contient très peu de mots, il a été facile de mener directement l'optimisation en reconnaissance multilocuteur. Des tests menés en reconnaissance interlocuteur auraient sûrement permis d'amplifier les phénomènes qui vont être rapportés. Les tests ont été effectués six fois pour différentes répétitions du vocabulaire utilisé. Dans chaque test, deux paramètres de l'équation (44) ont varié, le troisième étant fixe. Les valeurs fixes des paramètres ont été prises égales à : $W = 0.5$, $2Q + 1 = 7$ "frames" (70 ms), et $R=M$, où M est l'ordre du modèle *PLP*. Des tests ont été effectués pour un ordre du modèle *PLP* variant de cinq à huit. La figure 21 montre les résultats (moyennés) obtenus pour le modèle d'ordre cinq.

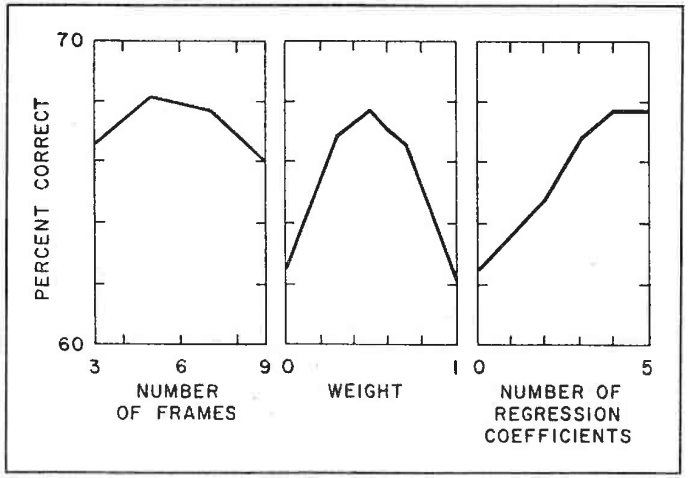


Figure 21 Effet de la fenêtre temporelle utilisée dans le calcul des coefficients de régression (à gauche), de la pondération des caractéristiques spectrales instantanées et dynamiques (au milieu), et du nombre de coefficients de régression (à droite) sur les scores de reconnaissance.

La taille de la fenêtre optimale $(2Q+1)$ est dans l'intervalle 50–70 ms et la pondération optimale des caractéristiques spectrales instantanées et dynamiques est $W=0.5$, en accord avec les résultats de Furui [Fur86b]. Toutefois, il a été observé que la troncature des

coefficients de régression à $R=4$ fournissait de meilleurs résultats que l'utilisation de tous les coefficients de régression ($R=M$) dans la mesure de distance définie par l'équation (44). Les tests effectués avec des modèles *PLP* d'ordre différent ont donné des résultats similaires. L'utilisation de caractéristiques spectrales dynamiques, à l'aide de la mesure de distance (44) optimisée, diminua l'erreur de reconnaissance de 14% par rapport au modèle *PLP_RPS* en reconnaissance multilocuteur sur la base de données *D3*.

E.2.8 Reconnaissance multilocuteur avec le nouveau modèle PLP

Afin de valider le nouveau modèle d'analyse acoustique déduit des tests d'optimisation précédents, deux tests supplémentaires ont été effectués :

A. reconnaissance multilocuteur sur la base de données *D1* (vocabulaire "keyboard"),
B. reconnaissance multilocuteur sur la base de données *D4* (vocabulaire des chiffres).

Ces tests ont utilisé à la fois la nouvelle mesure de distance (avec $S=0.6$) et les caractéristiques spectrales dynamiques (avec $R=4$, $W=0.5$, $2Q+1=70$ *ms*). Les propriétés de la reconnaissance multilocuteur correspondent à une intégration des propriétés observées en reconnaissance monolocuteur et interlocuteur. Si l'on se réfère à la figure 20, l'exposant optimum du "lifter" exponentiel se situe dans l'intervalle S=0.8–1.0 en reconnaissance monolocuteur et S=0.4–0.6 en reconnaissance interlocuteur (pour un ordre du modèle variant de cinq à huit). Par conséquent, pour la mesure de distance, la valeur $S=0.6$, qui est un compromis entre les optimums trouvés en reconnaissance monolocuteur et interlocuteur, a été utilisée. Les modèles *PLP* trouvés optimaux dans les précédentes études, i.e le modèle d'ordre huit pour le vocabulaire "keyboard" et le modèle d'ordre cinq pour le vocabulaire des chiffres [Her87], ont été sélectionnés. Les données de test comportaient huit locuteurs dans le test A et quarante huit locuteurs dans le test B. Enfin, le test A utilisait deux références par mot du vocabulaire et le test B, deux ou douze (deux tests ont été effectués). Les mots de référence ont été choisis aléatoirement dans le lot des données disponibles. Les tests furent effectués pour différentes combinaisons des mots de référence : deux fois pour la base de données *D1* ("keyboard"), quatre fois pour la base de données *D4* (chiffres) avec douze références par mot, et vingt quatre fois pour la base de données *D4* avec deux références par mot. Les résultats obtenus, en terme de diminution du taux d'erreur de reconnaissance par comparaison au modèle *PLP_RPS*, sont présentés dans la table 11.

base de données	D1 ("keyboard")	D4 (chiffres)	D4 (chiffres)
nombre de références par mot	2	2	12
diminution du taux d'erreur de reconnaissance	8.5%	10.6%	13.1%

Table 11 Pourcentages de diminution du taux d'erreur de reconnaissance (par comparaison au modèle PLP_RPS) obtenus par le modèle optimisé.

Les scores de reconnaissance étaient d'environ 80% pour la base de données *D1*, supérieurs à 90% pour la base de données *D4* en utilisant deux références par mot, et supérieurs à 98% pour la base de données *D4* en utilisant douze références par mot.

Enfin, le test A a été effectué pour S variant dans l'intervalle $<0.2, 1.2>$ et R variant dans l'intervalle $<0, 8>$. Les résultats obtenus sont présentés à la figure 22.

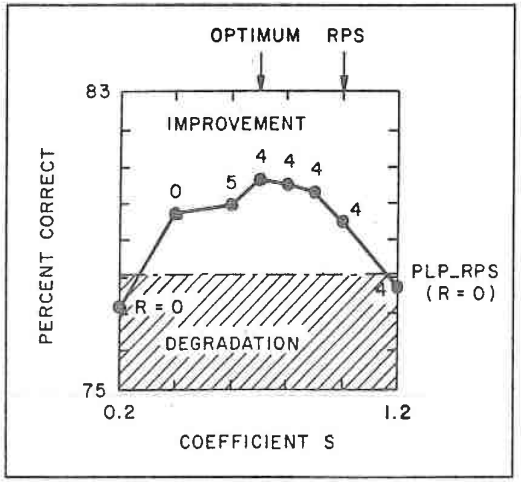


Figure 22 Scores de reconnaissance exprimés en fonction de la pondération exponentielle et du nombre de coefficients de régression utilisés dans la comparaison. Le nombre de coefficients de régression optimal est indiqué près de chaque point de mesure.

Les meilleures performances du nouveau modèle, obtenues pour $S=0.7$ et $R=4$, sont en accord avec celles rapportées dans les tests précédents.

E.2.9 Conclusions

Les conclusions qui peuvent être extraites de l'étude d'optimisation présentée sont les suivantes :

1. la distance *RPS* , lorsqu'elle est utilisée avec le modèle *PLP*, est trop sensible aux pics spectraux et pas assez sensible à la pente spectrale globale,
2. une nouvelle distance, qui pondère chaque coefficient cepstral par une puissance de son index, atténue les problèmes liés à la distance *RPS*,
3. des caractéristiques spectrales dynamiques sont particulièrement intéressantes avec le modèle *PLP* dans le cadre de vocabulaires comprenant des mots acoustiquement similaires.

La combinaison de la nouvelle mesure de distance et de caractéristiques spectrales dynamiques avec le modèle *PLP* diminue le taux d'erreur de reconnaissance d'environ 10% (en reconnaissance multilocuteur) par rapport au modèle *PLP_RPS*.

Plus généralement, une méthodologie pour examiner la sensibilité d'un modèle d'analyse acoustique à certaines caractéristiques spectrales a été proposée. Cette méthodologie n'est pas spécifique au modèle *PLP* et peut être appliquée à d'autres modèles d'analyse acoustique.

Chapitre E.3 EVALUATION ET AMELIORATION DES SYSTEMES DE RAP EN ENVIRONNEMENT BRUITE

Differences can arise because of our way of thinking : our conclusions are based on our premises, our premises are based on our conclusions. We think we mean what we think.

Gerard Nierenberg

E.3.1 Introduction

Un modèle d'analyse acoustique doit bien se comporter dans des conditions normales mais aussi dans le cadre de conditions environnantes difficiles, comme en présence de bruit. La section C.5.3 a présenté quelques travaux, effectués récemment, dont le but était d'améliorer la robustesse des systèmes de reconnaissance de parole en présence de bruit. Néammoins, beaucoup d'études sont encore nécessaires pour atteindre un niveau de performance acceptable.

Dans le chapitre E.1, il a été montré qu'un modèle utilisant l'analyse *PLP* fournissait de meilleurs scores de reconnaissance que plusieurs autres modèles proposés récemment. L'étude ci-après étend l'évaluation du modèle *PLP* à des environnements bruités.

L'effet de l'ordre du modèle *PLP* sur les scores de reconnaissance a tout d'abord été étudié. Ensuite, des modèles utilisant l'analyse *PLP* ont été comparés à d'autres modèles d'analyse acoustique utilisant différents "lifters" et mesures de distance. Etant donné les scores de reconnaissance obtenus à l'aide du meilleur modèle (environ 50% pour un *SB*=5 dB en reconnaissance monolocuteur), il a été décidé de développer une nouvelle méthode d'analyse, appelée *SLP*. Après une présentation de cette nouvelle technique, les résultats de son évaluation sont rapportés.

Enfin, une étude du comportement de plusieurs systèmes de reconnaissance lorsque la parole est *produite* en présence de bruit est présentée.

E.3.2 Préliminaires

Dans les tests qui suivent la base de données *D2* a eté utilisée, sauf pour les tests présentés dans la section E.3.6 qui ont été effectués sur la base de données *D5*. Afin de simuler deux types de bruit différents, du bruit blanc Gaussien et du bruit blanc Gaussien filtré (appelé dans la suite de ce chapitre bruit blanc filtré) ont été alternativement ajoutés au signal de parole. Le bruit blanc filtré a une pente spectrale moyenne similaire à celle de la parole [HW86]. De plus, ce type de bruit est une bonne approximation de celui enregistré dans une voiture, sur une autoroute, avec les fenêtres ouvertes et le ventilateur allumé. La fonction de transfert du filtre passe-bas utilisé lors des tests est : $1/\left(1-0.92z^{-1}\right)$. Pour tous les tests, les mots de référence ont été enregistrés dans des conditions normales et aucune connaissance sur les caractéristiques du bruit (type, *SB*, etc) n'a été utilisée. Les tests ont été effectués en reconnaissance monolocuteur et interlocuteur.

Dans la section C.5.3.2 la *mesure de distorsion cepstrale projetée*, proposée par Mansour et al. [MJ88a], a été décrite. Dans leur étude, Mansour et al. montrèrent que, lorsque du bruit blanc est ajouté au signal de parole, la norme des vecteurs cepstraux est réduite et la déviation angulaire entre deux vecteurs cepstraux n'est pas trop sensible aux dégradations engendrées. Compte tenu de ces résultats, ils proposèrent une famille de mesures cepstrales projetées. Certains des tests qui vont être décrits utilisent un cas particulier de cette famille de mesures, obtenu en limitant le paramètre α au cas α=1 (voir section C.5.3.2).

E.3.3 Evaluation de plusieurs modèles d'analyse acoustique en présence de bruit

E.3.3.1 Le modèle PLP_RPS : effet de l'ordre du modèle

Afin de déterminer l'ordre du modèle *PLP* qu'il est souhaitable d'utiliser en présence de bruit, des tests ont été effectués pour différents *SB* : 25, 15, 5 dB et pour de la parole non-bruitée. Le modèle *PLP_RPS*, qui s'est avéré être le modèle le plus performant parmi les modèles étudiés en environnement non-bruité (mis à part le modèle optimisé), a été sélectionné. Les deux types de bruit décrits dans la section précédente ont été alternativement ajoutés au signal de parole. Les modèles *PLP* d'ordre cinq et huit, trouvés optimaux lors de précédentes études ([Her87], et section E.1.3.2), ont été considérés et comparés au modèle *PLP* d'ordre quatorze. Les résultats obtenus avec la base de données *D2* sont présentés aux figures 23 et 24.

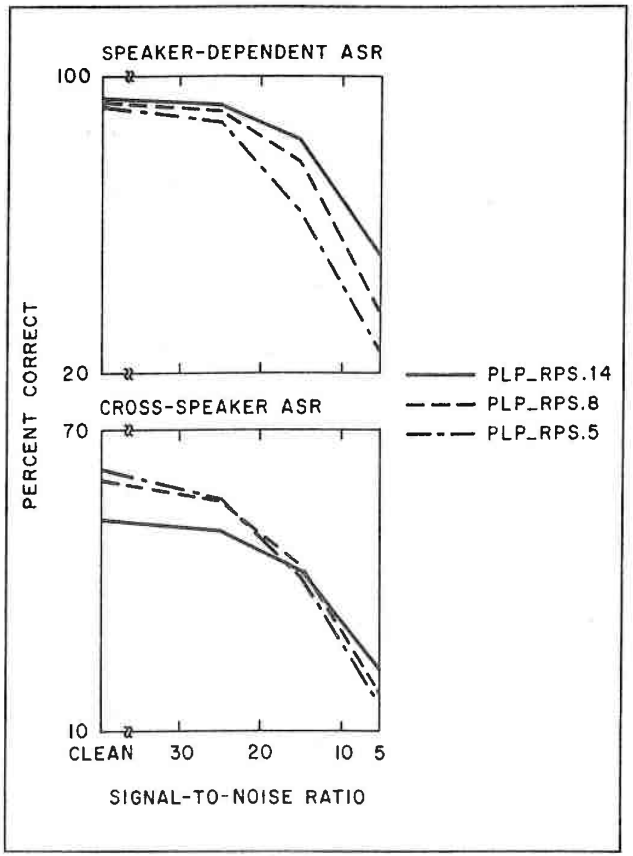


Figure 23 Effet de l'ordre du modèle PLP_RPS sur les scores de reconnaissance pour le cas du bruit blanc Gaussien.

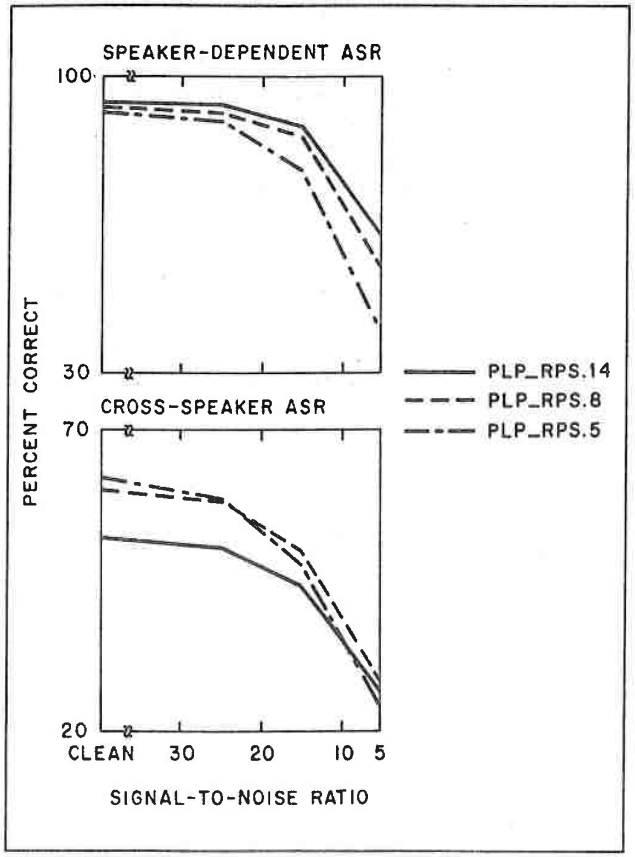


Figure 24 Effet de l'ordre du modèle PLP_RPS sur les scores de reconnaissance pour le cas du bruit blanc filtré.

En reconnaissance monolocuteur, un ordre élevé du modèle *PLP* est le plus performant. Ceci est en accord avec les résultats obtenus par Hanson et wakita [HW86]. Cependant, en reconnaissance interlocuteur et pour un *SB* faible, le modèle *PLP_RPS* d'ordre huit donne les meilleurs résultats (dans le cas du bruit blanc filtré) ou constitue une bonne alternative (dans le cas du bruit blanc Gaussien) au modèle *PLP_RPS* d'ordre quatorze. Dans son étude sur le comportement de l'analyse *LPC* en présence de bruit, Tierney [Tie80] rapporta que, afin de modéliser à la fois le signal et le bruit, un modèle d'ordre plus élevé que celui utilisé dans des environnements non-bruités devait être considéré. En présence de bruit, le modèle *PLP* nécessite néanmoins un ordre moins élevé

que celui requit par l'analyse *LPC*. Ces résultats sont à rapprocher avec ceux obtenus dans le cadre d'environnements non-bruités.

Enfin, remarquons que lorsque le signal de parole est dégradé par du bruit blanc filtré, les scores de reconnaissance obtenus sont beaucoup plus élevés que lorsque la dégradation est causée par du bruit blanc Gaussien. Le gain est de l'ordre de 5 à 10 dB.

Dans les tests suivants, compte tenu des résultats présentés ci-dessus, il a été décidé d'utiliser des modèles d'analyse acoustique d'ordre quatorze.

E.3.3.2 Comparaison entre les modèles CB_SM, LP et PLP en environnement bruité

Afin de clarifier quelles sont les performances du modèle *PLP* en présence de bruit, une étude comparative utilisant les techniques d'analyse *LP*, *PLP* et banc de filtres critiques a été effectuée. Le modèle utilisant le banc de filtres critiques a été décrit dans la section E.1.3.1. Deux mesures de distance : cepstrale Euclidienne et *RPS* ont été utilisées. Ces distances ont été sélectionnées à cause de leur très différente sensibilité à certaines caractéristiques, comme les pics spectraux ou la pente spectrale globale. Les tests ont été effectués en reconnaissance monolocuteur et interlocuteur. Les figures 25 et 26 présentent les résultats obtenus (moyennés sur l'ensemble des locuteurs de test) pour les modèles *CB_SM*, *LP_CEPS* et *PLP_CEPS* et les deux types de bruit étudiés.

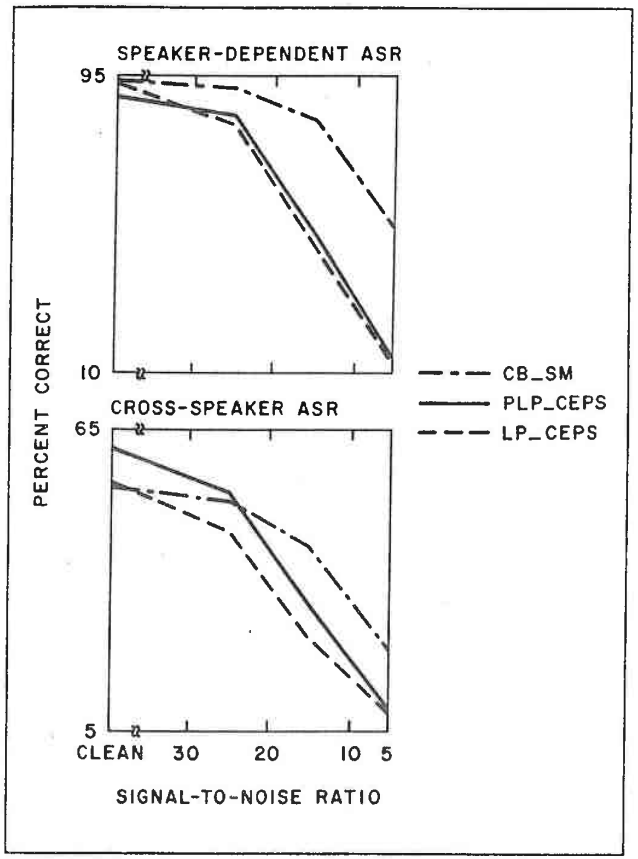


Figure 25 Comparaison des modèles CB_SM, LP_CEPS et PLP_CEPS pour le cas du bruit blanc Gaussien.

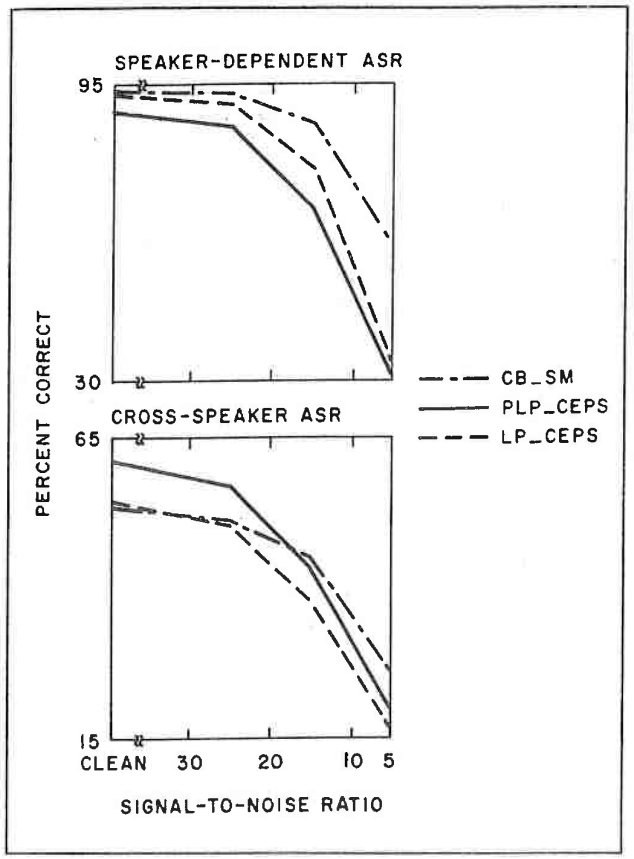


Figure 26 Comparaison des modèles CB_SM, LP_CEPS et PLP_CEPS pour le cas du bruit blanc filtré.

