ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DE LA
MAÎTRISE EN GÉNIE ÉLECTRIQUE
M.Ing.

PAR
Marc-Antoine RONDEAU BEAUCHAMP

EFFET DU CANAL SUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

MONTRÉAL, LE 8 JANVIER 2010

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE

M. Gheorghe Marcel Gabrea, directeur de mémoire
Département de génie électrique à l'École de technologie supérieure

M. Christian Gargour, président du jury
Département de génie électrique à l'École de technologie supérieure

M. Jean-Marc Lina, membre du jury
Département de génie électrique à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 15 DÉCEMBRE 2009

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens d'abord à remercier M. Gheorghe Marcel Gabrea, mon directeur de recherche, dont la disponibilité et l'accessibilité ont été très appréciées. Ses conseils et son assistance ont été des plus utiles durant la rédaction de ce mémoire.

Je souhaite également remercier ma famille et mes amis, qui m'ont soutenu, encouragé et calmé dans mes occasionnels moments de stress. L'assistance de Dimitri, dont la capacité à repérer les fautes de français aussi vite que je pouvais les corriger a été des plus utile, a été particulièrement appréciée.

Un remerciement particulier à Line et Jocelyn, sans qui la logistique de mes études aurait fréquemment été beaucoup plus complexe.

EFFET DU CANAL SUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Marc-Antoine RONDEAU BEAUCHAMP

RÉSUMÉ

Les variations dans les caractéristiques du canal augmentent le nombre d'erreurs dans la reconnaissance automatique de la parole. Une nouvelle méthode, utilisant la transformée en paquets d'ondelettes, est présentée dans ce travail. Cette transformée est utilisée pour approximer les bandes critiques du système auditif humain. On calcul ensuite des coefficients cepstraux perceptuels à spectre relatif pour caractériser la parole. Les résultats de tests utilisant ces coefficients ont montré une amélioration de la robustesse aux variations de canal comparativement aux coefficients obtenus à partir d'une batterie de filtres basée sur la transformée de Fourier.

Mots-clés : reconnaissance automatique de la parole, transformée en ondelettes, bruit convolutif, robustesse, bandes critiques

EFFET DU CANAL SUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Marc-Antoine RONDEAU BEAUCHAMP

ABSTRACT

Changes in the channel's characteristics increase the number of errors during automatic speech recognition. A new method, which uses the wavelet packet transform, is described in this text. This transform is used to approximate the critical bands of human auditive system. RASTA-PLP coefficients are produced from the energy of the approximated critical bands. Test results, using those coefficients, show an improvement of the robustness to channel's variation when compared to coefficients produced from a Fourier transform filters bank.

Keywords : automatic speech recognition, wavelet transform, convolutional noise, robustness, critical bands

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

La transmission d'une personne à l'autre d'idées et de connaissances se fait par l'intermédiaire de symboles : le langage. La parole est un des deux principaux mécanismes de transmission de ces symboles, l'autre étant l'écriture. Dans les deux cas, le symbole est représenté par quelque chose qui peut être perçu par un être humain. Le développement de la technologie a permis l'apparition de machines capables de percevoir. Entre autres, les microphones, qui « entendent », et les caméras, qui « voient ». En combinant cette capacité de perception à d'autres dispositifs, il est possible de *reconnaitre* la parole dans le son et l'écriture dans l'image. Dans les deux cas, on reconnait les symboles, mais il n'est pas nécessaire de comprendre l'idée qu'ils représentent.

L'utilisation des symboles dépend de l'application. Par exemple, il est possible d'utiliser la reconnaissance de la parole pour commander un ordinateur (ou tout autre système qui a une interface usager). Ainsi, l'usager peut simplement parler, ce qui laisse ses mains libres. On peut également utiliser la reconnaissance pour transcrire automatiquement des textes, par exemple des entrevues, et ainsi alléger la charge de travail de l'usager. Dans le premier cas, les symboles sont transformés en commandes. La représentation interne des commandes et, par conséquent, des symboles est arbitraire : il n'est pas utile qu'elle corresponde à un langage. Dans le second cas, les symboles seront transformés en mots, en écriture. Néanmoins, la représentation interne de l'écriture est tout aussi arbitraire que celle des commandes.