En reconnaissance monolocuteur, le modèle *CB_SM* obtient de bien meilleures performances que les autres modèles représentés dans les figures ci-dessus. En reconnaissance interlocuteur, le modèle *CB_SM* fournit les meilleurs résultats lorsque le *SB* est faible. Les bons scores de reconnaissance obtenus par ce modèle, particulièrement pour un *SB* faible, résultent de sa sensibilité aux pics spectraux. Cette hypothèse a été confirmée par les résultats obtenus avec la distance *RPS*. Ceux-ci sont présentés aux figures 27 et 28.

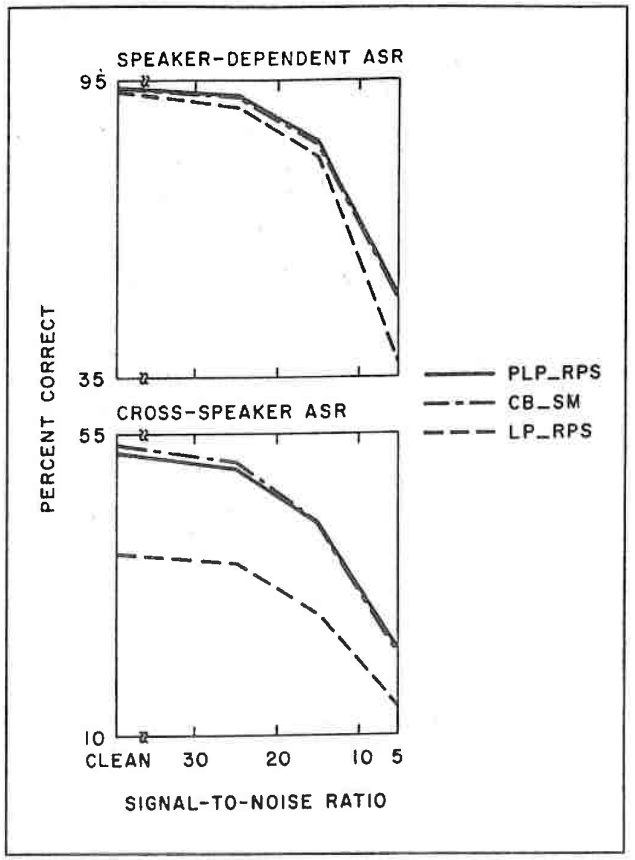


Figure 27 Comparaison des modèles CB_SM, LP_RPS et PLP_RPS pour le cas du bruit blanc Gaussien.

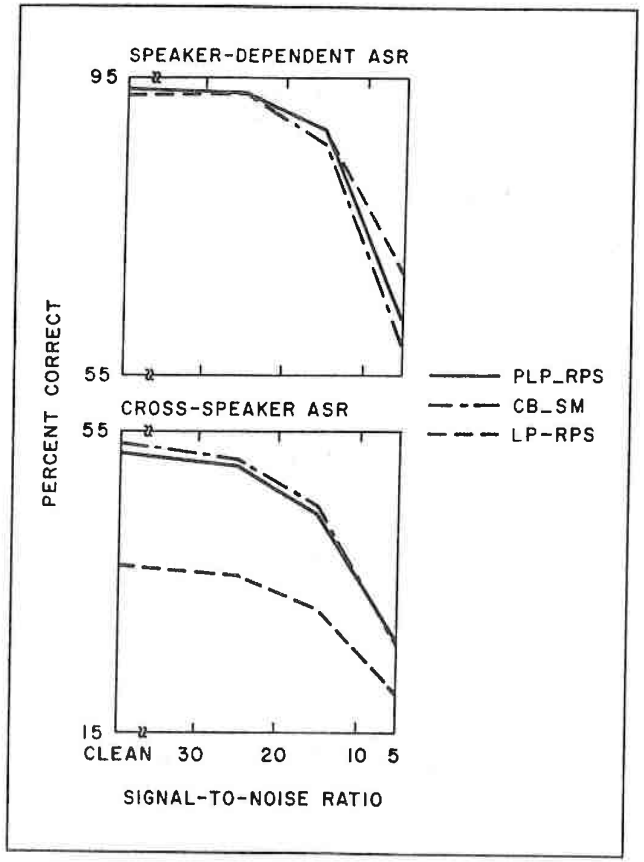


Figure 28 Comparaison des modèles CB_SM, LP_RPS et PLP_RPS pour le cas du bruit blanc filtré.

La mesure de distance *RPS*, qui est une mesure sensible aux variations de pente survenant autour des pics spectraux, améliore les performances obtenues avec les techniques d'analyse *LP* et *PLP*. Elle permet d'obtenir, avec le modèle *PLP*, des performances similaires à celles fournies par l'utilisation du modèle *CB_SM*. Notons, enfin, qu'en reconnaissance interlocuteur, le modèle *PLP_RPS* donne de bien meilleures performances que les modèles *LP_CEPS* et *LP_RPS*.

E.3.3.3 Optimisation du modèle PLP en présence de bruit

Afin de savoir si, en présence de bruit, la pondération des coefficients cepstraux par la distance *RPS* était optimale avec l'analyse *PLP*, de nouveaux tests, utilisant le "lifter" exponentiel proposé au chapitre E.2, ont été effectués. Le *SB* a été fixé à 15 dB et différentes valeurs de *S* (l'exposant du "lifter" exponentiel) ont été considérées pour les deux types de bruit décrits précédemment. Les deux environnements bruités conduirent aux mêmes conclusions. La figure 29 présente les résultats obtenus pour le cas du bruit blanc filtré.

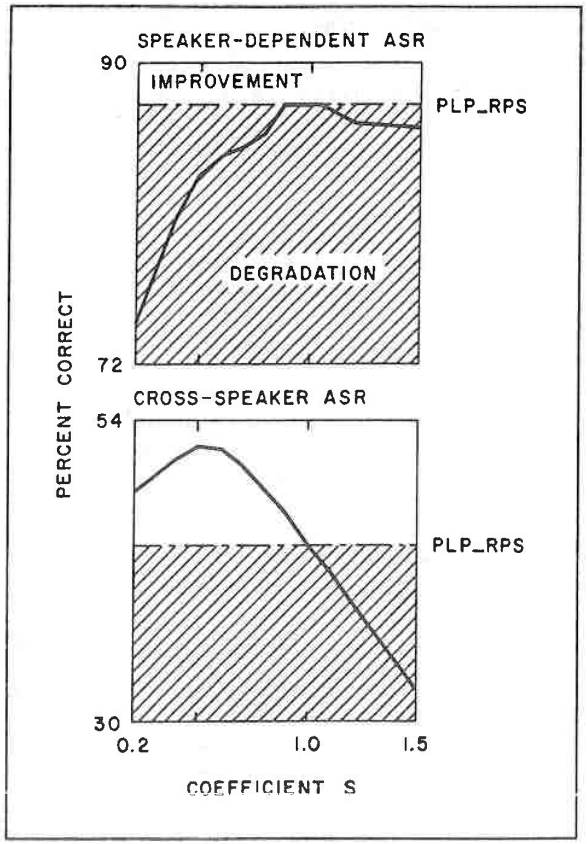


Figure 29 Effet de la pondération des coefficients cepstraux par le "lifter" exponentiel sur les scores de reconnaissance du modèle PLP d'ordre quatorze (SB=15 dB).

Comme pour le cas de la parole non-bruitée, un optimum a été trouvé, en reconnaissance interlocuteur, entre la distance cepstrale Euclidienne *(S=0)* et la distance *RPS (S=1)*. L'optimum se trouve aussi dans le même intervalle : *S=0.5–0.6*. Cependant, il a été observé que, lorsque le *SB* diminue, l'optimum se déplace vers la distance *RPS*. Cet optimum est en fait une fonction du *SB* des mots de référence et de test. Ceci a eté verifié expérimentalement par des tests supplémentaires. En reconnaissance monolocuteur la distance *RPS* est proche de l'optimum.

Une extension de cette étude à d'autres conditions pour les mots de test et de référence (e.g. mots de test et de référence bruités ou mots de référence bruités et mots de test non-bruités) montra que l'optimum varie en fonction des conditions de bruit. En particulier, une valeur faible de l'exposant *S* du "lifter" exponentiel est souhaitable lorsqu'à la fois les mots de test et de référence sont bruités. Les meilleurs scores de reconnaissance ont été obtenus lorsque les conditions de test étaient les mêmes que les conditions d'apprentissage. Plus de détails sur les tests effectués sont disponibles dans [JW89].

Afin de minimiser les temps de calcul et compte tenu des performances obtenues, dans la suite de cette étude nous avons décidé de nous concentrer sur la reconnaissance monolocuteur.

E.3.4 Développement d'un modèle auditif utilisant des connaissances physiologiques

E.3.4.1 Le modèle auditif à synchronisation temporelle SLP

En présence de bruit additif et pour des valeurs faibles du *SB*, les scores de reconnaissance fournis par le meilleur modèle étudié sont encore très inférieurs à ceux obtenus pour de la parole non-bruitée. Récemment, beaucoup d'études se sont orientées vers le développement de modèles auditifs utilisant des concepts physiologiques. De tels modèles ont déjà donné de bons résultats en milieu bruité [HL86, Ghi87]. Par conséquent, il fut décidé d'étudier une nouvelle méthode d'analyse qui simulerait des mécanismes physiologiques et qui aurait la caractéristique d'être plus robuste en milieu bruité.
Un modèle auditif peut être développé suivant deux approches [Del84] :

1. une approche *structurelle ou physiologique*,
2. une approche *fonctionnelle*.

Dans la première approche, tous les phénomènes observés essayent d'être expliqués. Inversement, l'approche fonctionnelle modélise des propriétés auditives jugées importantes afin de construire un modèle qui sera ensuite utilisé en tant que boîte noire dans un système plus général. Dans l'étude qui est présentée, la deuxième approche a été choisie et un nouveau modèle a été développé. En effet, le but visé était d'inclure le modèle développé dans un système de reconnaissance automatique. Certaines propriétés de l'oreille interne furent modélisées et intégrées dans un modèle appelé *SLP* (en abréviation de "time synchronous linear prediction analysis"). Même si les recherches menées en physiologie sont actives, nos connaissances des mécanismes prenant place dans le système auditif

ne nous permettent pas de décrire celui-ci dans son ensemble. Il fut donc décidé de modéliser les mécanismes jugés importants sous réserve qu'ils puissent être simulés par ordinateur sans trop de difficultés.

Le nouveau modèle auditif développé, *SLP*, décrit la propagation des sons dans l'oreille interne et la conversion de l'énergie acoustique dans une représentation au niveau des fibres nerveuses. Les mécanismes liés au passage des sons dans le canal auditif de l'oreille moyenne n'ont pas été considérés. En effet ceux-ci peuvent être grossièrement modélisés par un filtre passe-bas (voir section A.2.1.3) et sont supposés peu importants en reconnaissance de parole. Quant au mécanisme de contrôle automatique de gain souvent attribué à l'oreille moyenne, nous avons choisi de ne pas le modéliser à ce niveau de l'étude. Le modèle *SLP* est similaire au niveau de la détection de l'enveloppe spectrale à celui proposé par Ghitza [Ghi87]. Toutefois, il diffère de celui-ci par le banc de filtres utilisé pour simuler les vibrations de la membrane basilaire (nous avons développé nos propres filtres et adopté une échelle de fréquence non-linéaire). De plus, nous avons introduit des mécanismes de cross-corrélation des canaux adjacents et de filtrage passe-bas qui, nous le verrons par la suite, sont deux étapes particulièrement utiles.

La technique d'analyse *SLP* simule les vibrations mécaniques de la membrane basilaire (*BM*) et les effets de filtrage correspondants ainsi que la transformation de ces vibrations en potentiels électriques. C'est un modèle de l'*oreille interne* et plus particulièrement un modèle de la cochlée. La figure 30 présente un diagramme de cette nouvelle méthode d'analyse.

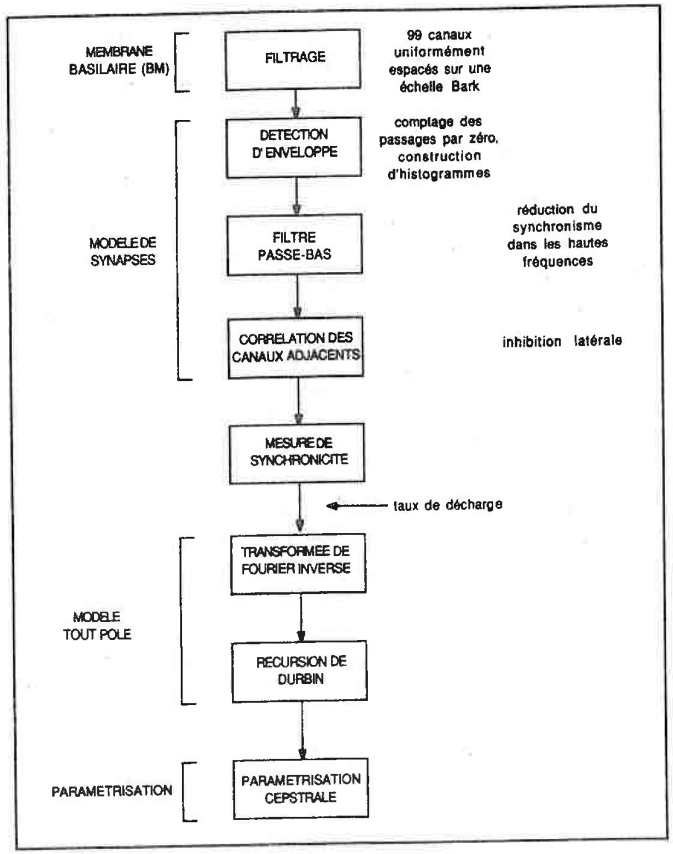


Figure 30 Diagramme fonctionnel de l'analyse SLP.

La partie filtrage est formée de 99 filtres uniformément espacés sur une échelle Bark (de 0 à 5 kHz). Les filtres sont symétriques avec des pentes de 10 dB/Bark. Ces valeurs ont été optimisées grâce à des tests de reconnaissance préliminaires. En particulier, des filtres asymétriques ont fourni des scores de reconnaissance inférieurs à ceux obtenus avec les filtres sélectionnés. La principale fonction de cette partie filtrage, qui représente une simulation des fonctions de la membrane basilaire, est de séparer un mélange complexe de sons en régions ayant un *SB* élevé.

La deuxième partie du modèle *SLP* est composée de trois étapes. Tout d'abord une détection d'enveloppe, utilisant une méthode de passages par zéro et de construction

d'histogrammes [All85, Ghi87], détermine pour chaque canal l'enveloppe du spectre de fréquence. Ensuite un filtre passe-bas ayant une pente progressive (3 dB à 2 kHz, 9 dB à 4 kHz et 13 dB à 6 kHz) est chargé de réduire le synchronisme (réalisé dans la troisième partie) pour les hautes fréquences [Sen87]. En effet, le synchronisme des taux de décharge des fibres nerveuses dans la région des hautes fréquences n'est pas très évident [Joh80]. Ce filtrage permet, en particulier, d'estomper le spectre haute fréquence ce qui le rend plus lisse. Enfin, un algorithme permettant de corréler les canaux adjacents [All85, HL86] accentue les pics spectraux (ce qui est particulièrement utile en présence de bruit) et simule certaines propriétés du mécanisme d'inhibition latérale. Ce mécanisme de corrélation des canaux adjacents est similaire à celui utilisé par Allen, Deng et Hunt [All85, DGG88, HL86] et différent de celui utilisé par Shamma [Sha88] qui effectue une soustraction entre canaux. Le but de ce traitement est de fournir une représentation de l'enveloppe spectrale qui soit relativement invariante dans une plage assez grande de *SB* [DGG88].

La troisième partie définit une mesure de synchronicité (en sommant tous les canaux) dont le but est de combiner tous les canaux qui fournissent des informations à une fréquence donnée en même temps. La largeur de chaque région (correspondant à un intervalle d'histogramme) est utilisée comme une estimation de l'intensité spectrale dans cette région [Ghi87]. Comme le modèle de Ghitza [Ghi87], le modèle proposé ne respecte pas le principe de tonotopie.

La dernière partie est un modèle tout pôle (comme dans les techniques *LP* et *PLP*) suivi d'une paramétrisation cepstrale.

Dans les tests réalisés, trois différentes versions de cette nouvelle méthode d'analyse ont été utilisées :

1. *SLP1* qui n'inclut pas le filtre passe-bas et le mécanisme de corrélation,
2. *SLP2* qui n'inclut pas le mécanisme de corrélation,
3. *SLP3* (appelé aussi *SLP*) qui est l'analyse telle qu'elle a été décrite ci-dessus.

E.3.4.2 Evaluation du modèle SLP

Tout d'abord les performances du modèle auditif *SLP1* ont été évaluées par comparaison aux modèles *LP* et *PLP* en reconnaissance monolocuteur et en présence de bruit blanc Gaussien. La figure 31 indique les résultats obtenus avec la base de données *D2*.

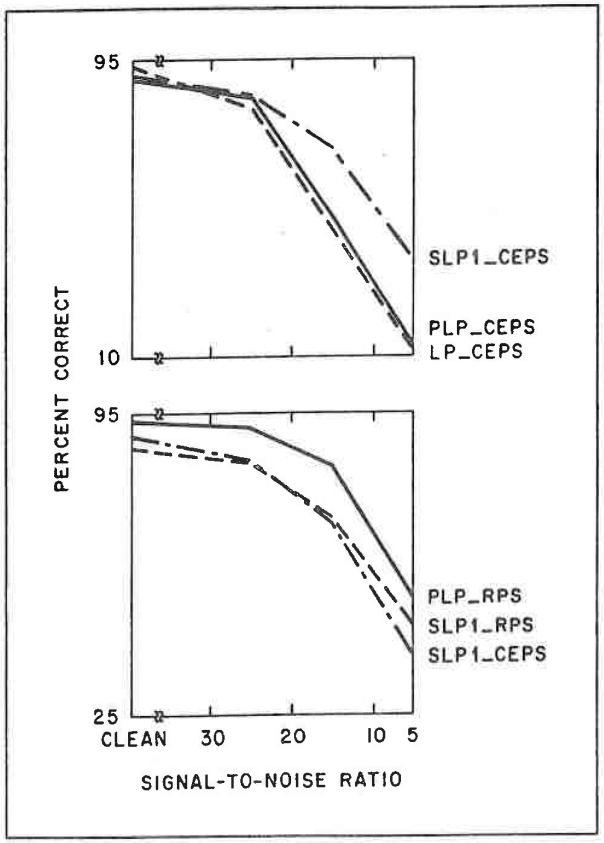


Figure 31 Evaluation du modèle auditif SLP1 en reconnaissance monolocuteur.

Lorsque les méthodes d'analyse *LP*, *PLP* et *SLP1* sont combinées avec la distance cepstrale Euclidienne le modèle *SLP1* donne les meilleurs résultats. Cependant, lorsque la distance *RPS* est utilisée le modèle *PLP_RPS* fournit les meilleurs scores de reconnaissance. Ces résultats montrent, une fois encore, que la technique d'analyse et la mesure de distance doivent être étudiées en même temps. Une étude séparée peut amener des conclusions incomplètes sinon erronées. La distance *RPS*, qui est bien adaptée à l'analyse *PLP*, ne semble pas adéquate pour l'analyse *SLP1*.

Les tests présentés ont été suivis de tests complémentaires destinés à évaluer l'influence des mécanismes de corrélation des canaux adjacents et de filtrage passe-bas.

Ces tests ont été effectués pour un *SB*=5 dB. Les résultats obtenus sont présentés dans la table 12.

modèles étudiés	scores de reconnaissance
PLP_RPS	51.4%
LP_RPS	37.8%
SLP1_RPS	45.0%
SLP2_RPS	51.7%
SLP3_RPS	58.3%

Table 12 Comparaison des scores de reconnaissance de plusieurs modèles d'analyse acoustique pour un SB=5 dB dans le cas de bruit blanc Gaussien additif.

Chacun des nouveaux mécanismes introduits entraîne une amélioration des scores de reconnaissance. Cette fois, le modèle auditif *SLP* donne les meilleurs résultats (même si ceux-ci sont encore insuffisants). La corrélation des canaux adjacents accentue les pics spectraux, ce qui est désirable en présence de bruit, et le filtrage passe-bas lisse le spectre haute fréquence, ce qui facilite la mise en correspondance des mots par l'algorithme de programmation dynamique.

E.3.5 Etude comparative de "lifters" cepstraux et de mesures de distance associés à des modèles tout pôle en milieu bruité.

Afin de compléter l'évaluation des différents modèles étudiés, des tests visant à clarifier les performances des trois modèles tout pôle *LP*, *PLP* et *SLP* utilisés avec différents "lifters" ("lifter" exponentiel et bandpass (*BP*) défini à la section C.3.2.4) et mesures cepstrales (Euclidienne et projetée) ont été effectués. Le but recherché était d'étudier expérimentalement l'effet sur les modèles tout pôle *PLP* et *SLP* de la mesure de distorsion cepstrale projetée et du "lifter" *BP* qui ont été utilisés avec succès en combinaison avec l'analyse *LP*. Deux types de conditions expérimentales ont été simulés : milieu non-bruité et milieu bruité avec un *SB*=5 dB. Les conditions de bruit sélectionnées furent sévères afin d'amplifier la contribution du "lifter" et de la mesure de distance (ou de distorsion) en milieu bruité. Les résultats obtenus avec la base de données *D2* sont présentés dans la table 13.

conditions expérimentales	modèle tout pôle	S=0	S=0.6	S=1	BP
test 1	LP	92.5%	94.7%	92.5%	94.4%
(mots de référence et de test non-bruités,	PLP	88.9%	92.8%	93.3%	93.6%
distance cepstrale Euclidienne)	SLP	90.0%	88.9%	87.8%	88.3%
test 2	LP	11.9%	19.7%	37.8%	28.1%
(mots de référence non-bruités et mots de test bruités,	PLP	13.6%	32.5%	51.4%	36.1%
distance cepstrale Euclidienne)	SLP	32.5%	54.7%	58.3%	54.7%
test 3	LP	90.0%	93.9%	92.5%	94.2%
(mots de référence et de test non-bruités,	PLP	86.1%	90.8%	90.8%	90.3%
distorsion cepstrale projetée)	SLP	87.5%	88.9%	86.1%	89.5%
test 4	LP	13.1%	44.2%	70.0%	50.8%
(mots de référence non-bruités et mots de test bruités,	PLP	10.3%	35.0%	66.7%	45.3%
distorsion cepstrale projetée)	SLP	36.1%	58.3%	62.5%	60.3%

Table 13 Scores de reconnaissance obtenus par différents modèles (d'ordre 14) utilisant les techniques d'analyse LP, PLP et SLP en reconnaissance monolocuteur. Pour la parole bruitée le SB a été fixé à 5 dB.

Pour les tests en présence de bruit, les meilleurs résultats sont obtenus avec la technique d'analyse *SLP* et la pondération *RPS* lorsque la mesure de distorsion cepstrale projetée n'est pas utilisée (test 2). Cependant, en milieu non-bruité, le modèle *SLP* est un petit peu moins performant que les autres modèles étudiés. L'utilisation de la mesure de distorsion cepstrale projetée améliore, dans tous les tests effectués, les scores de reconnaissance. Les meilleurs résultats sont obtenus avec l'analyse *LP* (test 4). Enfin, il faut remarquer que la mesure de distorsion cepstrale projetée, en milieu non-bruité, diminue généralement les scores de reconnaissance de quelques % (test 3 comparé au test 1). Comme le mentionne Mansour et Juang [MJ88a], l'égalisation adaptative par "frame" que réalise la mesure de distorsion cepstrale projetée crée probablement des discontinuités qui perturbent l'algorithme de programmation dynamique.

E.3.6 Reconnaissance automatique de la parole produite en milieu bruité

De précédentes études ont montré [PBNY85, BMM88, SJA88] que des différences systématiques existent au niveau de la structure phonétique de mots isolés (ou de la parole continue) produits dans des environnements non-perturbés et bruités. Il est important de mesurer l'incidence de ces changements phonétiques sur les systèmes de reconnaissance. Dans ce but, le comportement des techniques d'analyse *LP* et *PLP* a été étudié lorsque les locuteurs de test étaient exposés à du bruit ambient. La technique *SLP* n'a pas été considérée dans ces tests afin de limiter les temps de calcul. De plus, cette nouvelle analyse nécessite encore des améliorations. Les techniques *LP* et *PLP* ont été comparées en faisant varier l'ordre des modèles tout pôle et en utilisant les "lifters" exponentiel et *BP*. Dans tous les cas les mots de test ont été prononcés en présence de bruit (voir la description de la base de données *D5* à la section E.1.1). Le premier test a été effectué pour des mots de test non-bruités (aucun bruit n'a été ajouté, ce qui revient à tenir compte uniquement de l'effet Lombard). Pour le deuxième, du bruit blanc Gaussien a été ajouté au signal de parole afin d'obtenir un *SB* de 20 dB. Les conditions de bruit sélectionnées n'ont pas été trop sévères à cause de la dégradation déjà occasionnée par l'effet Lombard. Dans les deux tests les mots de référence étaient produits dans un environnement non-bruité et ne comportaient aucun bruit additif. Les résultats obtenus sont présentés à la figure[6] 32.

[6] Dans cette figure SNR est une abréviation de "signal-to-noise ratio".

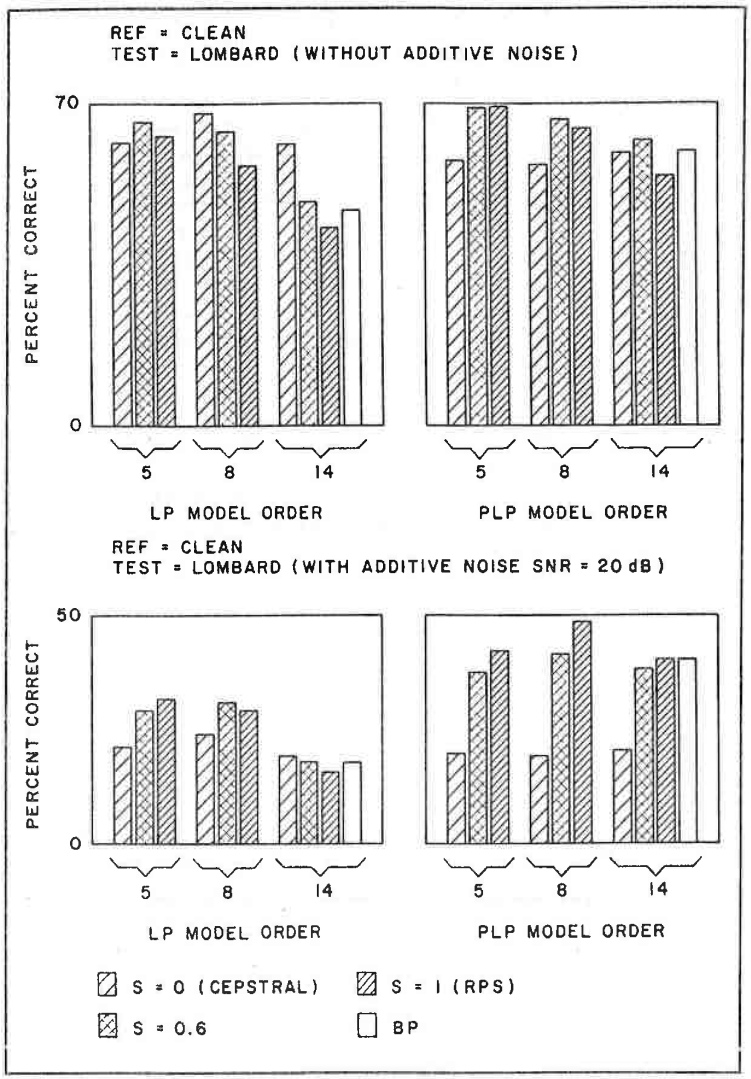


Figure 32 Effet de l'ordre du modèle tout pôle sur les techniques d'analyse LP et PLP lorsque la parole est produite dans du bruit (pour les mots de test). Pour les deux figures du haut, les mots de test étaient non-bruités alors que pour les deux figures du bas les mots de test étaient bruités (SB=20 dB).

La degradation provoquée par l'effet Lombard est importante. Notons qu'un modèle tout pôle d'ordre réduit, qui fournit une approximation grossière de la forme du spectre de fréquence, donne les meilleurs résultats. Cependant, lorsque du bruit blanc est ajouté aux mots de test l'ordre du modèle utilisé doit être augmenté. Dans ce cas, les meilleurs scores de reconnaissance sont obtenus avec le modèle *PLP_RPS* d'ordre huit.

Enfin, des tests supplémentaires effectués avec la mesure de distorsion cepstrale projetée indiquèrent que cette mesure n'améliore pas de façon significative (et même diminue dans le cas de l'analyse *PLP*) les scores de reconnaissance en présence de l'effet Lombard. La figure 33 présente les résultats de ces tests.

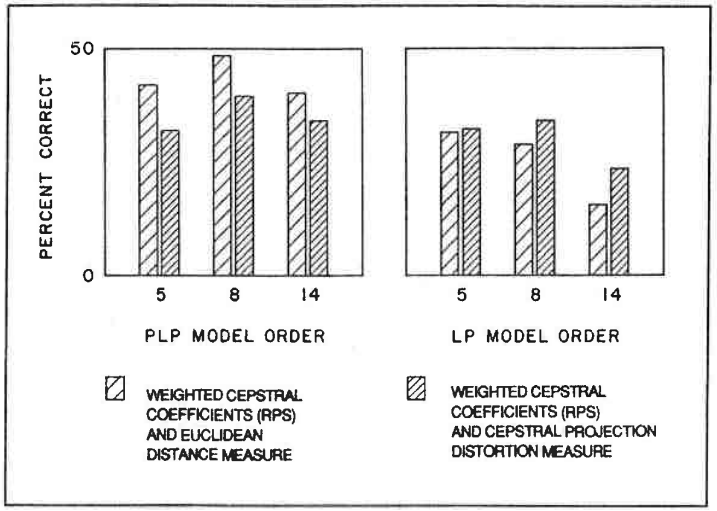


Figure 33 Effet de la mesure de distorsion cepstrale projetée sur les modèles LP et PLP en présence de l'effet Lombard (mots de test bruités, SB=20 dB).

E.3.7 Résumé et discussion

Une étude comparative en environnement bruité de plusieurs modèles d'analyse utilisant différents "lifters" cepstraux et mesures de distance (ou de distorsion) a été présentée. Pour l'analyse *PLP*, et de façon similaire à ce qui avait été observé pour l'analyse *LP*, l'ordre du modèle tout pôle doit être augmenté en présence de bruit par rapport à des conditions non-bruitées. L'augmentation de l'ordre du modèle permet de représenter à la fois le bruit et le signal de parole. Une comparaison de modèles dérivés de la technique *PLP* par rapport à d'autres modèles proposés récemment a montré que le modèle *PLP_RPS* est plus performant que des modèles utilisant l'analyse *LP* en reconnaissance interlocuteur. Ceci est dû, en particulier, aux propriétés de la distance *RPS* et à l'adéquation entre cette distance et l'analyse *PLP*. La sensibilité aux pics spectraux de la distance *RPS* est désirable en présence de bruit (particulièrement lorsque le *SB* est faible). Grâce à l'utilisation du "lifter" exponentiel, il a été observé que la pondération optimale des coefficients cepstraux est une fonction du *SB* des mots de test et de référence. En fait, la modification du *SB* modifie la sensibilité optimale de la mesure de distance aux pics spectraux.

Afin d'améliorer les performances des systèmes de reconnaissance en présence de bruit, une nouvelle méthode d'analyse, *SLP*, a été développée. Ce modèle, fondé sur des concepts physiologiques, simule certains mécanismes de la cochlée. Cette nouvelle méthode d'analyse est encore en développement. Toutefois, grâce à une comparaison avec les méthodes d'analyse *LP* et *PLP*, il a été montré que le modèle *SLP* fournissait déjà des résultats encourageants. En particulier, le modèle *SLP_RPS* a fourni les meilleurs résultats, en reconnaissance monolocuteur, lorsqu'aucun mécanisme d'égalisation adaptative par "frame" n'a été utilisé. Il a été montré que des mécanismes de filtrage passe-bas, permettant de lisser les hautes fréquences, et de corrélation des canaux adjacents étaient deux étapes importantes du modèle. Cependant, afin d'obtenir les performances désirées, il est souhaitable de modifier ce modèle afin, par exemple, d'obtenir un spectre de fréquence plus lissé. En particulier, des caractéristiques dynamiques de la parole comme l'adaptation à court terme et le masquage postérieur doivent être pris en considération. Enfin, une pondération cepstrale adéquate et une mesure de distance adaptée à cette nouvelle technique d'analyse doivent être étudiées.

Lorsque les conditions de bruit sont simulées et que la parole est produite dans des conditions normales, il a été montré que, en reconnaissance monolocuteur, la mesure de distorsion cepstrale projetée permet d'améliorer les scores de reconnaissance, les meilleurs résultats étant obtenus avec l'analyse *LP*.

En présence de l'effet Lombard, c'est-à-dire lorsque la parole est produite dans du bruit, les scores de reconnaissance chutent rapidement. Un modèle d'analyse acoustique qui fournit une approximation grossière de la forme du spectre de fréquence donne les meilleurs résultats. Ceux-ci sont obtenus avec un modèle *PLP_RPS* d'ordre réduit. Les bonnes performances de ce modèle, par rapport au modèle *LP*, sont dues au fait qu'il est moins sensible aux changements pouvant survenir au niveau des pics du spectre de fréquence [Her87]. Etant donné une méthode d'analyse, il est apparu que l'obtention

des meilleurs résultats passait par un ajustement de la pondération cepstrale et de l'ordre du modèle tout pôle au *SB* des mots de test. De plus, la mesure de distorsion cepstrale projetée, qui a été testée avec succès sans la présence de l'effet Lombard, n'améliore pas de façon significative et même diminue les scores de reconnaissance lorsque cet effet est présent. C'est un point important de notre étude. Lorsque la parole est produite dans du bruit, la dégradation provoquée par l'effet Lombard est plus importante que lorsque du bruit est ajouté au signal de parole. Les modèles qui ont été ajustés pour donner de bonnes performances en milieu bruité ne sont pas aussi performants lorsque des changements dans la structure phonétique interviennent. Par conséquent, afin d'améliorer les performances des systèmes de reconnaissance en présence de bruit, l'effet Lombard doit être étudié en priorité.

A cause des temps de calcul induits par les tests effectués, les études présentées ont souvent fixé la valeur du *SB*. Cette valeur a été choisie en fonction des hypothèses qui voulaient être vérifiées. Pour être complets, les résultats obtenus doivent être généralisés à d'autres valeurs du *SB*.

Chapitre E.4 UTILISATION DE CONNAISSANCES PHONETIQUES EN RECONNAISSANCE AUTOMATIQUE DE MOTS ISOLES MULTILOCUTEUR

Les mots sont les vêtements de nos pensées qui, pas plus que notre propre personne, ne devraient être revêtues de haillons et de loques, ni garnies de poussière.

Lord Chesterfield

E.4.1 Introduction

Un mot peut être caractérisé par un ensemble de propriétés acoustiques. Ces propriétés sont partiellement déterminées par le système de phonation et par la façon dont le système auditif réagit en fonction des différents sons. Elles permettent d'effectuer des distinctions sur la base d'informations phonétiques.

Une importante caractéristique de la réprésentation d'un mot en terme d'indices acoustiques est que généralement davantage d'indices que le nombre nécessaire pour procéder à l'identification sont utilisés. Cette représentation est *redondante*. Cela vient du fait que les indices acoustiques sélectionnés ne sont pas toujours présents dans la parole produite. En effet, la parole est hautement variable avec le locuteur et le contexte [Ste86]. La redondance introduite permet de traiter cette variabilité. Toutefois, afin de réduire la variabilité associée aux indices acoustiques extraits du signal de parole, Stevens [Ste87] proposa l'utilisation d'indices exprimant des *relations* entre d'autres indices ou propriétés (e.g. différence entre les valeurs des formants, contour du fondamental, etc). Ces indices sont qualifiés de *relationnels*.

Un système de reconnaissance à base d'indices est généralement constitué de deux étapes :

1. identification des indices distinctifs à partir d'une représentation acoustique, généralement un spectrogramme,
2. identification des unités lexicales à partir des indices acoustiques.

La deuxième étape est généralement réalisée à partir d'un système à base de règles de production.

Dans cette étude, les avantages d'un système utilisant des indices acoustiques comme unités principales de reconnaissance, dans le cas de mots acoustiquement similaires, sont présentés. Le principal niveau de variabilité pris en compte est celui qui est dû à la

différence de morphologie entre les locuteurs. En effet, le but de cette étude est la reconnaissance de mots isolés aussi les différences de style et de vitesse d'élocution ne sont pas très importants. De plus, le vocabulaire choisi (*E-SET* dans un premier temps) limite beaucoup l'influence du contexte. Comme il est indiqué au chapitre D.2, l'approche proposée est fondée sur l'utilisation de plusieurs sources de connaissances dans un système hybride à deux passes, *ORION*, qui utilise des connaissances phonétiques pendant la deuxième passe. Le but de ce système à deux passes est de faciliter la discrimination entre les mots acoustiquement similaires du vocabulaire tout en conservant de bonnes performances dans les autres cas.

E.4.2 Acquisition et représentation des connaissances

E.4.2.1 Utilisation de distinctions phonétiques

En regardant de plus près les erreurs de reconnaissance des précédentes évaluations utilisant le modèle d'analyse acoustique *PLP*, quatre sous-groupes ou classes (appartenant à la base de données *D1*) de mots confondus facilement furent identifiés : *E-SET*={B, C, D, E, G, P, T, V, Z, FEED}, {M, N}, {F, S, X} et {LINE, NINE}. Lorsqu'un mot de ces classes est mal reconnu, il a été observé que les premiers mots reconnus appartiennent à la même classe. L'étude présentée dans ce chapitre s'est intéressée à la classe *E-SET* mais les conclusions peuvent être généralisées aux autres classes.

Le modèle *PLP* optimisé, décrit au chapitre E.2, fournit un score de reconnaissance de 68% en reconnaissance multilocuteur (avec 9 références par mot) sur la classe *E-SET*. Le problème rencontré avec l'algorithme de programmation dynamique et plus généralement dans les algorithmes de comparaison de formes est que toutes les parties du mot testé ont le même poids pendant la reconnaissance. Le mot "FEED" est quelquefois confondu avec les autres mots de la même classe alors qu'il n'y a pas de confusion possible en regardant un spectrogramme de ce mot (un indice distinctif est, par exemple, la présence très fréquente d'une période de silence et d'une barre d'explosion à cause du phonème /d/).

Des techniques de discrimination ont déjà été proposées pour améliorer la reconnaissance de mots acoustiquement similaires [RW81, CV88, CSL85, BCL82] (voir la section C.4.2). Dans l'approche proposée, afin d'effectuer la discrimination, le système de reconnaissance invoque une deuxième passe dont le but est d'extraire des indices acoustiques distinctifs menant à une identification correcte. Les mots sont décrits en terme de propriétés acoustiques. Ces propriétés ont été déterminées par l'étude visuelle de spectrogrammes traditionnels et de spectrogrammes auditifs (ou pseudo-spectrogrammes) obtenus à partir de l'analyse *PLP*. Ces spectrogrammes ont été visualisés et plus généralement "édités" à l'aide du système d'analyse *STAR* qui a été décrit dans le chapitre D.3.

E.4.2.2 Définition et extraction des indices discriminants

Afin de ne pas engendrer des calculs trop importants, des indices grossiers ont été sélectionnés. Grâce à l'observation de spectrogrammes et pseudo-spectrogrammes neuf indices de base ont été considérés :

1. présence d'une barre d'explosion accompagnée de ses caractéristiques (force, position, etc.),
2. durée de la partie non-voisée au début de chaque mot,
3. présence d'une barre de voisement,
4. présence d'une période de silence (à l'intérieur du mot) accompagnée de ses caractéristiques (durée, position),
5. mouvements du deuxième et troisième formant calculés au commencement de la partie voisée de chaque mot à l'aide de l'analyse *LPC* d'ordre quatorze,
6. énergie dans certaines bandes de fréquence avant le début de la voyelle,
7. durée de la partie consonne et de la partie restante,
8. taux de passages par zéro,
9. mouvements des deux pics spectraux obtenus par le modèle *PLP* d'ordre cinq.

Pour la classe *E-SET*, l'information qui est importante se trouve au début des mots. Afin d'être pertinents les indices qui ont été décrits nécessitent une segmentation, au préalable, des mots en consonnes et voyelles. Par conséquent, dans le but de déterminer les frontières associées aux différents segments, un système de segmentation automatique (*SAIPH*), utilisant la technique d'analyse *PLP* et des caractéristiques spectrales dynamiques, a été développé. Ce système de segmentation automatique a été décrit au chapitre D.4.

Comme le montre la figure 34, les pics de la courbe de transition du mot "V" (choisi comme exemple) indiquent la frontière entre les parties consonne et voyelle ainsi que le début et la fin du mot.

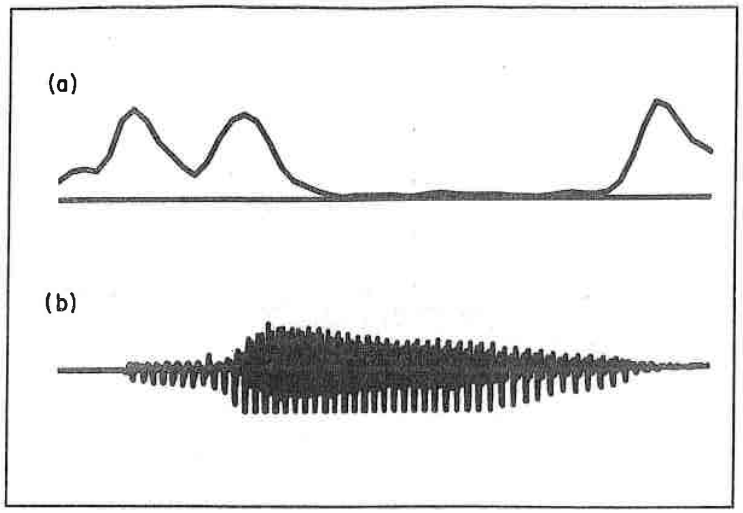


Figure 34 Signal temporel (a) et courbe de transition (b) associés au mot "V".

Cette mesure de transition ne dépend pas du locuteur et peut être calculée, en même temps que le modèle d'analyse acoustique utilisé, durant la première passe.

Le modèle PLP d'ordre cinq a aussi été choisi pour aider à la discrimination. Comme le montre la figure 35 pour les mots ("V", "G", "B"), une représentation grossière du spectre de fréquence, à l'aide d'un modèle d'ordre réduit de l'analyse *PLP*, constitue un indice particulièrement intéressant dans le cadre du vocabulaire étudié.

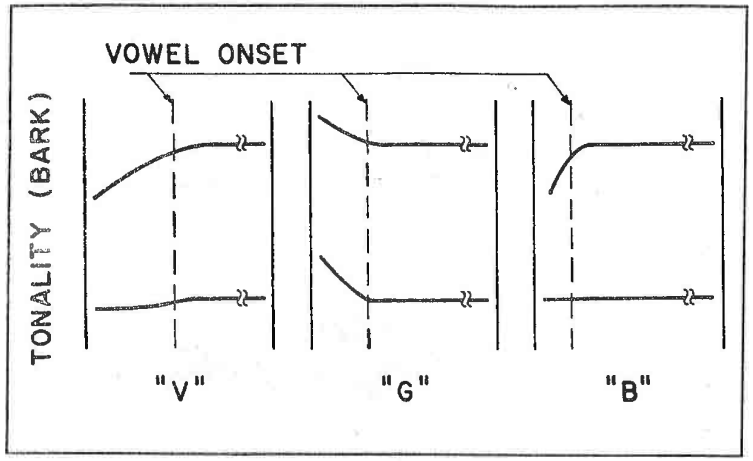


Figure 35 Mouvements des pics spectraux du modèle PLP d'ordre cinq pour les mots "V", "G", "B".

Des caractéristiques spectrales grossières ont déjà été employées par Makhoul [Mak73] en reconnaissance de parole. Dans cette étude, un modèle *LPC* à deux pôles était utilisé.

E.4.2.3 Représentation des connaissances et raisonnement incertain

Les connaissances phonétiques ont été encodées à l'aide d'un langage de représentation qui décrit la connaissance en terme de règles et de faits. La représentation utilisée par ce langage est exprimée par une syntaxe à base de "frames" (qui dans ce cas désigne un ensemble d'informations contenues dans une structure commune). Deux sortes de connaissances sont distinguées :

1. *connaissances schématiques*,
2. *connaissances applicatives*.

où les *connaissances schématiques* décrivent quelles sont les valeurs qui peuvent être prises par un indice, et les *connaissances applicatives* représentent un ensemble de faits et de règles qui manipulent ces indices. Les connaissances sont encodées à l'aide d'une structure à base d'arbre de décision pour chaque règle traitant d'un mot particulier.

Cette étude a été dirigée vers l'acquisition et la représentation des connaissances plutôt que sur le développement d'un moteur d'inférence. Pour ce dernier, un produit commercial a été sélectionné. Il s'agit de *KWB* (en abréviation de "knowledge workbench") qui est en fait un environnement de programmation pour développer des systèmes experts. Le langage de représentation des connaissances utilisé fait aussi partie de cet ensemble logiciel. Le moteur d'inférence (écrit en prolog) permet, en particulier, de communiquer avec des fonctions écrites dans des langages usuels (par exemple C) afin de

récupérer les indices extraits automatiquement. De plus, il fournit des outils intéressants de mise au point et de manipulation de *connaissances incertaines*.

Les indices ont été décrits de façon qualitative (variables linguistiques) en terme de haut, bas, moyen, etc... afin de construire une base de connaissances facilement modifiable par des chercheurs qui ne seraient pas familiés avec le système. C'est aussi la façon la plus naturelle de décrire des spectrogrammes.

Le système manipule des connaissances incertaines en utilisant un mécanisme défini dans *MYCIN* [BS85] qui associe à chaque indice extrait un nombre compris dans un intervalle donné. Ainsi, chaque indice est caractérisé par une variable linguistique et une valeur d'incertitude comprise entre 0 et 1. Ceci permet de fournir au moteur d'inférence des informations continues et non discrètes. La valeur d'incertitude correspond au degré de confiance que l'on attribue à la présence d'un indice donné dans le signal manipulé.

Enfin, une valeur d'incertitude (entre 0 et 1) est aussi associée à chaque branche de l'arbre de décision représentant une règle. Cette valeur d'incertitude a été ajustée en fonction des résultats de reconnaissance obtenus et de la confiance que l'on attribue à chaque règle. Par exemple, une des branches de l'arbre de décision définissant le mot "B" est :

if burst_strength='weak' and voice_bar='yes'
and plp_peak_1='flat' and plp_peak_2='increase_fast'
and begin_no_voiced='very_small'
then word='b' cf(0.7).

où *burst_strength* indique l'énergie avec laquelle une barre d'explosion est présente, *voice_bar* indique l'existence (ou l'absence) d'une barre de voisement, *begin_no_voiced* représente la durée de la partie non-voisée en début de mot, et *plp_peak_1* et *plp_peak_2* représentent les mouvements des pics spectraux du modèle PLP d'ordre cinq dans la transition consonne-voyelle.

Si l'on compare ces règles à celles utilisées par un système de décodage acoustico-phonétique comme *APHODEX* nos règles sont beaucoup plus simples. En effet, les règles développées utilisent uniquement des informations extraites sur le signal manipulé alors que les règles d'*APHODEX* tiennent compte, en particulier, des phonèmes candidats pour le segment précédent et le segment suivant (contexte gauche et contexte droit). Dans notre cas, le contexte étant similaire, cette information n'est pas nécessaire. La simplicité de nos règles découle de l'application considérée.

E.4.3 Combinaison d'un système à règles de production et d'une approche reconnaissance des formes

Les deux approches à base de règles de production et de comparaison de formes ont été combinées à l'aide d'une *stratégie de décision*. La décision finale est prise en tenant compte des résultats fournis par les deux passes du système. Aux hypothèses générées par l'algorithme de programmation dynamique est associé un facteur de confiance dépendant des mots qui ont été reconnus et d'une matrice de confusion obtenue par apprentissage. Quant à la deuxième passe (fondée sur l'extraction d'indices) elle effectue une discrimination entre trois classes : {B, D, E}, {G, T, P}, {V, Z, C, FEED}. Cela signifie que lorsque les candidats trouvés par cette deuxième passe sont fournis, il y a aussi une décision prise quant à l'appartenance du mot de test à une de ces trois classes. La décision finale considère les candidats du système à base de connaissances et les candidats, parmi les trois premiers, de l'algorithme de programmation dynamique qui appartiennent à une des trois classes identifiée par le système à base de connaissances. Un facteur de confiance est calculé grâce à une fonction combinant les différents facteurs de confiance des hypothèses générées pendant les deux passes. Le nouveau facteur de confiance est défini par :

$$nouveau\ CF = CF1 + CF2 - CF1 \times CF2.$$

où 1=*première passe* et 2=*deuxième passe*.

Cette fonction, fondée sur l'hypothèse que les résultats générés par les deux passes sont indépendants, permet d'accumuler des évidences progressivement. Le mot reconnu est celui auquel est associé le facteur de confiance le plus élevé. Cette stratégie de décision suit le principe d'*information prépondérante*, utilisé dans le système de reconnaissance de parole continue *HEARSAY II* [EHRLR80], qui donne plus d'importance à la source de connaissance qui est la plus fiable. Dans le système proposé, pour les classes de mots difficiles les candidats générés par la deuxième passe sont plus fiables que ceux générés par l'algorithme de programmation dynamique.

Ce système hybride a été testé sur le vocabulaire *E-SET*, qui est un sous-ensemble de la base de données *D1*. Les scores de reconnaissance obtenus sont très proches de 90%, ce qui correspond à une diminution de l'erreur de reconnaissance de plus de 60% par rapport aux résultats obtenus lors de la première passe. Dans la version actuelle du système, les facteurs de confiance associés aux indices extraits sont calculés mais ne sont pas encore pris en compte par le moteur d'inférence. De plus, les facteurs de confiance qui ont été utilisés (ceux qui sont associés aux règles) ont besoin d'être ajustés par l'intermédiaire de tests plus intensifs. Ces dernières remarques laissent entrevoir un facteur supplémentaire de progression qui peut aussi être obtenu par le raffinement des connaissances utilisées.

E.4.4 Résumé et conclusions

ORION est un système hybride qui utilise plusieurs sources de connaissances : *psychoacoustiques* (dans l'analyse *PLP*), *physiologiques* (par l'introduction de caractéristiques spectrales dynamiques) et *phonétiques*. Un tel système tient compte de notre connaissance sur la parole sans pour autant négliger ses limitations. Pour un vocabulaire de mots acoustiquement non-similaires, un modèle qui simule des propriétés du système auditif humain et accentue les transitions spectrales a fourni les meilleurs résultats parmi les modèles étudiés. Pour des mots acoustiquement similaires, l'introduction de connaissances phonétiques, l'utilisation de modèles perceptuels (pour la segmentation automatique et l'extraction d'indices acoustiques) et le développement d'une stratégie de décision élaborée améliorent considérablement les scores de reconnaissance et permettent de s'affranchir des limitations associées à l'algorithme de programmation dynamique.

Cependant, la version actuelle du système nécessite quelques améliorations (citées dans la section précédente) et doit être étendu à d'autres classes acoustiquement similaires du vocabulaire étudié.

La parole est la soeur jumelle de la vision,
elle est incapable de se mesurer

Walt Whitman

PARTIE F
EXTENSION A D'AUTRES APPLICATIONS

Une extension de l'algorithme de segmentation automatique (*SAIPH*) à une application d'étiquetage automatique de la parole est présentée. L'approche proposée a été conçue autour d'un modèle *"blackboard"* qui facilite la communication entre plusieurs sources de connaissances et permet l'implantation et le test de stratégies complexes. Après une présentation générale du système, les différentes sources de connaissances utilisées sont décrites et les premiers résultats intermédiaires sont rapportés.

Chapitre F.1 UNE ARCHITECTURE "BLACKBOARD" POUR L'ETIQUETAGE AUTOMATIQUE DE LA PAROLE

*Respecter dans chaque homme **l'homme,***
sinon celui qu'il est, au moins celui
***qu'il pourrait** être, qu'il **devrait** être.*

Henri-Frédéric Amiel

F.1.1 Introduction

L'intégration de plusieurs sources de connaissances pour résoudre un problème donné semble être une fonction naturelle de l'être humain. Comme le dit K. Church [Chu87] *"In understanding an utterance, a native speaker (subconsciously) invokes his knowledge of the language, the environment, and the context. The sources of knowledge (KS) include parametrization of speech signals (signal processing and electrical engineering), the physiology of the vocal tract and the ear (articulatory phonetics), descriptive accounts of the acoustic characteristics of speech sounds (phonetics), positional constraints of phonemes within syllables (phonotactics), theoretical explanations of the variability in pronunciations (phonology), the stress and intonation patterns of words and phrases (prosodics), lexical constraints, the composition of words (morphology), the grammatical structure of the language (syntax), the meaning of words and sentences (semantics), and the context of conversation (pragmactics)"*. Un système de compréhension de la parole idéal devrait utiliser toutes ces sources de connaissances (ou contraintes) afin de décoder le message émis. En pratique, des compromis sont faits. Dans la plupart des systèmes, les connaissances énoncées ci-dessus sont incomplètes ou absentes.

Dans le chapitre D.4, un système de segmentation automatique (*SAIPH*) a été présenté comme une composante du système hybride *ORION*. Encouragé par les résultats obtenus, une extension de ce système à l'étiquetage automatique de la parole a été étudiée. L'étiquetage automatique consiste à détecter, dans le signal de parole, les changements acoustiques et phonétiques et à identifier la nature des différents segments. Dans ce but, des paramètres types de représentation de la parole, des indices acoustiques, ainsi que diverses sources de connaissances sont utilisés.

L'étiquetage automatique peut être appliqué à une base de données ou peut constituer une étape de traitement d'un système de reconnaissance automatique de la parole. Un des buts visés, était de concevoir un système qui puisse être utilisé dans les deux types d'applications. Par conséquent, afin de prendre en compte différentes sources de connaissances (dépendant de l'application) et de réaliser un système souple et extensible,

un modèle "blackboard" a été utilisé. Dans les prochaines sections, l'intérêt d'un système d'étiquetage automatique d'une base de données est discuté, et l'architecture du système proposé (sources de connaissances et contrôle) est présentée. Enfin, l'état actuel du système, qui est encore sous développement, est indiqué.

F.1.2 Etiquetage automatique d'une base de données

Dans les dernières années, d'importants progrès ont été réalisés en reconnaissance automatique de la parole. Un des buts principaux est d'étendre les systèmes de reconnaissance à des vocabulaires importants et à la parole continue. Dans ce cas, la parole est souvent manipulée à l'aide d'unités de base : phonèmes (e.g. [LZ84]), syllabes (e.g. [FHLS85]), classes grossières (e.g. [Foh88]), événements phonétiques (e.g. [PC85]). Avec les progrès de la technologie, des bases de données plus importantes sont traitées [Car84, LKS86, SKT87, PC87]. Dans le cadre de la reconnaissance de parole ou la synthèse à partir du texte, un nombre important de données segmentées et étiquetées est nécessaire. Traditionnellement, ces traitements ont été exécutés manuellement grâce à l'observation visuelle d'un certain nombre de paramètres : signal temporel, énergie, mouvements des formants, etc. Une telle méthode a plusieurs inconvénients, comme cela a déjà été rapporté dans la littérature [LZ84, Wag81] :

1. elle est très coûteuse en temps,
2. les décisions sont très subjectives et demandent une grande connaissance sur la parole pour une tâche qui est fastidieuse,
3. les décisions sont influencées par la connaissance à priori des chaînes phonétiques qui doivent théoriquement être trouvées. Ceci est souvent un obstacle à la prise en compte des mauvaises prononciations,
4. les résultats ne peuvent pas être reproduits.

Quelques tentatives pour réaliser des systèmes d'étiquetage automatique ont été effectuées [CB83, HCLR83, LZ84, MW86, Wag81, Hun84]. Ces essais ont tenté d'aligner la parole à étiqueter avec une référence segmentée manuellement (naturelle ou synthétique) en utilisant un algorithme de programmation dynamique [CB83, HCLR83, Hun84], de segmenter et d'étiqueter la parole en classes grossières avant le recalage temporel [LZ84, Wag81], ou d'utiliser des formes de référence obtenues par apprentissage [MW86]. Dans [LZ84], Leung et Zue utilisèrent des connaissances phonétiques afin de raffiner les marques de segmentation. Plus récemment des méthodes statistiques (appliquées à la segmentation automatique) [Obr88] ou à base de connaissances [Kat88] ont été présentées. Glass et Zue [GZ88] proposèrent une description à plusieurs niveaux, appelée dendrogram, pour représenter les résultats de la segmentation automatique. Citons enfin les travaux menés dans le cadre du *GRECO* "Communication Parlée" du C.N.R.S [GRE88a] et du projet *ESPRIT SAM* (en abréviation de "speech assessment methodologies") au niveau étiquetage automatique ou semi-automatique.

Dans les sections qui suivent, une méthode d'étiquetage automatique qui a été conçue autour d'un modèle "blackboard" [EL80, EHRLR80, Nii86b, Nii86a] et qui ne nécessite

pas d'unités de référence est proposée. Cette méthode peut être appliquée à l'étiquetage automatique d'une base de données mais aussi à la reconnaissance automatique de la parole. Le modèle "blackboard" a permis de développer le système modulairement, de le rendre facilement extensible et de faciliter la communication entre les différentes sources de connaissances. Dans le futur, il permettra de concevoir et de tester des stratégies de décision élaborées.

F.1.3 L'architecture "blackboard"

F.1.3.1 Présentation générale

Le système a été conçu avec l'idée que le processus d'étiquetage, afin d'être mené à bien, nécessite une coopération effective et efficace entre plusieurs sources de connaissances. Partant de cette hypothèse, le problème principal était la conception d'un système facilitant l'interaction entre les différentes sources de connaissances. Plusieurs paradigmes ont été proposés, en traitement de la parole, pour manipuler des sources de connaissances. Citons, par exemple, le modèle "blackboard" utilisé dans *HEARSAY II* [EL80, EHRLR80], le système à plans perceptuels [MLG87] et la société de spécialistes [GH88].

Dans le système proposé, le modèle "blackboard" a été choisi. L'architecture "blackboard" permet d'ajouter, de supprimer ou de modifier des sources de connaissances aisément et d'explorer diverses stratégies de décision. De plus, les sources de connaissances peuvent être développées et testées séparément. Le "blackboard" est un type d'architecture plus qu'un modèle particulier. Aussi, suivant les applications, l'interprétation qui a été faite du "blackboard" varie. La figure 36 montre un diagramme du modèle utilisé pour l'application considérée, à savoir l'étiquetage automatique.

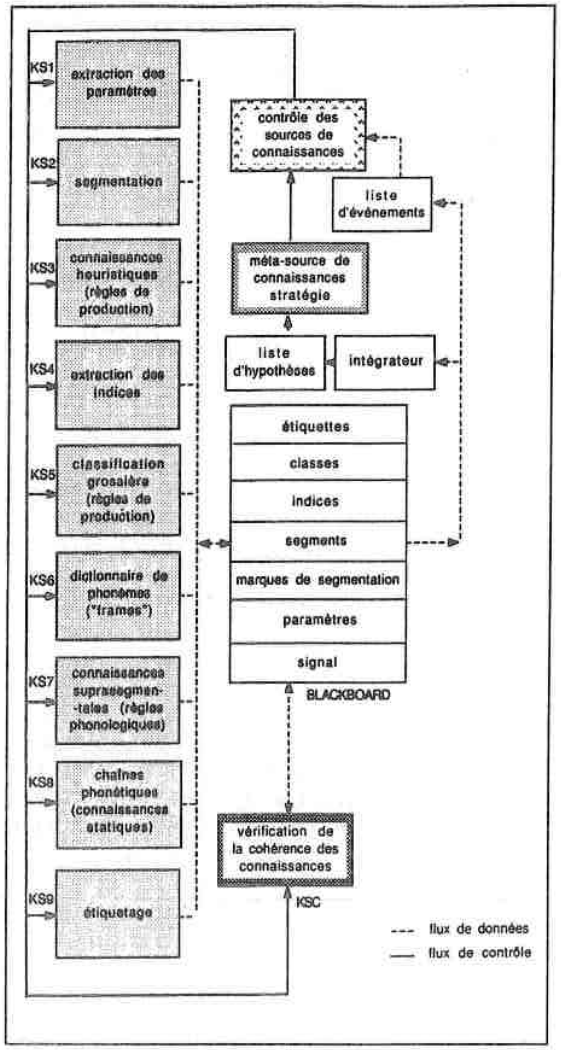


Figure 36 L'architecture "blackboard" du système d'étiquetage automatique.

Le système est organisé en trois niveaux comme dans *ATOME* [HMLM87, LMH88]. Les deux niveaux supérieurs (stratégie et contrôle des sources de connaissances) con-

stituent le contrôle hiérarchique du système. Le niveau *stratégie* analyse la situation courante et s'occupe de déterminer l'espace des hypothèses à explorer pour progresser dans la recherche de la solution (exprimé autrement, quelles sont les sources de connaissances à invoquer). Le rôle du *module de contrôle des sources de connaissances* est de synchroniser l'activation des sources de connaissances spécialistes (le troisième niveau). Chaque fois qu'une hypothèse susceptible de modifier la cohérence des informations contenues dans le "blackboard" est générée le *module de vérification de cohérence* vérifie que toutes les informations sont compatibles. En fonction des informations trouvées dans le "blackboard" une hypothèse peut être ajoutée ou supprimer.

Neuf sources de connaissances, qui communiquent par l'intermédiaire d'une zone commune : le "blackboard", sont utilisées. Dans le "blackboard" plusieurs niveaux d'abstraction sont définis : *signal, paramètres, marques de segmentation, segments, indices, classes, et étiquettes.* Chaque source de connaissances utilise un ou plusieurs niveaux d'entrée et de sortie dans le "blackboard". Par exemple, la source de connaissances *classification grossière* utilise les niveaux d'entrée *indices* et *segments* et le niveau de sortie *classes*. La figure 37 montre comment les niveaux du "blackboard" sont utilisés par les différentes sources de connaissances.

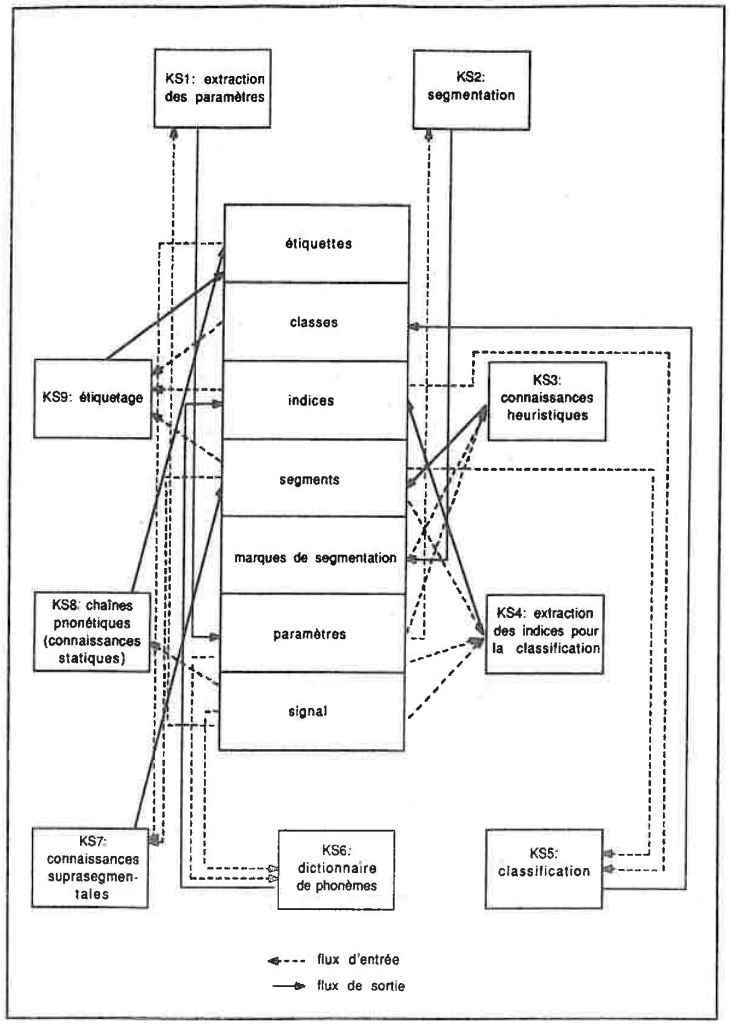


Figure 37 Schéma des communications entre les différentes sources de connaissances et les niveaux du "blackboard".

Enfin, dans le but d'améliorer l'efficacité du système, une *liste d'événements* (comme dans *ATOME*), dont le rôle est de mémoriser les événements générés par les sources de connaissances, et un *intégrateur* qui utilise une liste d'hypothèses pour résumer les

informations du "blackboard" dans des informations de plus haut niveau, ont été inclus dans le système. L'intégrateur est un filtre dont le rôle est de maintenir à jour la liste d'hypothèses qui contient des informations moins nombreuses mais souvent de plus haut niveau que celles du "blackboard".

F.1.3.2 La stratégie de contrôle

F.1.3.2.1 La méta-source de connaissances stratégie Le contrôle est composé de deux niveaux : le niveau *stratégie* et le niveau *contrôle des sources de connaissances*. La stratégie est le niveau le plus élevé. C'est en fait une méta-source de connaissances qui analyse l'état courant de la solution et détermine quelles sont les sources de connaissances à activer afin de poursuivre le processus. Cette source de connaissances contient, en particulier, un ensemble de règles de production. La partie gauche de ces règles prend ses informations dans la liste d'hypothèses. A ce niveau, les informations qui sont manipulées sont celles qui sont jugées importantes par le cogniticien afin de progresser dans l'élaboration de la solution. La partie droite des règles contient une ou plusieurs sources de connaissances à exécuter. Pour d'autres applications, comme la reconnaissance automatique de la parole où la chaîne phonétique correspondant au signal de parole n'est pas connue, la source de connaissances stratégie devra être modifiée.

F.1.3.2.2 Le contrôle des sources de connaissances Cette source de connaissances, qui correspond au niveau le plus bas du contrôle hiérarchique utilisé, est chargée de supprimer les conflits pouvant intervenir lors de l'exécution des sources de connaissances et de gérer l'ordre des appels. Dans une version future, elle prendra aussi en compte la gestion du parallélisme pouvant exister dans la recherche d'une solution. Le parallélisme peut être introduit au niveau de l'exécution des sources de connaissances mais aussi pour explorer en même temps les différentes formes allophoniques du mot à étiqueter (dans le cas où plusieurs variantes de la chaîne phonétique d'entrée sont utilisées).

Lorsque l'exécution d'une source de connaissances est terminée ou lorsque des résultats fournis par une autre source de connaissances sont nécessaires pour avancer dans la recherche de la solution, un événement est envoyé dans la liste des événements. A l'inverse de la source de connaissances stratégie, cette partie contrôle de bas niveau est dirigée par les événements.

F.1.3.3 Vérification de la cohérence des informations du "blackboard".

Afin de prendre en compte des données ambigues et incomplètes et de fusionner plusieurs sources d'informations tout en maintenant la cohérence de celles-ci, un module de vérification de la consistance des hypothèses émises a été introduit (source de connaissances *KSC*). Ce module est invoqué par le module stratégie lorsque des hypothèses, pouvant modifier la cohérence des informations, sont ajoutées dans le "blackboard". Le principal but de ce module est de faciliter la gestion de connaissances hypothétiques qui ne sont pas indépendantes. C'est en fait une version simpliée d'un système de maintenance de la vérité comme *TMS* [Doy79] ou *ATMS* [Kle86]. Ces systèmes fournissent des

moyens pour maintenir les dépendances entre données. Ils permettent les retours-arrière, en explorant les hypothèses les unes après les autres comme dans *TMS*, ou le parallélisme, en explorant les hypothèses en parallèle comme dans *ATMS*.

Le module de vérification de cohérence présenté est très primaire (à cause de la simplicité relative de l'application) comparé aux autres systèmes cités. En étiquetage automatique de la parole, une solution unique doit être trouvée. Ainsi, le système proposé se rapproche davantage de *TMS* que d'*ATMS*. En particulier, les retours-arrière et la suppression d'hypothèses sont permis. Au cours d'une autre application, comme la reconnaissance de parole, un autre mécanisme pourra être envisagé. Ce module, constitué d'un ensemble de règles de production, vérifie que les informations contenues dans le "blackboard" sont cohérentes mais permet aussi l'addition et la suppression d'hypothèses. Par exemple, la règle,

si (segment A=fricative)
et (prochain_segment=fricative)
et (il n'y a pas deux fricatives consécutives dans la chaîne phonétique en entrée)
et (le facteur de confiance de la frontière droite du segment A est faible)
alors (supprimer la frontière droite du segment A)

supprime une marque de segmentation provenant d'une sur-segmentation dans une fricative. De la même façon,

si (durée_segment A > durée_segment A' + C *)*
et (la chaîne phonétique en entrée contient l, r ou w)
et (une transition faible existe dans le segment A)
alors (ajouter une marque de segmentation dans le segment A)

** où A' représente le segment correspondant du segment A dans la chaîne phonétique d'entrée et C est une constante.*

ajoute une marque de segmentation dans l'ensemble des marques de segmentation, car une liquide ou une semi-consonne (l, r, w) n'a pas été détectée.

Lorsqu'une incohérence est trouvée par ce module, le contenu du "blackboard" est modifié. Par conséquent, la liste d'hypothèses est aussi changée pour indiquer à la méta-source de connaissances stratégie qu'un changement est intervenu. Ceci donne naissance à un nouveau mode de raisonnement prenant en compte les modifications effectuées par le module de vérification de la cohérence.

Enfin, ce module bénéficie d'un modèle d'incertitude développé durant les phases de segmentation et de classification. Ceci est réalisé par la prise en compte lors du raisonnement des facteurs de confiance associés aux instances des objets *marques de segmentation* et *classes*.

F.1.3.4 Le "blackboard" et les objets manipulés

Le "blackboard" est une zone commune qui contient les données produites et utilisées par les sources de connaissances. Ces données sont générées progressivement jusqu'à l'établissement d'une solution. Le contenu du "blackboard" est formé d'instances d'objets organisées hiérarchiquement en niveaux d'analyse.

Chaque objet est caractérisé par un ensemble de propriétés (ou attributs) et de relations (ou liens) avec d'autres objets. Le signal de parole est représenté par un arbre dont chaque noeud est associé à un niveau de représentation (paramètre, marque de segmentation, segment, etc). La figure 38 présente les différents niveaux de représentation.

Dans le système de compréhension de la parole *HEARSAY II*, le "blackboard" est divisé en niveaux qui sont liés aux représentations intermédiaires des processus de décodage (syllabe, mot, phrase). L'application qui est visée est beaucoup plus simple. Aussi, les niveaux du "blackboard" sont davantage liés à une représentation objet de la parole (jusqu'à l'objet étiquette) qu'à des niveaux intermédiaires de décodage.

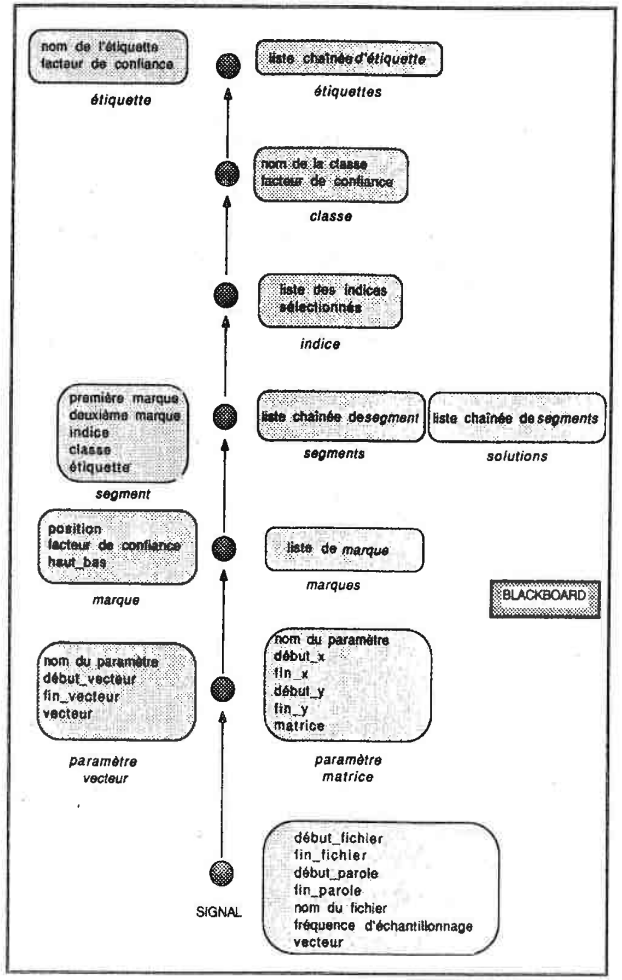


Figure 38 La représentation objet de la parole.

F.1.4 Les sources de connaissances spécialistes

F.1.4.1 Les paramètres de segmentation et les connaissances heuristiques

Pour segmenter un mot en unités élémentaires, une mesure de transition qui modèle des caractéristiques spectrales dynamiques (voir section D.4.2) a eté définie. Une première étape de segmentation automatique est effectuée à l'aide de la mesure de transition, le logarithme de l'énergie normalisée, et le taux de passages par zéro. Pendant cette phase de segmentation, les marques trouvées sont séparées en deux classes. Les marques qui appartiennent à la première classe sont détectées par des pics ou plateaux de la mesure de transistion qui ont une amplitude supérieure à un seuil (*SL*) défini expérimentalement. Elles représentent les marques de segmentation sur lesquelles le raisonnement va être effectué. La deuxième classe, comportant les autres pics et plateaux, représente des marques incertaines considérées comme des informations supplémentaires par le processus de recherche de la solution.

Un facteur de confiance, dérivé directement de la mesure de transition, est associé à chaque marque de segmentation. Il est fonction de l'amplitude de la transition (*val*) normalisée par l'amplitude de la plus importante transition dans le mot (*valmax*), la durée entre la marque de segmentation considérée et la précédente (loc_i - loc_{i-1}, plus les marques de segmentation sont rapprochées, plus elles sont suspectes) et du précédent minimum local. Ceci a conduit à la formule suivante :

$$\frac{val}{valmax} \times \frac{(loc_i - loc_{i-1}) \times \ val - min}{(loc_i - loc_{i-1}) \times \ val + min}. \qquad (46)$$

Pour les plateaux, qui aident à détecter les transitions lentes, la durée du plateau normalisée par une constante (*maxlength*) est utilisée. Si la durée du plateau est supérieure à la valeur de *maxlength* alors elle est prise égale à *maxlength*. Ceci a conduit à la formule suivante :

$$\frac{val}{valmax} \times \frac{length}{maxlength} \times \alpha \quad (0 < \alpha < 1). \qquad (47)$$

Le facteur de confiance est pondéré par la valeur de α (fixée à 0.5) afin de donner moins d'importance aux marques de segmentation trouvées à l'aide des plateaux par rapport à celles trouvées avec les pics de la courbe de transition. Enfin, pour les marques de segmentation insérées à l'aide des règles heuristiques (*KS3*), un facteur de confiance est associé à chaque règle. Dans tous les cas, ce facteur de confiance est compris entre 0 et 1. La figure 39 montre, pour le cas du mot "insert" produit par un locuteur féminin, la courbe de transition et le facteur de confiance accompagnant chaque marque de segmentation.

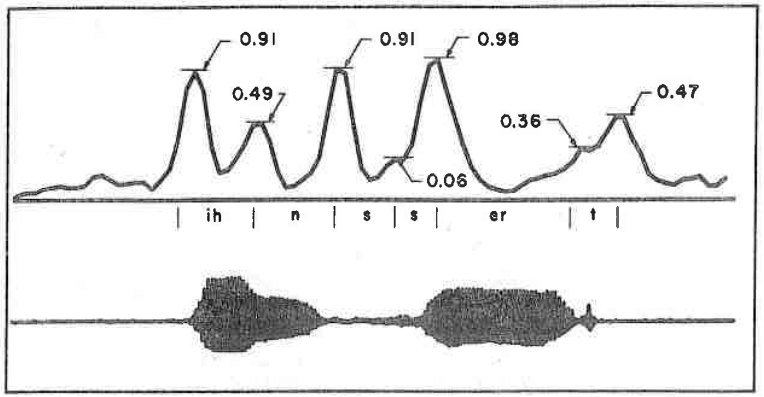


Figure 39 Signal temporel, courbe de transition et facteurs de confiance associés aux marques de segmentation du mot "insert" (locuteur féminin).

Notons que pour la marque de segmentation parasite se trouvant dans "s" le facteur de confiance est faible. Cette marque sera éliminée par les niveaux supérieurs lorsque d'autres informations confirmeront cette hypothèse.

Les algorithmes et les connaissances présentés dans cette section forment le contenu des sources de connaissances *KS1* et *KS2*. Afin de raffiner les marques de segmentation générées par *KS2*, la source de connaissances *KS3* (ensemble de règles de production) est utilisée.

F.1.4.2 Classification grossière à l'aide de connaissances phonétiques.

Grâce à l'examination des résultats fournis par la segmentation de base (voir la section D.4.4), les raisons liées à la sous-segmentation (environ 7%) et à la sur-segmentation (environ 22%) ont été identifiées. Il y a généralement sous-segmentation dans les liquides et les semi-consonnes (r, l et w) et dans les plosives où la segmentation automatique ne fait pas toujours la distinction entre la période de silence et le spectre de bruit suivant la barre d'explosion (en particulier, à la fin des mots lorsqu'il n'y a pas un relâchement net de la constriction du conduit vocal). De plus, les nasales à la fin des mots ont quelquefois des transitions lentes qui ne sont pas détectées. Quant à la sur-segmentation elle se trouve généralement dans les fricatives et les affriquées (s, f, ch) et dans les plosives qui ont une longue aspiration (k, p). A la fin des mots, du bruit provoqué par des clics ou des fluctuations dans le rythme d'élocution entraînent aussi quelquefois des marques de segmentation parasites. Etant donné la régularité des problèmes rencontrés, il a été décidé d'inclure dans le système des connaissances phonétiques dont le but est d'essayer d'éliminer quelques-unes des erreurs et de fournir des ilots de confiance pour les traitements suivants. Les segments de parole ont été classifiés en cinq catégories :

1. silence ou murmure nasal,

2. plosive,
3. bruit fricatif,
4. bruit (segments qui ne représentent pas de la parole),
5. autre.

Afin de réaliser cette classification, les indices suivants ont été extraits automatiquement (source de connaissances *KS4*) : présence d'une barre d'explosion avec ses caractéristiques, taux moyen et maximum de passages par zéro dans chaque segment, durée, énergie dans différentes bandes de fréquence, concentrations d'énergie, minimums de l'énergie du signal temporel. Les indices extraits ont ensuite été combinés à l'aide de règles (source de connaissances *KS5*) qui fournirent en sortie les classes indentifiées.

Cette classification grossière permet de diminuer le nombre de marques de segmentation parasites, apparaissant quelquefois dans les fricatives et les plosives, et de raffiner les frontières de mots.

F.1.4.3 Les autres sources de connaissances

Le *dictionnaire de phonèmes* (source de connaissances *KS6*) est une base de données qui est en développement. A chaque phonème pouvant apparaître dans la transcription phonétique du mot à segmenter sont associés divers attributs comme 1) des propriétés articulatoires distinctives, 2) la durée minimale et maximale, 3) le nom de la procédure permettant d'extraire des indices phonétiques distinctifs, etc. Ces connaissances sont représentées à l'aide de "frames". Une telle représentation facilite l'implantation de structures adaptées à l'évolution des connaissances et l'interaction entre des connaissances déclaratives et procédurales.

Les *connaissances suprasegmentales* (source de connaissances *KS7*) représentent des règles phonologiques qui fournissent, en particulier, les différentes formes allophoniques des mots qui sont à étiqueter. La figure 40 montre comment les différentes hypothèses, correspondant aux variations allophoniques d'un mot donné, sont générées.

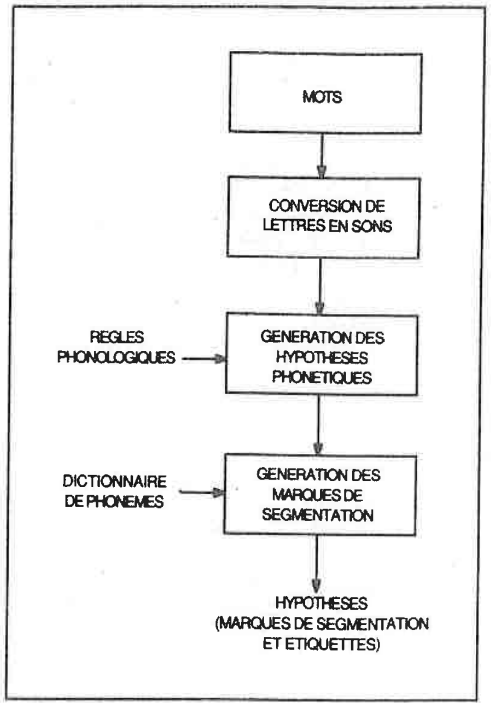


Figure 40 Diagramme de la génération des hypothèses correspondant aux variations allophoniques d'un mot donné.

Dans le cas de la segmentation automatique d'une base de données, la source de connaissances *KS8* contient les chaînes phonétiques des différents mots à étiqueter.

Enfin, le rôle de la source de connaissances *KS9* est d'étiqueter les segments trouvés automatiquement en utilisant, en particulier, les résultats de la classification grossière.

F.1.5 Evaluation partielle

Comme le mentionne la section D.4.4, les erreurs de segmentation sont difficiles à définir car elles dépendent du choix de la segmentation supposée correcte auquelle les résultats sont comparés. En reconnaissance automatique de la parole, une possibilité est de comparer la segmentation automatique à l'usage qui en est fait dans les niveaux supérieurs du système. Dans l'étude présentée, l'évaluation a été faite par comparaison à une segmentation manuelle de référence. Les résultats de cette évaluation ont été présentés à la section D.4.4. 93% des segments corrects ont été identifiés après un traitement utilisant les trois sources de connaissances *KS1*, *KS2* et *KS3*. La classification grossière, qui utilise les sources de connaissances *KS4* et *KS5*, a montré que plus de 90% des segments étaient correctement classifiés. Les autres sources de connaissances ainsi que la stratégie qui les manipule sont encore en développement. Aussi, aucune évaluation complète du système n'a été réalisée.

F.1.6 Résumé et conclusions

Un système d'étiquetage automatique de la parole utilisant un modèle perceptuel (*PLP*), des caractéristiques spectrales dynamiques et plusieurs autres sources de connaissances a été décrit. Comme aucune forme de référence n'est utilisée, ce système peut aussi être appliqué, en adaptant certaines sources de connaissances et la stratégie qui les manipule, à la reconnaissance automatique de la parole. Le passage d'une application à l'autre est facilité par la modularité offerte par le modèle "blackboard". Ce système d'étiquetage automatique de la parole comporte un certain nombre d'avantages comparé aux autres méthodes présentées dans la littérature :

1. il utilise une architecture modulaire et facilement extensible,
2. le spectre lissé fournit par l'analyse *PLP* limite les marques de segmentation parasites,
3. la mesure de transition développée est indépendante du locuteur,
4. le système n'utilise aucune forme de référence,
5. des connaissances phonétiques sont introduites seulement pour les cas difficiles.

Le but visé était de développer une architecture générique pouvant être appliquée à plusieurs applications liées au traitement de la parole. Par conséquent, lors de la conception, l'accent a été mis sur la souplesse et l'extensibilité. Le système utilise un contrôle hiérarchique à deux niveaux : stratégie et contrôle des sources de connaissances, ainsi que des sources de connaissances spécialistes. Actuellement, la partie contrôle et quelques-unes des sources de connaissances (*KS6*, *KS7* et *KS9*) sont encore en développement.

Le sage passe la seconde partie de sa vie à redresser les bêtises, les préjugés qu'il a acquis dans la première

Jonathan Swift

PARTIE G
CONCLUSIONS ET PERSPECTIVES

Après un résumé des points importants du travail présenté, les perspectives directement liées aux développements rapportés dans ce document sont discutées. Ensuite, de façon plus générale, certaines directions de recherche qui paraissent prometteuses en reconnaissance automatique de la parole sont évoquées.

Chapitre G.1 CONCLUSIONS

La nuit tombera bientôt. Les ombres annonciatrices de la nuit balaient la terre, balaient le continent comme si une grande main tirait le rideau pour que s'achève encore un jour. Bientôt, ce jour sera mort et, avec sa mort, naîtra un nouveau jour.
Toutes les créatures doivent, elles aussi, affronter le crépuscule de la vie avant que la nuit qui descend les balaie pour faire place à une autre vie encore à vivre, quelque part dans un quelconque monde ou sur un quelconque plan d'existence.

Extrait de "Crepuscule" Lobsang Rampa

La reconnaissance automatique de mots isolés est une tâche difficile qui n'a pas encore été totalement résolue. L'expérience montre que des systèmes de reconnaissance automatique de la parole qui donnent de bons résultats dans des environnements de laboratoire voient leurs performances se dégrader lorsqu'ils opèrent dans des conditions réelles. Ceci est lié à l'existence de nombreux problèmes souvent sous-estimés comme 1) la variabilité intra- et interlocuteur 2) l'existence de bruit de fond et 3) la difficulté des vocabulaires étudiés. Après une étude visant à déterminer le modèle d'analyse acoustique le moins sensible a la variabilité intra- et interlocuteur, les problèmes 2 et 3 ont été étudiés. Certaines des méthodes développées ont ensuite été appliquées à l'étiquetage automatique de la parole.

Dans ce travail, l'analyse par prédiction linéaire perceptivement fondée (*PLP*) a joué un rôle important. Grâce à une évaluation de plusieurs modèles utilisant l'analyse *PLP*, il a été montré que le modèle *PLP_RPS* donnait d'aussi bons ou de meilleurs scores de reconnaissance que d'autres modèles proposés récemment. Ensuite, un nouveau modèle, utilisant des caractéristiques spectrales dynamiques et une nouvelle mesure de distance bien adaptée à l'analyse *PLP* a été développé. Plus qu'un nouveau modèle, une méthodologie permettant d'examiner la réponse d'un modèle donné à certaines caractéristiques spectrales a été présentée. L'étude réalisée n'est pas spécifique au modèle *PLP* et peut être étendue à d'autres types de modèle.

Grâce à l'utilisation de bruit additif, le comportement de plusieurs modèles en présence de bruit a été étudié. Les résultats obtenus indiquent qu'aujourd'hui les systèmes de reconnaissance de parole sont très sensibles aux environnements bruités. Cependant,

il a été observé que lorsqu'il y a adéquation entre le modèle utilisé et le *SB* des mots de référence et de test, la dégradation en terme de scores de reconnaissance peut être minimisée. De plus, une série de tests a montré que la mesure de distorsion cepstrale projetée, qui donne les meilleurs résultats avec l'analyse *LP*, est profitable en présence de bruit additif.

Lorsque la parole est produite dans du bruit, il a été observé que la dégradation est beaucoup plus importante que lors des tests avec du bruit additif. Les techniques qui donnent de bons résultats pour du bruit additif (comme la mesure de distorsion cepstrale projetée) ne prennent pas en considération les changements phonétiques dus à l'effort vocal effectué, ce qui conduit à des performances très faibles. Dans un tel contexte, il a été montré qu'une approximation grossière du spectre de fréquence donne les meilleurs résultats.

Dans ce travail, le spectre lissé fourni par l'analyse *PLP* a conduit à de bons résultats. Par exemple, l'introduction dans la représentation paramétrique de caractéristiques spectrales dynamiques a été facilitée par ce specre lissé. De plus, le lissage déjà présent au niveau de l'analyse ne doit pas être effectué au niveau de la mesure de distance. Ces considérations amenèrent la proposition du "lifter" exponentiel ainsi que le développement du système de segmentation automatique *SAIPH*. Les résultats obtenus sont très encourageants.

Une approximation grossière du spectre de fréquence s'est avérée très profitable pour extraire des indices utilisés comme unités de reconnaissance. C'est une direction à poursuivre avec l'analyse *PLP*. En effet, ses possibilités d'intégration semblent meilleures que celles de l'analyse *LP*. La variabilité interlocuteur est ainsi moins importante.

Cependant, si dans cette étude les avantages de l'analyse *PLP* ont été mis à jour, certaines de ses limitations sont aussi apparues. En présence de bruit, même si les meilleures performances sont obtenues avec l'analyse *PLP*, celles-ci sont encore très faibles. De plus, une méthode utilisant l'analyse *PLP* et un algorithme de programmation dynamique ne permet pas d'effectuer la discrimination entre les mots du vocabulaire pour des mots acoustiquement similaires. Dans ce cas, le système dans son ensemble doit être revu. Ces deux points, mis en évidence dans le travail présenté, amenèrent le développement d'une nouvelle méthode d'analyse, *SLP*, qui utilise des concepts physiologiques et d'une méthode de discrimination qui tire partie de connaissances phonétiques.

Etant donné les résultats obtenus grâce à la simulation de phénomènes perceptuels dans la méthode *PLP*, mais aussi compte tenu des limitations de cette dernière, l'analyse *SLP* a été développée. L'analyse *PLP* est plus robuste que l'analyse *LP* à cause de l'intégration par banc de filtres critiques. Cependant, cela semble être aussi une de ses limitations. Une méthode fondée sur l'estimation du spectre de fréquence par une évaluation du synchronisme du taux de décharge des fibres nerveuses au niveau de la cochlée semble plus à même de résoudre les problèmes posés par le bruit qui est par nature aléatoire. Les résultats obtenus avec cette méthode sont déjà encourageants, même si celle-ci nécessite des améliorations (voir le prochain chapitre).

Pour aborder le problème des vocabulaires de mots acoustiquement similaires, un système hybride à deux passes, *ORION*, a été développé sur la base de l'utilisation de connaissances phonétiques pour faciliter la discrimination. Grâce à l'étude de spectrogrammes, des indices grossiers ont été identifiés et combinés en utilisant un système à règles de production afin de fournir des candidats potentiels. Une stratégie de décision a alors été utilisée pour combiner les candidats générés par les deux passes et fournir le mot reconnu. Le système développé s'est avéré très intéressant, tout d'abord par les résultats obtenus, mais aussi par le cadre de travail qu'il a fourni. Son développement modulaire a permis le test de divers algorithmes et l'interaction entre plusieurs modules conçus séparément.

En résumé, nous nous sommes intéressés à trois aspects de la reconnaissance de mots isolés :

1. la variabilité intra- et interlocuteur,
2. la robustesse des systèmes en présence de bruit.
3. les problèmes posés par les vocabulaires de mots acoustiquement similaires,

Les études se rapportant à ces problèmes et développées dans ce document ont mis en évidence les principaux points suivants :

1. grâce à une étude comparative de plusieurs modèles d'analyse acoustique, il a été montré que le modèle *PLP_RPS* donnait d'aussi bons ou de meilleurs résultats que les autres modèles étudiés,
2. la distance *RPS* , lorsqu'elle est utilisée avec le modèle *PLP*, est trop sensible aux pics spectraux et pas assez sensible à la pente spectrale globale,
3. une nouvelle distance, qui pondère chaque coefficient cepstral par une puissance de son index, atténue les problèmes liés à la distance *RPS*,
4. des caractéristiques spectrales dynamiques sont particulièrement intéressantes avec le modèle *PLP* dans le cadre de vocabulaires comprenant des mots acoustiquement similaires.
5. la combinaison de la nouvelle mesure de distance et de caractéristiques spectrales dynamiques avec le modèle *PLP* diminue le taux d'erreur de reconnaissance d'environ 10% (en reconnaissance multilocuteur) par rapport au modèle *PLP_RPS*,
6. en présence de bruit, l'ordre du modèle doit être plus élevé qu'en milieu non-bruité et la sensibilité de la mesure de distance doit être adaptée au *SB* des mots de test et de référence,
7. la mesure de distorsion cepstrale projetée, qui donne les meilleurs résultats avec l'analyse *LP*, est profitable en présence de bruit additif,
8. l'effet Lombard dégrade davantage les performances que le bruit additif et les techniques qui donnent de bons résultats en présence de bruit additif conduisent à des performances très faibles lorsque la parole est produite dans du bruit à cause des changements phonétiques occasionnés. L'effet Lombard doit donc être étudié en priorité,

9. en présence de l'effet Lombard, une approximation grossière du spectre de fréquence est souhaitable,
10. une nouvelle méthode d'analyse, *SLP*, fondée sur des concepts physiologiques a été développée. Celle-ci a donné de meilleurs scores de reconnaissance que les autres modèles étudiés (sous certaines conditions) pour un faible *SB*,
11. un système hybride à deux passes utilisant plusieurs sources de connaissances a été proposé afin de faciliter la discrimination entre des mots acoustiquement similaires. Ce système a permis de diminuer de plus de 60% le taux d'erreur de reconnaissance obtenu lors de la première passe.

Enfin, une extension du système de segmentation automatique, *SAIPH*, a été appliquée à l'étiquetage automatique. Ce système se distingue par sa grande modularité, sa souplesse et son architecture évoluée. Bien qu'il soit toujours en développement, les idées qui ont menées à sa conception semblent prometteuses.

Dans ce travail, des connaissances sur la parole ont été prises en compte sans pour autant négliger notre ignorance. Le système *ORION*, mais aussi le développement du modèle fonctionnel *SLP* et la méthode d'étiquetage automatique, où des connaissances phonétiques ont été prises en compte uniquement pour traiter les cas difficiles, illustrent cette approche. Les systèmes de reconnaissance de parole doivent être adaptés à notre connaissance sur la parole et inversement. Ils doivent aussi être flexibles et extensibles afin de pouvoir intégrer facilement l'évolution de notre savoir.

Ce travail souligne quelques problèmes importants qui nécessitent d'être étudiés afin d'améliorer la robustesse des systèmes de reconnaissance automatique de mots isolés. Il propose aussi des solutions tout en suggérant des idées pour des recherches futures. Le prochain chapitre indique tout d'abord quels sont les travaux qui doivent être accomplis afin de compléter l'étude présentée et ensuite discute les axes de recherche qui, à la lumière de cette étude, paraissent importants et prometteurs.

Chapitre G.2 PERSPECTIVES

L'esprit humain est un grenier tellement impossible à remplir que, du point de vue de la connaissance, il représente un abîme.

Jan Amos Coménius

G.2.1 Perspectives directement en relation avec les études présentées

Comme il est dit dans les avant-propos, encore beaucoup d'efforts restent à faire afin de compléter certains des travaux présentés. En effet, comme cela a été explicité tout au long de ce document, nos travaux ont amené des solutions, des améliorations, mais aussi de nouvelles questions dont certaines sont exprimées ci-après.

Au niveau techniques d'analyse, nous avons vu que la méthode *PLP* donnait de bonnes performances. Cependant, l'analyse *PLP* effectue une intégration en bandes critiques après une transformée de Fourier afin d'obtenir le spectre auditif. L'utilisation d'un banc de filtres non-simulé permettrait d'obtenir une bonne résolution spectrale aux basses fréquences et une bonne résolution temporelle aux hautes fréquences (comme l'oreille humaine). Il semble souhaitable d'intégrer cette modification dans l'analyse *PLP*.

Afin d'obtenir un modèle d'analyse moins sensible au bruit de fond, la technique *SLP* a été développée. Même si les résultats obtenus sont encourageants, de nombreuses améliorations restent possibles. Le spectre de fréquence généré par cette nouvelle méthode n'est pas assez lisse. Ceci est une conséquence de l'estimation du spectre de fréquence par une technique fondée sur le comptage des passages par zéro. L'augmentation des scores de reconnaissance grâce à l'introduction d'un filtre passe-bas confirme cette hypothèse. De plus, d'autres mécanismes comme le masquage postérieur ou l'adaptation à court terme méritent d'être introduits afin de prendre en compte les caractéristiques dynamiques de la parole. Enfin, de bonnes performances sont obtenues lorsqu'il y a adéquation entre la technique d'analyse et la mesure de distance. Par conséquent, une nouvelle mesure de distance bien adaptée à la technique *SLP* doit être étudiée.

L'étude de l'effet Lombard a permis de mesurer l'incidence qu'il a sur les scores de reconnaissance. La prochaine étape, qui a déjà été initialisée, consiste à acquérir la connaissance sur les changements phonétiques qui interviennent lorsque la parole est produite dans du bruit. Ainsi, il sera possible de corréler les performances du modèle d'analyse acoustique aux phénomènes observés sur la base de données étudiée. Il est très important d'avoir une meilleure compréhension des facteurs qui dégradent les

performances des systèmes de reconnaissance. Ainsi, les modèles d'analyse acoustique pourront être modifiés en tenant compte des perturbations dues à l'effet Lombard. Dans le même ordre d'idée, Hattori et Yoshida [HY88] proposèrent une méthode de normalisation des voyelles sur la base d'une étude des spectres de voyelles produites en présence de bruit. L'étude de la variabilité occasionnée par le bruit et le développement de modèles moins sensibles à cette variabilité font partie des travaux en cours.

En ce qui concerne le système hybride *ORION*, la discrimination doit être étendue à d'autres classes que celle étudiée (*E-SET*). De plus, les facteurs de confiance associés aux indices extraits doivent être parfaitement intégrés dans le système.

Le système d'étiquetage automatique présenté est encore en développement. L'implantation des mécanismes de contrôle et la constitution de plusieurs sources de connaissances (dictionnaire de phonèmes, règles phonologiques, etc) doivent être terminées. Ce système, qui à l'origine était écrit en langage *C*, est présentement réécrit dans le langage orienté objet *EIFFEL*. Ce langage est bien adapté à la représentation objet choisie pour le système. Ensuite, les performances du système doivent être évaluées et comparées à d'autres systèmes d'étiquetage automatique comme *APHODEX* réalisé au C.R.I.N. L'un des buts du système est de faciliter la coopération entre différentes sources de connaissances et d'intégrer notre savoir de façon incrémentale. Aussi, les travaux futurs vont porter sur l'amélioration et l'augmentation des connaissances utilisées par le système. En particulier, le développement d'un modèle de durée est une extension possible. Enfin, ce système doit être appliqué à la reconnaissance automatique de la parole.

G.2.2 Perspectives générales

Tout au long de ce travail quelques outils, ayant pour but de faciliter les recherches effectuées, ont été développés (*ORION, STAR, SAIPH*). A cause de la variété des travaux réalisés mais aussi des liens qui existaient entre eux, un environnement de recherche intégré, facilitant l'évaluation des méthodes développées et le développement de nouveaux algorithmes, s'est fait énormément sentir. Actuellement, des environnements de travail, généralement conçus pour des applications de traitement de signal, sont disponibles mais ils sont souvent de bas niveau et difficiles à personnaliser sans l'aide du concepteur. Des outils de recherche qui combineraient les propriétés : souplesse, réutilisabilité et convivialité sont nécessaires. Plus qu'une collection de programmes d'analyse (comme *ILS*), ce sont des méthodes de conception de tels environnements qui sont à étudier. Dans le passé, ceci était difficile à imaginer à cause des supports de programmation disponibles. Aujourd'hui, avec l'apparition des langages orientés objets, de standards comme Xwindow, Postscript, etc... et les progrès réalisés au niveau des interfaces utilisateurs, il doit être possible de développer des outils qui soient plus souples et mieux adaptés à des environnements de recherche.

Les *premières étapes* d'un système de traitement de la parole sont très importantes. Ce sont, par exemple, la méthode d'analyse et la mesure de distance pour un système de reconnaissance de mots isolés ou le système de décodage acoustico-phonétique d'un système de reconnaissance de parole continue. Tous les traitements ultérieurs dépendent

de ces premières étapes. Aussi, elles doivent s'attacher à préserver l'information contenue dans le signal de parole. Nous avons vu que les modèles auditifs avaient récemment soulevé beaucoup d'enthousiasme. En effet, les êtres humains sont les meilleurs systèmes de décodage de messages que nous connaissons. C'est la raison pour laquelle la modélisation du système auditif humain a suscité autant d'intérêt. Même si les performances obtenues à l'heure actuelle par les modèles auditifs ne sont pas systématiquement meilleures que celles obtenues par des méthodes plus traditionnelles, les résultats qu'ils fournissent sont encourageants. Notre connaissance du système auditif périphérique doit progresser afin de réaliser des modèles fonctionnels qui soient performants. Les techniques d'analyse de la parole sont directement liées à la suppression du bruit, l'extraction automatique d'indices acoustiques et beaucoup d'autres traitements de niveau supérieur. Aussi, le développement d'une méthode d'analyse capable de préserver l'information de parole, robuste et efficace est une voie de recherche dans laquelle il faut persévérer.

Enfin, comme notre connaissance sur les mécanismes de la parole n'est encore pas très importante et que la parole est un phénomène complexe, il est souvent nécessaire de combiner plusieurs sources d'informations afin de résoudre un problème donné. Ceci incite le développement d'architectures adaptées au partage d'informations et à la coopération entre plusieurs sources de connaissances. De plus, comme notre connaissance ne s'exprime pas en terme de vrai ou de faux, des modèles d'incertitude doivent être développés. On entre dans le domaine de l'ingénieur cogniticien. Si l'on se réfère à la définition de Negoita et Ralescu [NR87] "*knowledge engineering is concerned with the problem of managing an accumulated body of truths and facts integrated into computer systems to solve complex problems normally requiring a high level of human expertise*". L'intégration de connaissances liées à différents domaines est un axe de recherche privilégié. Les problèmes posés ont trait au raisonnement hypothétique, à la vérification de la cohérence des informations manipulées et plus généralement impliquent l'utilisation de techniques d'intelligence artificielle. Le système d'étiquetage automatique qui a été présenté est fondé sur ces idées.

Cette section a présenté quelques directions de recherche qui, à la vue du travail réalisé, sont apparues comme très importantes. De notre point de vue, il est essentiel de poursuivre nos efforts dans cette direction. Certainement, cela ne veut pas dire que les autres domaines de recherche liés à la reconnaissance de parole sont sans intérêt. Notre but est seulement de mettre l'accent sur certains points particuliers qui, à notre avis, sont de première importance pour les travaux futurs.

Bibliographie

Un classique est quelque chose
que tout le monde veut avoir lu
et que personne ne veut lire.

Extrait de la Sagesse de Mark Twain

[AH71]B.S. Atal and S.L. Hanauer. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *J. Acoust. Soc. Am.*, 50:637–655, August 1971.

[AHW87]T.H. Applebaum, B.A. Hanson, and H. Wakita. Weighted Cepstral Distance Measures in Vector Quantization Based Speech Recognizers. In *ICASSP-87*, 1987.

[Ali73]P. Alinat. Reconnaissance des Phonèmes au Moyen d'une Cochlée Artificielle, 1973. Thèse de Docteur Ingénieur.

[All85]J.B. Allen. Cochlear Modeling. *IEEE ASSP Magazine*, pages 3–29, 1985.

[AR77]H.L. Alder and E.B. Roessler, editors. *Introduction to Probability and Statistics*. W. H. Freeman and Company, 1977.

[AS68]B.S. Atal and M.R. Schroeder. Predictive Coding of Speech Signals. In *6th International Congress on Acoustic, Tokyo*, pages 21–28, August 1968.

[ASS85]K. Aikawa, M. Sugiyama, and K. Shikano. Spoken Word Recognition Based on Top-Down Phoneme Segmentation. In *ICASSP-85*, pages 33–36, 1985.

[Ata74]B.S. Atal. Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification. *J. Acoust. Soc. Amer.*, 55:1304–1312, June 1974.

[Bak75]J.K. Baker. Stochastic Modeling for Automatic Speech understanding. In R. Reddy , editor, *Speech Recognition*, pages 521–542. Academic Press, Inc, 1975.

[BBM88]C. Bourjot, A. Boyer, and J.F. Mari. Methodology about Assessment of Large Vocabulary Systems. In *7th FASE Symposium, Edinburgh*, pages 161–169, 1988.

[BCEG84]M. Blomberg, R. Carlson, K. Elenius, and B. Granström. Auditory Models in Isolated Word Recognition. In *ICASSP-84*, pages 17.9.1–17.9.4, 1984.

[BCL82]G.L. Bradshaw, R. Cole, and Z. Li. A Comparison of Learning Techniques in Speech Recognition. In *ICASSP-82*, pages 554–557, 1982.

[BE67]L.E. Baum and J.A. Eagon. An Inequality with Applications to Statistical Estimations for Probabilistic Functions of Markov Processes and to a Model of Ecology. *Amer. Math Soc. Bulletin*, 73:360–362, 1967.

[Bér85]D. Béroule. Un Modèle de Mémoire Adaptative, Dynamique et Associative pour le Traitement Automatique de la Parole. Université de Paris-Sud, Mai 1985. Thèse d'Université.

[BGGM80]A.A. Buzo, H. Gray, Jr.R.M. Gray, and J.D. Markel. Speech Coding Based on Vector Quantization. *IEEE Trans. ASSP-28*, (5):562–574, 1980.

[Bla83]A. Bladon. Two-Formant Models of Vowel Perception: Shortcomings and Enhancements. *Speech Communication*, 2:305–313, 1983.

[Bla85]A. Bladon. Acoustic Phonetics, Auditory Phonetics, Speaker Sex and Speech Recognition: A Thread. In F. Fallside and W. A. Woods, editor, *Computer Speech Processing*, pages 29–39. Prentice Hall International, 1985.

[Bla87]A. Bladon. The Auditory Modelling Dilemma, and a Phonetic Response. In *11th ICphS Tallinn*, 1987.

[BMM88]Z.S. Bond, T.J. Moore, and K. McCreight. Some Phonetic Characteristics of Sentences Produced in Noise. In *J. Acoust. Soc. Am.*, number 83 S1, page S67, Seattle 1988.

[Bol70]Bolt et al. Speaker Identification by Speech Spectrograms: Some Further Observations. *J. Acoust. Soc. Am.*, 47(2):531–534, February 1970.

[Boy87]A. Boyer. Application des Techniques de Programmation Dynamique et de Quantification Vectorielle à la Reconnaissance des Mots Isolés et des Mots Enchaînés. Université de Nancy I, Avril 1987. Thèse d'Université.

[Bri86]G. Bristow, editor. *Electronic Speech Recognition: Techniques, Technology, and Applications*. McGraw-Hill, 1986.

[Bro86]D.J. Broad. Vowels in Context: Dynamics, Statistics, and Recognition. In W.A. Lea, editor, *Towards Robustness in Speech Recognition*. To appear in Apple Valley, Mn, Speech Science Publications, 1986.

[BS85]B.G. Buchanan and E.H. Shortliffe. *Rule-Based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, 1985.

[BW87]H. Bourlard and W.J. Wellekens. Multilayer Perceptrons and Automatic Speech Recognition. In *1st International Conference on Neural Networks, San Diego CA*, volume 4, pages 407–416, 1987.

[BW88]H. Bourlard and W.J. Wellekens. Speech Pattern Discrimination and Multilayer Perceptrons. *To appear in Computer, Speech and Language*, 1988.

[Cad82]J.A. Cadzow. ARMA Modeling of Time Series. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (2), March 1982.

[Cae79]J. Caelen. Un modèle d'Oreille; Analyse de la Parole Continue; Reconnaissance Phonémique. Université Paul Sabatier de Toulouse, 1979. Thèse d'Etat.

[Cae85]J. Caelen. Space/Time Data-Information in the ARIAL-Project Ear Model. *Speech Communication*, 4:163–180, 1985.

[Car84]Carre et al. The French Language Database: Defining, Planning, and Recording a Large Database. In *ICASSP-84*, pages 42.10.1–42.10.4, 1984.

[Car86]Carbonell et al. APHODEX, Design and Implementation of an Acoustic-Phonetic decoding Expert System. In *ICASSP-86*, pages 1201–1204, 1986.

[CB83]R.M. Chamberlain and J.S. Bridle. Zip: a dynamic programming algorithm for time-aligning two indefinitely long utterances. In *ICASSP-83*, pages 816–819, 1983.

[CFG75]R. Carlson, G. Fant, and B. Granström. Two-Formant Models, Pitch and Vowel Perception. In G. S. Fant and M .A .A. Tatham, editor, *Auditory Analysis and Perception of Speech*, pages 55–82. Academic Press, New York and London, 1975.

[CG82]R. Carlson and B. Granström. Towards an Auditory Spectrogram. In R. Carlson and B. Granström , editor, *The Representation of Speech in the Peripheral Auditory System*, pages 109–114. Elsevier Biomedical Press, 1982.

[CGF70]R. Carlson, B. Granström, and G. Fant. Some Studies Concerning Perception of Isolated Vowels. Technical Report STL-QPRS 2-3, Royal Institute of Technology, Stockholm, 1970.

[Chi82]Chistovich et al. Temporal Processing of Peripheral Auditory Patterns of Speech. In R. Carlson and B. Granström, editors, *The Representation of Speech in the Peripheral Auditory System*, pages 165–180. Elsevier Biomedical Press, Amsterdam, 1982.

[Chu87]K.W. Church. *Phonological Parsing in Speech Recognition*. Kluwer Academic, 1987.

[CM81]R.V. Cox and D. Malah. A Technique for Perceptually Reducing Periodically Structured Noise in Speech. In *ICASSP-81*, pages 1089–1092, 1981.

[CM82]P.L. Chu and D.G. Messerschmitt. A Frequency Weighted Itakura-Saito Spectral Distance Measure. *IEEE ASSP-30*, (30), 1982.

[Coh85]J.R. Cohen. Application of an Adaptive Auditory Model to Speech Recognition. In *Montreal Workshop on Speech Recognition*, 1985.

[Cok76]C.H. Coker. A Model of Articulatory Dynamics and Control. *IEEE Trans. ASSP*, 64(4):453–460, April 1976.

[Col83]Cole et al. Feature-Based Speaker-Independent Recognition of Isolated English Letters. In *ICASSP-84*, pages 731–733, 1983.

[CRZR80]R.A. Cole, R. Rudnicky, V.W. Zue, and D.R. Reddy. Speech as Patterns on Paper. pages 3–50. Lawrence Erlbaum Associates , 1980.

[CSL78]L.A. Chistovich, R.L. Sheikin, and V.V. Lublinskaja. 'Centers of Gravity' and Spectral Peaks as the Determinants of vowel Quality. In B. Lindblom and S. Ohman, editor, *Frontiers of Speech Communication Research*, pages 145–157. Academic Press, New York and London, 1978.

[CSL85]R.A. Cole, R.M. Stern, and M.J. Lasry. Performing Fine Phonetic Distinctions. In J.S. Perkell and D.H. Klatt, editors, *Variability and Invariance in Speech Processes*, pages 325–345. Hillsdale, NJ, Lawrence Erlbaum Assoc, 1985.

[CV88]F. Casacuberta and E. Vidal. Speech Recognition with Difficult Vocabularies. In H. Niemann et al., editor, *Recent Advances in Speech Understanding and Dialog Systems*, pages 279–283. Springer-Verlag Berlin Heidelberg, 1988.

[CZ80]R.A. Cole and V.W. Zue. Speech as Eyes See it. pages 475–494. Lawrence Erlbaum Associates , 1980.

[Dal72]Dallos et al. Cochlear Inner and Outer Hair Cells: Functional Differences. *Science*, 177:356–358, 1972.

[Del82]B. Delgutte. Some Correlates of Phonetic Distinctions at the Level of the Auditory Nerve. In R. Carlson and B. Granström, editors, *The representation of Speech in the Peripheral Auditory System*, pages 131–149. Elsevier Biomedical Press, 1982.

[Del84]B. Delgutte. Codage de la Parole dans le Nerf Auditif, 1984. Thèse de Doctorat d'Etat.

[Del86]B. Delgutte. Comment on the Use of Peripheral Auditory Models in Speech Recognition. In J. S. Perkell and D. H. Klatt, editor, *Variance and Variability in Speech Processes*, pages 320–323. Lawrence Erlbaum Associates, 1986.

[DGG88]L. Deng, C.D. Geisler, and S. Greenberg. A Composite Model of the Auditory Periphery for the Processing of Speech. *Journal of Phonetics*, 16:93–108, 1988.

[DH73]R.O. Duda and P.E. Hart. Wiley-Interscience, New York, 1973.

[DM80]S.B. Davis and P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. ASSP*, 28(4):357–366, 1980.

[Dol80]J.M. Dolmazon. Contribution aux Recherches sur l'Appareil Auditif - Elaboration et Exploitation d'un Modèle du Fonctionnement du Système Auditif Périphérique, 1980. Thèse d'Etat.

[Dol82]J.M. Dolmazon. Representation of Speech-like Sounds in the Peripheral Auditory System in Light of a Model. In R. Carlson and B. Granström, editors, *The Representation of Speech in the Peripheral Auditory System*, pages 151–164. Elsevier Biomedical Press, Amsterdam, 1982.

[Doy79]J. Doyle. A Truth Maintenance System. *Artificial Intelligence*, 12:231–272, 1979.

[DS77]N.R. Dixon and H.F. Silverman. The 1976 Modular Acoustic Processor (MAP). *IEEE Trans. ASSP-25*, (5):367–379, 1977.

[EB82]K. Elenius and M. Blomberg. Effects of Emphasizing Transitional or Stationary Parts of the Speech Signal in a Discrete Utterance Recognition System. In *ICASSP-82*, pages 535–538, 1982.

[EHRLR80]L.D. Erman, F. Hayes-Roth, V.R. Lesser, and D.R. Reddy. The HEARSAY-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *Computing Surveys*, 12(2):213–253, 1980.

[EL80]L.D. Erman and V.R. Lesser. The HEARSAY-II Speech Understanding System. In W. LEA, editor, *Trends in Speech Recognition*, pages 361–381. Prentice-Hall, 1980.

[Ell62]L.L. Elliot. Backward and Forward Masking of Probe Tones of Different Frequencies. *J. Acoust. Soc. Am.*, 34:1116–1117, 1962.

[EMJ88]Y. Ephraim, D. Malah, and B.H. Juang. On the Application of Hidden Markov Models for Enhancing Noisy Speech. In *ICASSP-88*, pages 533–536, 1988.

[Eva75]E. Evans. The Sharpening of Cochlear Frequency Selectivity in the Normal and Abnormal Cochlea. *Audiology*, 14:419–442, 1975.

[EWR87]Y. Ephraim, J.G. Wilpon, and L.R. Rabiner. A Linear Predictive Front-End Processor for Speech Recognition in Noisy Environments. In *ICASSP-87*, pages 1324–1327, 1987.

[FA86]S. Furui and M. Akagi. On the Role of Spectral Transition in Phoneme Perception and its Modeling. In *ICA 12*, pages A2–6, 1986.

[Fan73]G. Fant. *Speech Sounds and Features*. MIT Press, Cambridge, MA, 1973.

[FHLS85]D. Fohr, J.P. Haton, F. Lonchamp, and L. Sauter. Méthodes de Segmentation Syllabique en Reconnaissance de la Parole. In *XIV JEP PARIS*, pages 164–167, 1985.

[Fla55a]J.L. Flanagan. A Difference Limen for Vowel Formant Frequency. *J. Acoust. Soc. Amer.*, 27:613–716, 1955.

[Fla55b]J.L. Flanagan. Difference Limen for the Intensity of a Vowel Sound. *J. Acoust. Soc. Amer*, 27:1223–1225, 1955.

[Fla72]J.L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, 2nd ed. New York, 1972.

[Fle40]H. Fletcher. Auditory Patterns. *Review of Modern Physics*, pages 47–65, 1940.

[FM33]H. Fletcher and W.A. Munson. Loudness, its Definition, Measurement, and Calculation. *J. Acoust. Soc. Am.*, 5:82–108, 1933.

[Foh86]D. Fohr. APHODEX: Un Système Expert en Décodage Acoustico-Phonétique de la Parole Continue. Université de Nancy I, Mars 1986. Thèse d'Université.

[Foh88]Fohr et al. Paramétrisation Acoustique et Décodage Phonétique Fondé sur des Connaissances, pour la Parole Continue Multilocuteur. In GRECO "Communication Parlée du C.N.R.S, editor, *Décodage Acoustico-Phonétique*. 1988.

[Fon84]P. Fonsale. Feature-Based Speaker-Independent Word Recognition without Oral Learning. In *ICASSP-84*, pages 17.7.1–17.7.4, 1984.

[Fur81]S. Furui. Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Trans. ASSP-29*, pages 254–272, April 1981.

[Fur86a]S. Furui. On the role of spectral transition for speech perception. *J. Acoust. Soc. Am.*, (80):1016–1025, 1986.

[Fur86b]S. Furui. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. ASSP*, 34:52–59, 1986.

[GH88]Y. Gong and J.P. Haton. A specialist society for continuous speech understanding. In *ICASSP-88*, pages 627–630, 1988.

[Ghi86]O. Ghitza. Auditory Nerve Representation as a Front-End for Speech Recognition in a Noisy Environment. *Computer Speech and Language*, 1986(1):109–130, 1986.

[Ghi87]O. Ghitza. Robustness Against Noise: The Role of Timing-Synchrony Measurement. In *ICASSP-87*, pages 2372–2375, 1987.

[Ghi88]O. Ghitza. Auditory Neural Feedback as a Basis for Speech Processing. In *ICASSP-88*, pages 91–94, 1988.

[Gil84]Gillet et al. SERAC: Un Système Expert en Reconnaissance Acoustico-Phonétique. In *4 ième Congrès AFCET-RFIA*, 1984.

[GLM84]V. Gupta, M. Lennig, and P. Mermelstein. Decision Rules for Speaker-Independent Isolated Words Recognition. In *ICASSP-84*, pages 9.2.1–9.2.4, 1984.

[GM76]A.H. Gray and J.D. Markel. Distance Measures for Speech Processing. *IEEE Trans. ASSP-24*, (5):380–391, 1976.

[GM87]Y. Gu and J.S.D. Mason. Vocal Tract and Auditory Feature Analysis Using Chinese Utterance in ASR System. In *Int. Conf. on Chinese Inf. Processing*, 1987.

[GRE88a]GRECO "Communication Parlée du C.N.R.S, editor. *Décodage Acoustico-Phonétique*. Septembre 1988.

[Gre88b]S. Greenberg. A Special Issue on the "Representation of Speech in the Auditory Periphery". *Journal of Phonetics*, 15(4), January 1988.

[Gre88c]S. Greenberg. The Ear as a Speech Analyzer". *Journal of Phonetics*, 15(4):139–149, January 1988.

[GZ88]J.R. Glass and V. Zue. Multi-level acoustic segmentation of continuous speech. In *ICASSP-88*, pages 429–432, 1988.

[HCLR83]H.D. Hohne, C. Coker, S.E. Levinson, and L.R. Rabiner. On temporal alignment of sentences of natural and synthetic speech. *IEEE ASSP-31*, pages 807–813, 1983.

[HD79]D.M. Harris and P. Dallos. Forward Masking of Auditory Nerve Fiber Responses. *Journal of Neurophysiology*, 42:1083–1107, 1979.

[Her87]H. Hermansky. An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory perception. In *ICASSP-87*, pages 1159–1162, 1987.

[HHW85a]B.A. Hanson, H. Hermansky, and H. Wakita. Root-power sums and spectral slope distortion measures for all-pole models of speech. In *J. Acoust. Soc. Am.*, number 78 S1, page S49, 1985.

[HHW85b]H. Hermansky, B.A. Hanson, and H. Wakita. Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain. *Speech Communication*, (4):181–187, 1985.

[HJ88]H. Hermansky and J.C. Junqua. Optimization of perceptually-based ASR front-end. In *ICASSP-88*, pages 219–222, 1988.

[HL86]M.J. Hunt and C. Lefebvre. Speech Recognition Using a Cochlear Model. In *ICASSP-86*, pages 1979–1982, 1986.

[HL87]W.Y. Huang and R.P. Lippmann. Neural Networks and Traditional Classifiers. In *Conference on Neural Information Processing Systems, Boulder CO- Natural and Synthetic, IEEE*, 1987.

[HL88]M.J. Hunt and C. Lefebvre. Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model. In *ICASSP-88*, pages 215–218, 1988.

[HMLM87]J.P. Haton, B. Maître, H. Laasri, and T. Mondot. ATOME: Another TOol for Developing Multi-Expert Systems. In *Workshop on Blackboard Systems: AAAI-87*, Seattle Wa, July 13 1987.

[Hou72]T. Houtgast. Psychophysical Evidence for Lateral Inhibition in Hearing. *J. Acoust. Soc. Am.*, 51(6.2):1885–1894, 1972.

[Hun84]M.J. Hunt. Time alignment of natural speech to synthetic speech. In *ICASSP-84*, pages 2.5.1–2.5.4, 1984.

[HW86]B.A. Hanson and H. Wakita. Spectral Slope Based Distortion Measures for All-Pole Models of Speech. In *ICASSP-86*, pages 757–780, 1986.

[HY88]H. Hattori and K. Yoshida. Recognition of Speech Produced in a Simulated Noisy Environment. In *J. Acoust. Soc. Am.*, volume 84 S1, Honolulu 1988.

[IS68]F. Itakura and S. Saito. Analysis Synthesis Thelephony Based on the Maximum Likelihood Method. In *6th International Congress on Acoustic, Tokyo*, August 1968.

[IS70]F. Itakura and S. Saito. A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies. *Electron. Com. Japan.*, 53-A:36–43, 1970.

[Ita75]F. Itakura. Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Trans. ASSP-23*, pages 67–72, 1975.

[IU87]F. Itakura and T. Umezaki. Distance Measure for Speech Recognition Based on the Smoothed Group Delay Spectrum. In *ICASSP-87*, pages 1257–1280, 1987.

[Jav83]Javel et al. Suppression of Auditory-Nerve Responses. Suppression Threshold and Growth, Iso-Suppression Contours. *J. Acoust. Soc. Am.*, 74(3):801–813, 1983.

[Jel76]F. Jelinek. Continuous Speech Recognition Using Statistical Methods. *IEEE Trans. ASSP-64*, pages 532–556, 1976.

[Joh80]D.H. Johnson. The Relationship between Spike Rate and Synchrony in Responses of Auditory-Nerve Fibers to Single Tones. *J. Acoust. Soc. Am.*, 68(4):1115–1122, 1980.

[Jon57]D. Jones. *An Outline of English Phonetics*. Cambride, Mass, Heffer, 1957.

[JRW86]B. H. Juang, L.R. Rabiner, and J.G. Wilpon. On the Use of Bandpass Liftering in Speech Recognition. In *ICASSP-86*, pages 765–768, 1986.

[Jun87]J.C. Junqua. Evaluation of ASR front-ends in speaker-dependent and speaker-independent recognition. *J. Acoust. Soc. Am.*, (81 S1):S93, 1987.

[JW88]J.C. Junqua and H. Wakita. Speaker-independent ASR Using Perceptual Cues in a Knowledge-Based Approach. In *EUSIPCO-88*, 1988.

[JW89]J.C. Junqua and H. Wakita. A Comparative Study of Cepstral Lifters and Distance Measures for All-pole Models of Speech in Noise. In *ICASSP-89*, 1989.

[Kam75]I. Kameny. Comparison of Formant Spaces of Retroflexed and Nonretroflexed Vowels. *IEEE Trans. ASSP*, 23:38–49, 1975.

[Kar87]M. Karjalainen. Auditory Models for Speech Processing. In *11th ICphS Tallinn*, 1987.

[Kat88]S. Katagiri. Relaxation-Based Speech Labeling. *J. Acoust. Soc. Am.*, (84 S1), 1988.

[Kav84]Kavaler et al. A Dynamic Time Warp IC for One Thousand Word Recognition System. In *ICASSP-84*, 1984.

[Kay80]S.M. Kay. Noise Compensation for Autoregressive Spectral Estimates. *IEEE ASSP-28*, (3):292–302, 1980.

[Kia68]N.Y.S. Kiang. A Survey of Recent Developments in the Study of Auditory Physiology. *Ann. Otol. Rhinol. Laryngol*, 77:656–675, 1968.

[KK84]J. Koljonen and M. Karjalainen. Use of Computational Psychoacoustical Models in Speech Processing: Coding and Objective Performance Evaluation. In *ICASSP-84*, 1984.

[Kla82]D.H. Klatt. Prediction of Perceived Phonetic Distance from Critical-Band Spectra: a First Step. In *ICASSP-82*, pages 1278–1281, 1982.

[Kla85]D.H. Klatt. The Problem of Variability in Speech Recognition and in Models of Speech Perception. In J.S. Perkell and D.H. Klatt, editors, *Variability and Invariance in Speech Processes*. Hillsdale, NJ, Lawrence Erlbaum Assoc, 1985.

[Kle86]J. De Kleer. An Assumption-Based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.

[Koh88]T. Kohonen. The Neural Phonetic typewriter. *Computer IEEE*, 21(3):11–22, Mars 1988.

[Kos87]B. Kosko. Constructing an Associative Memory. *Byte*, pages 137–144, September 1987.

[Kuh84]G.M. Kuhn. Description of a Color Spectrogram. *J. Acoust. Soc. Am.*, 76(3):682–685, 1984.

[KWTC65]N.Y.S. Kiang, T. Watanabe, E.C. Thomas, and L.F. Clark. *Discharge Patterns of Single Fibres in the Cat's Auditory Nerve*. MIT Press, Cambridge, MA, 1965.

[Lad82]P. Ladefoged. *A Course in phonetics*. Second Edition. Harcourt Brace Jovanovich New York, 1982.

[Lad85]P. Ladefoged. The Phonetic Basis for Computer Speech Processing. In F. Fallside and W. A. Woods , editor, *Computer Speech Processing*, pages 3–27. Prentice Hall, 1985.

[Leb88]J. Leboeuf. Un système Connexionniste Appliqué au Traitement Automatique de la Parole. Université de Paris-Sud, Octobre 1988. Thèse d'Université.

[Leh69]I. Lehiste, editor. *Readings in Acoustic Phonetics*. M.I.T. Press, 1969.

[Lim78]J.S. Lim. Estimation of LPC Coefficients from Speech Waveforms Degraded by Additive Random Noise. In *ICASSP-78*, pages 599–601, 1978.

[Lip87]R.P. Lippmann. An Introduction to Computing with Neural Nets. *IEEE Trans. ASSP Magazine*, pages 4–22, April 1987.

[Lip88]R.P. Lippmann. Neural Nets for Computing. In *ICASSP-88*, pages 1–6, 1988.

[LKS86]L.F. Lamel, R.H. Kassel, and S. Seneff. Speech Database Development: Design and Analysis of The Acoustic-Phonetic Corpus. In *Proceedings of the DARPA Speech Recognition Workshop, Palo Alto*, pages 100–109, February 1986.

[LMH88]H. Lâasri, B. Maître, and J.P. Haton. Hybrid control to achieve flexibility and efficiency in blackboard-based systems. In *Workshop on Blackboard Systems: AAAI-88*, 1988.

[Lom11]E. Lombard. Le Signe de l'Elévation de la Voix. *Ann. Maladiers Oreille, Larynx, Nez, Pharynx*, 37:101–119, 1911.

[Lon88]F. Lonchamp. Etudes sur la production et la perception de la parole. Université de Nancy II, Institut de Phonétique, Avril 1988. Thèse d'état.

[LRRW79]S. Levinson, L. Rabiner, A. Rosenberg, and J. Wilpon. Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition. *IEEE Trans. ASSP-27*, pages 134–141, 1979.

[LRS83]S. Levinson, L. Rabiner, and M. Sondhi. An Introduction to the Application of the Theory of Probalistic Functions of a Markov Process to Automatic Speech Recognition. *Bell Sys. Tech. J. 62*, pages 1035–1074, 1983.

[LT71]H.L. Lane and B. Tranel. The Lombard Sign and the Role of Hearing in Speech. *Speech and Hearing*, 14, 1971.

[Lyo82]R. F. Lyon. A Computational Model of Filtering, Detection, and Compression in the Cochlea. In *ICASSP-82*, pages 1282–1285, 1982.

[Lyo83]R. F. Lyon. A Computational Model of Binaural Localization and Separation. In *ICASSP-83*, pages 1148–1151, 2983.

[LZ82]L.F. Lamel and V.W. Zue. Performance improvement in a dynamic-programming-based isolated word recognition system for the alpha-digit task. In *ICASSP-82*, pages 558–561, 1982.

[LZ84]H.C. Leung and V.W. Zue. A Procedure for Automatic Alignment of Phonetic Transciptions with Continuous Speech. In *ICASSP-84*, pages 2.7.1–2.7.4, 1984.

[LZ88]H.C. Leung and V.W. Zue. Some Phonetic Recognition Experiments Using Artificial Neural Nets. In *ICASSP-88*, pages 422–425, 1988.

[Mah36]P.C. Mahalanobis. On the Generalized Distance in Statistics. In *Nat. Inst. Sci. India.*, volume 12, pages 49–55, 1936.

[Mak73]J. Makhoul. Spectral analysis of speech by linear prediction. *IEEE Trans. ASSP-21*, (3):140–148, 1973.

[Mak74]J.I. Makhoul. Selective Linear Prediction and Analysis-by-Synthesis in Speech Analysis. Technical Report 2578, Bolt Beranek and Newman Inc., Cambridge, Mass., April 1974.

[Mak75a]J. Makhoul. Linear Prediction: a Tutorial Review. *IEEE Trans. ASSP*, 63:561,580, 1975.

[Mak75b]J. Makhoul. Spectral Linear Prediction: Properties and Applications. *IEEE Trans. ASSP*, 23:283–296, June 1975.

[Mar75]T.B. Martin. Applications of Limited Vocabulary Recognition Systems. In D.R. Reddy, editor, *Speech Recognition*, pages 55–71. Academic Press, Inc, 1975.

[Mar82]J.J. Mariani. Un système de Compréhension de la Parole Continue. Université de Pierre et Marie Curie, Juillet 1982. Thèse d'Etat.

[Mar85]J.F. Mari. Reconnaissance de mots enchaînés à l'aide de modèles markoviens discrets. In *AFCET-RFIA-85*, pages 859–867, 1985.

[Mar87]J. Mariani. Les Technologies de Reconnaissance Automatique de la Parole. In INRIA, editor, *Research and Developpment in Language Processing*, pages 115–143. 1987.

[Mas87]D.W. Massoro. *Speech Perception by Ear and Eye*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, London, 1987.

[MC76]J. Makhoul and L. Cosell. LPCW: An LPC Vocoder with Linear Predictive Warping. In *ICASSP-76*, pages 466–469, 1976.

[MG76]J.D. Markel and A.H. Gray. *Linear Prediction of Speech*. Springer-Verlag, New York, 1976.

[MHM86]S. Morishima, H. Harashima, and H. Miyakawa. A Proposal of a Knowledge Based Isolated Word Recognition. In *ICASSP-86*, pages 713–716, 1986.

[MI86]H. Matsumoto and H. Imai. Comparative Study of Various Spectrum Matching Measures on Noise Robustness. In *ICASSP-86*, pages 769–772, 1986.

[MJ88a]D. Mansour and B.H. Juang. A Family of Distortion Measures Based upon Projection Operation for Robust Speech Recognition. In *ICASSP-88*, pages 36–39, 1988.

[MJ88b]D. Mansour and B.H. Juang. The Short-Time Modified Coherence Representation and its Application for Noisy Speech Recognition. In *ICASSP-88*, pages 525–528, 1988.

[MLG87]R. De Mori, L. Lam, and M. Gilloux. Learning and plan refinement in a knowledge-based system for automatic speech recognition. *IEEE ASSP-PAMI-9*, (2):289–305, 1987.

[MN55]G.A. Miller and P.E. Nicely. Analysis of Perceptual Confusions among some English Consonants. *J. Acoust. Soc. Am.*, 27:338–353, 1955.

[Moo77]R. Moore. Evaluating Speech Recognizers. *IEEE Trans. ASSP-25*, pages 178–183, 1977.

[MRR80]C. Myers, L.R. Rabiner, and A.E. Rosenberg. Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition. *IEEE Trans. ASSP*, 28(6):623–634, 1980.

[MS85]J. Makhoul and R. Schwartz. Ignorance Modeling: comments from Performing Fine Phonetic Distinctions, R. Cole, R. M. Stern, and M. J. Lasry. In J.S. Perkell and D.H. Klatt, editors, *Variability and Invariance in Speech Processes*. Hillsdale, NJ, Lawrence Erlbaum Assoc, 1985.

[MW86]S. Makino and H. Wakita. Automatic Labeling System for Large Speech Database. In *12th ICA*, pages A4–8, 1986.

[Nii86a]H.P. Nii. Blackboard Systems: Blackboard Applications Systems from a Knowledge Engineering Perspective. *Artificial Intelligence*, 7(3):82–106, 1986.

[Nii86b]H.P. Nii. Blackboard Systems: The Blackboard Model of Problem Solving and the Evolution of Blackboard Architectures. *Artificial Intelligence*, 7(2):38–53, 1986.

[NMW83]G. Neben, R.J. McAulay, and C.J. Weinstein. Experiments in Isolated Word Recognition Using Noisy Speech. In *ICASSP-83*, pages 1156–1159, 1983.

[Nod88]H. Noda. Frequency-Warped Spectral Distance Measures for Speaker Verification in Noise. In *ICASSP-88*, pages 576–579, 1988.

[NR87]C.V. Negoita and D. Ralescu. *Simulation, Knowledge-Based Computing, and Fuzzy Statistics*. Van Nostrand Reinhold Company Inc, 1987.

[NSRK85]N. Nocerino, F.K. Soong, L.R. Rabiner, and D.H. Klatt. Comparative Study of Several Distortion Measures for Speech Recognition. In *ICASSP-85*, pages 25–28, 1985.

[Obr88]R.A. Obrecht. A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals. *IEEE Trans. ASSP*, 36(1):29–40, 1988.

[O'S87]D. O'Shaughnessy, editor. *Speech Communication, Human and Machine.* Addison-Wesley, 1987.

[Pal82]K.K. Paliwal. On the Performance of the Quefrency-Weighted Cepstral Coefficients in Vowel Recognition. *Speech Communication*, (1):151–154, 1982.

[Pal88]K.K. Paliwal. A Study of Line Spectrum Pair Frequencies for Speech Recognition. In *ICASSP-88*, pages 485–488, 1988.

[PB52]G. Peterson and H. Barney. Control Methods Used in a Study of Vowels. *J. Acoust. Soc. Am.*, 24:175–184, 1952.

[PB81]T.L. Petersen and S.F. Boll. Acoustic Noise Suppression in the Context of a Perceptual Model. In *ICASSP-81*, pages 1086–1089, 1981.

[PBNY85]D.B. Pisoni, R.H. Bernacki, H.C. Nusbaum, and M. Yuchtman. Some Acoustic-Phonetic Correlates of Speech Produced in Noise. In *ICASSP-85*, pages 1581–1584, 1985.

[PC85]G. Perennou and M. De Calmes. Segmentation en Evénements Phonétiques et en Unités Syllabiques. In *XIV JEP PARIS*, pages 142–146, 1985.

[PC87]G. Perennou and M. De Calmes. Lexical Data and Knowledge Base of Spoken and Written French. In *European Conference on Speech Technology*, September 1987.

[Pfe74]L.L. Pfeifer. Inverse Filter for Speaker Identification. *Speech Communication Res. Lab., Santa Barbara, CA, Final Rep.*, RADC-TR-74-214, 1974.

[Phi87]M.S. Phillips. Speaker-Independent Classification of Vowels and Diphthongs in Continuous Speech. 1987.

[Pic57]J.M. Pickett. Perception of Vowels Heard in Noises of Various Spectra. *J. Acoust. Soc. Am.*, 29:613–620, 1957.

[PP89]A. Paolini and P. Pocci. Performance Assessment of Speaker-Independent Recognition Devices Using Italian Database. In *MELECON'89*, pages 245–248, 1989.

[RCM77]A.L. Rupert, D.M. Caspary, and G. Moushegian. Response Characteristics of Cochlear Nucleus Neurons to Vowel Sounds. *Ann. Otol.*, 86:37–48, 1977.

[RD56]D.W. Robinson and R.S. Dadson. A Redetermination of the Equal-Loudness relations for Pure Tones. *British Journal of Applied Physics*, 7:166–181, 1956.

[RD85]P.K. Rajasekaran and G.R. Doddington. Speech Recognition in the F-16 Cockpit Using Principal Spectral Components. In *ICASSP-85*, pages 882–885, 1985.

[RDP86]P.K. Rajasekaran, G.R. Doddington, and J.W. Picone. Recognition of Speech under Stress and in Noise. In *ICASSP-86*, pages 733–736, 1986.

[Red76]D.R. Reddy. Speech Recognition by Machine: A Review. *IEEE Trans. ASSP*, pages 501–531, April 1976.

[RHAB71]J.E. Rose, J.E. Hind, D.J. Anderson, and J.F. Brugge. Some Effects of Stimulus Intensity on Response of Auditory Nerve Fibers in the Squirrel Monkey. *Neurophysiol*, 34:685–699, 1971.

[RHW86]D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning Internal Representations by Error Propagation. In D. E. Rumelhart and J. L. McCelland, editor, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT PRESS, 1986.

[RJ86]L.R. Rabiner and B.H. Juang. An Introduction to Hidden Markov Models. *IEEE Trans. ASSP Magazine*, 3(1):4–16, 1986.

[RL81]L.R. Rabiner and S.E. Levinson. Isolated and Connected Word Recognition Theory and Selected Applications. *IEEE Trans. ASSP-29*, pages 621–659, 1981.

[RLS83]L. Rabiner, S. Levinson, and M. Sondhi. On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated-Word Recognition. *Bell Sys. Tech. J. 62*, pages 1075–1105, 1983.

[RRWK83]A.E. Rosenberg, L.R. Rabiner, J.G. Wilpon, and D. Kahn. Demisyllable-Based Isolated Word Recognition System. *IEEE Trans. ASSP-31*, (3):713–726, 1983.

[RS75]L.R. Rabiner and M.R. Sambur. An Algorithm for Determining the Endpoints of Isolated Utterances. *Bell Syst. Tech. J.*, 54(2):297–315, 1975.

[RW79]L. Rabiner and J. Wilpon. Considerations in Applying Clustering Techniques to Speaker-Independent Word Recognition. *J. Acoust. Soc. Am. 66*, pages 663–673, 1979.

[RW81]L.R. Rabiner and J.G. Wilpon. Isolated Word Recognition Using a Two-Pass Pattern Recognition Approach. In *ICASSP-81*, pages 724–727, 1981.

[SC71]H. Sakoe and S. Chiba. A Dynamic Programming Approach to Continuous Speech Recognition . In *ICASSP-71*, pages Paper 20C–13, 1971.

[Sch81a]M.R. Schroeder. Direct(Nonrecursive) Relations between Cepstrum and Predictor Coefficients. *IEEE ASSP-29*, pages 297–301, 1981.

[Sch81b]J.L. Schwartz. Apport de la Psychoacoustique à la Modélisation du Système Auditif chez l'Homme. I.N.P de Grenoble, 1981. Thèse de 3ième cycle.

[Sch86]R.M. Schwartz. Probabilistic Methods for Modeling Acoustic Variability in Speech Recognition. In W.A. Lea , editor, *Towards Robustness in Speech Recognition.* To appear in Apple Valley, Mn, Speech Science Publications, 1986.

[SEM86]P.E Stern, M. Eskenazi, and D. Memmi. An Expert System for Speech Spectrogram Reading. In *ICASSP-86*, pages 23.1.1–23.1.4, 1986.

[Sen84]S. Seneff. Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model. In *ICASSP-84*, pages 36.2–36.5, 1984.

[Sen86]S. Seneff. A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research. In *ICASSP-86*, pages 1983–1986, 1986.

[Sen87]S. Seneff. A Model for the Transduction Stage of Auditory speech Processing. *J. Acoust. Soc. Am.*, 82 S1:S83, 1987.

[Sha86]S. Shamma. Encoding the Acoustic Spectrum in the Spatio-Temporal Responses of the Auditory Nerve. In B. C. J. Moore and R. D. Patterson, editor, *Auditory Frequency Selectivity*, pages 289–296. New York, Plenum, 1986.

[Sha88]S. Shamma. The Acoustic Features of Speech Sounds in a Model of Auditory Processing: Vowels and Voiceless Fricatives. *Journal of Phonetics*, 16:77–91, 1988.

[SJ76]M.R. Sambur and N.S. Jayant. LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise. *IEEE Trans. ASSP*, 24:488,494, December 1976.

[SJA88]B.J. Stanton, L.H. Jamieson, and G.D. Allen. Acoustic-phonetic analysis of loud and lombard speech in simulated cockpit conditions. In *ICASSP-88*, pages 331–334, 1988.

[SKT87]Y. Sagisaka, S. Katagiri, and K. Takeda. Phonetic Labeling and Acoustic Correlates for Building Japanese Speech Database. *J. Acoust. Soc. Am.*, (81 S1), 1987.

[SM75]R. Schwartz and J. Makhoul. Where the Phonemes Are: Dealing with Ambiguity in Acoustic-Phonetic Recognition. *IEEE ASSP-23*, pages 50–53, 1975.

[SR75]M.R. Sambur and L.R. Rabiner. A Speaker-Independent Digit-Recognition System. *Bell Syst. Tech. J.*, 54:81–102, Janvier 1975.

[SR86]F.K. Soong and A.E. Rosenberg. On the Use of Instantaneous and Transitional Spectral Information in Speaker recognition. In *ICASSP-86*, pages 877–880, 1986.

[SS81]M. Sugiyama and K. Shikano. LPC Peak Weighted Spectral Matching Measures. *Electron. Commun. Japan*, 64-A(5):50–58, 1981.

[SS87]F.K. Soong and M.M. Sondhi. A Frequency-Weighted Itakura Spectral Distortion Measure and its Application to Speech Recognition in Noise. In *ICASSP-87*, pages 625–628, 1987.

[Ste57]S.S. Stevens. On the Psychophysical Law. *Psychological Review*, 64:153–181, 1957.

[Ste86]K.N. Stevens. Models of Phonetic Recognition II: An Approach to Feature-Based Recognition. In *Montreal Symposium of Speech Recognition*, pages 21–22, 1986.

[Ste87]K.N. Stevens. Relational Properties as Perceptual Correlates of Phonetic Features. In *11th ICphS Tallinn*, pages 352–356, 1987.

[SZ76]R.M. Schwartz and V.W. Zue. Acoustic-Phonetic Recognition in BBN SPEECHLIS. In *ICASSP-76*, pages 21–24, 1976.

[Tie80]J. Tierney. A Study of LPC Analysis of Speech in Additive Noise. *IEEE ASSP-28*, (4), 1980.

[Toh85]Y. Tohkura. Speaker-Independent ASR of Isolated Digits Using a Weighted Cepstral Distance. *J. Acoust. Soc. Am.*, (77 S1):S11, 1985.

[Toh87]Y. Tohkura. A Weighted Cepstral Distance Measure for Speech Recognition. *IEEE ASSP-35*, pages 1414–1422, 1987.

[TW88]S. Tamura and A. Waibel. Noise Reduction Using Connectionist Models. In *ICASSP-88*, pages 553–556, 1988.

[Vai85]J. Vaissiere. Speech Recognition: A Tutorial. In F. Fallside and W.A. Woods, editors, *Computer Speech Processing*, pages 191–226. Prentice Hall International, 1985.

[Vin68]T.K. Vintsyuk. Speech Discrimination by Dynamic Programming. *Kibernetika, Cybernatics*, 4(1), 1968.

[Wag81]M. Wagner. Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms. In *ICASSP-81*, pages 1156–1159, 1981.

[Wai87]Waibel et al. Phoneme Recognition Using Time-Delay Neural Networks. Technical Report TR-1-006, ATR Interpreting Telephony Research Laboratory Technical Report, October 1987.

[Wak73]H. Wakita. Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms. *IEEE Trans. ASSP*, AU-21(5):417–427, 1973.

[Wak81]H. Wakita. Linear Prediction Voice Synthesizers. *Speech Tech., Fall*, pages 17–22, 1981.

[WB73]M.D. Wang and R.C. Bilger. Consonant Confusions in Noise: a Study of Perceptual Features. *J. Acoust. Soc. Am.*, 54:1248–1266, 1973.

[Whi76]G.M. White. Speech Recognition: a Tutorial Overview. *Computer*, pages 40–53, May 1976.

[WR87]J.G. Wilpon and L.R. Rabiner. Application of Hidden Markov Models to Automatic Speech Endpoint Detection. *Computer Speech and Language*, 2:321–341, 1987.

[YR79]B. Yegnanarayana and R. Reddy. A Distance Measure Based on the Derivative of Linear Prediction Phase Spectrum. In *ICASSP-79*, pages 744–747, 1979.

[YS79]E.D. Young and M.B. Sachs. Representation of Steady-State Vowels in the Temporal Aspects of the Discharges Patterns of Populations of Auditory-Nerve Fibers. *J. Acoust. Soc. Am.*, 66:1381–1403, 1979.

[YSR84]B. Yegnanarayana, D.K. Saikia, and R. Reddy. Significance of Group Delay Functions in Signal Reconstruction from Spectral Magnitude or Phase. *IEEE Trans. ASSP*, 32(3):610–623, 1984.

[YT86]H. Ye, , and D. Tuffelli. Evaluation de Distances en Utilisant des Sons Synthétiques et la Perception Humaine. In *XV JEP AIX en PROVENCE*, pages 147–152, 1986.

[ZC79]V.W. Zue and R.A. Cole. Experiments on Spectrogram Reading. In *ICASSP-79*, pages 116–119, 1979.

[ZF81]E. Zwicker and R. Feldtkeller. *Psychoacoustique: l'Oreille Récepteur d'Informations.* Masson, 1981.

[ZL86]V.W. Zue and L.F. Lamel. An Expert Spectrogram Reader: A Knowledge-Based Approach To Speech Recognition. In *ICASSP-86*, pages 1197–1200, 1986.

[ZR81]S.A. Zahorian and M. Rothenberg. Principal-Component Analysis for Low-Redundancy Encoding of Speech Spectra. *J. Acoust. Soc. Am.*, 69:832–845, 1981.

[ZS65]E. Zwicker and B. Scharf. A Model of Loudness Summation. *Psychological Review*, 72(1):3–26, 1965.

[ZT79]E. Zwicker and E. Terhardt. Automatic Speech Recognition Using Psychoacoustic Models. *J. Acoust. Soc. Am.*, 65(2), February 1979.

[ZT80]E. Zwicker and E. Terhardt. Analytical Expressions for Critical Band Rate and Critical Bandwidth as a Function of Frequency. *J. Acoust. Soc. Am.*, 68:1523–1525, 1980.

Index

Les numéros de page qui apparaissent en caractères gras pointent sur des noms d'auteurs qui ont été référencés dans les pages en question. Uniquement le premier auteur est indiqué.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

THE ROAD NOT TAKEN

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that, the passing here
had worn them really about the same,

And both that morning equally lay
in leaves no step has trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two road diverged in a wood, and I-
I took the one less traveled by,
And that has made all the difference

Robert Frost

NOM DE L'ETUDIANT : Monsieur JUNQUA Jean-Claude

NATURE DE LA THESE : DOCTORAT DE L'UNIVERSITE DE NANCY I

EN INFORMATIQUE

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le 19 MAI 1989 n° 779

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

M. BOULANGE

RESUME

L'expérience montre que des systèmes de reconnaissance automatique de la parole qui donnent de bons résultats dans des environnements de laboratoire voient leurs performances se dégrader lorsqu'ils opèrent dans des conditions réelles. Ceci est lié à l'existence de nombreux problèmes souvent sous-estimés comme 1) la variabilité intra- et interlocuteur 2) l'existence de bruit de fond, et 3) la difficulté des vocabulaires étudiés. Les recherches liées à ces problèmes ont été regroupées sous la rubrique "amélioration de la robustesse des systèmes de traitement de la parole". Le travail présenté s'inscrit dans le cadre de "*l'amélioration de la robustesse des systèmes de reconnaissance de mots isolés*". Après une étude visant à déterminer le modèle d'analyse acoustique le moins sensible à la variabilité intra- et interlocuteur, les problèmes 2 et 3 sont étudiés. Certaines des méthodes développées sont ensuite appliquées à l'étiquetage automatique de la parole.

Initialement, un modèle auditif appelé, analyse par prédiction linéaire perceptivement fondée (*PLP*), est étudié. Après une évaluation de ce modèle par comparaison à d'autres modèles utilisés récemment, il est montré que la technique *PLP* permet d'obtenir d'aussi bons ou de meilleurs résultats que les autres modèles étudiés. Une optimisation de ce modèle, grâce à l'utilisation de caractéristiques spectrales dynamiques du signal de parole et d'une nouvelle mesure de distance, est ensuite proposée.

Une étude comparative de plusieurs modèles d'analyse acoustique en environnement bruité montre que ce modèle est néammoins très sensible au bruit. Ceci a amené le développement d'un modèle physiologique à synchronisation temporelle (*SLP*). Il est montré que ce modèle donne de meilleurs résultats que les autres modèles étudiés (sous certaines conditions) lorsque le rapport signal-sur-bruit est faible. Une extension de cette étude à de la parole prononcée dans du bruit (effet Lombard), montre que les changements phonétiques occasionnés par l'effort vocal dégradent beaucoup les performances. Dans un tel contexte, la technique *PLP* donne de meilleurs résultats que la technique *LP*.

Afin de prendre en compte le problème des vocabulaires difficiles, un système hybride: *ORION*, permettant la discrimination entre des mots acoustiquement similaires, a été développé. Grâce à l'utilisation de connaissances phonétiques, ce système améliore considérablement les performances obtenues à l'aide de systèmes conventionnels.

Enfin, un système de segmentation automatique, réalisé dans le cadre du système *ORION*, a été étendu à l'étiquetage automatique de la parole. La particularité de ce système est d'utiliser plusieurs sources de connaissances qui communiquent à travers un système "blackboard" administré par un contrôle hiérarchique.

Mots clés

Reconnaissance de la parole, mots isolés, robustesse, modèle d'analyse acoustique, mesure de distance, discrimination, source de connaissances, physiologie, psychoacoustique, phonétique, segmentation automatique.

I S B N - 2 - 7261 - 0586 - 6

TU_876