On voit donc que la reconnaissance de la parole est totalement indépendante de la compréhension de la parole. La signification du symbole n'entre pas en compte. On reconnait simplement un ensemble de caractéristiques, présentes dans le son, que l'on a associées à un symbole donné. Ce symbole est tout aussi arbitraire que sa représentation interne. Il sont habituellement définis à partir du langage, mais c'est simplement parce que la parole est une représentation du langage. Seules les caractéristiques des symboles à reconnaitre sont importantes.

Comme la reconnaissance dépend des caractéristiques, tout ce qui les modifie est un obstacle.

Il faut donc faire en sorte que le système de reconnaissance soit robuste face aux variations. Les variations sont nombreuses. Il y a des variations causées par la production de la parole. L'être humain n'étant pas une machine dont les mouvements sont contrôlés avec une précision absolue, la production d'un symbole ne sera donc pas toujours identique. Comme la parole est produite par plusieurs éléments du corps humain et comme nous ne sommes pas tous identiques, il y a également des différences d'une personne à l'autre.

Il y a également des variations causées par l'environnement. Comme l'environnement est rarement parfaitement silencieux, il y a du bruit. Le bruit va être perçu en même temps que la parole. Les caractéristiques du bruit vont donc être présentes en plus de celle de la parole. En plus du bruit, l'environnement peut modifier la parole avant qu'elle soit perçue. L'atmosphère ne transporte pas toutes les ondes de la même façon, les murs vont produire des échos et les microphones ne sont pas tous identiques et ne vont pas tous percevoir le son identiquement. Ce sont les variations de canal. Ces variations vont également changer les caractéristiques de la parole. Il existe des techniques qui sont robustes à toutes ces variations. Malheureusement, ces techniques ne sont pas parfaites, et il est encore possible de réduire les erreurs durant la reconnaissance de la parole.

Le développement d'une nouvelle technique plus robustes aux variations de canal est l'objet ce mémoire. Cette technique est une nouvelle méthode de caractérisation de la parole, qui devrait être moins affectée par les variations de canal que les caractérisations actuelles. Cette méthode utilise la transformée en paquets d'ondelettes pour simuler le système auditif humain. Elle s'ajoute donc à une longue lignée de techniques qui utilisent ce système comme modèle. Comme les changements de canaux affectent fondamentalement le comportement en fréquence décrit par la transformée de Fourier, on espère que la résolution temporelle variable de la transformée en ondelettes permettra de conserve de l'information supplémentaire. Cette information supplémentaire dans le temps pourrait ainsi remplacer l'information fréquentielle distordue par le changement de canal.

La parole sera présentée en plus de détails dans le chapitre 1. Quelques particularités de la parole seront présentées en premier. Sa production, qui détermine les caractéristiques à reconnaître, sera ensuite présentée. Nous verrons ensuite sa perception et quelques notes sur son traitement.

Le chapitre 2 présentera différents outils appropriés au traitement de la parole en général, et à sa reconnaissance en particulier. Les transformées de Fourier et en ondelettes seront présentées dans ce chapitre. Nous y verrons également les outils utilisés pour modéliser les symboles à reconnaitre. Ce chapitre présente des notions de base pour le chapitre suivant.

Le chapitre 3 présentera un système de reconnaissance de la parole et les variations qui affectent ses performances. On y verra comment utiliser les notions présentées au chapitre 2 pour construire un système de reconnaissance de la parole. Deux méthodes de caractérisation de la parole y seront également présentées.

Le chapitre 4 présentera la méthode de caractérisation proposée. Nous verrons les détails de cette méthodes et les différents choix que sa conception a demandés.

Finalement, le chapitre 5 présentera les tests de cette méthode. L'environnement de test sera décrit. Les mesures de performances utilisées seront définies. Les résultats des tests et les conclusions tirées de ces résultats seront présentés en dernier.

CHAPITRE 1

LA PAROLE ET SON TRAITEMENT

1.1 Introduction

La parole est un moyen de communication utilisé par les êtres humains. Ce fait, plutôt évident, nous indique que la parole doit être produite par le système vocal humain et perçue par le système auditif humain. Les limitations de ces deux systèmes imposeront des limitations à la parole. L'étude de ces deux systèmes peut donc guider les techniques utilisées pour la reconnaissance automatique de la parole.

Comme l'objectif de la parole est la communication, elle est structurée : comme un système de communication digital qui produit des formes d'ondes pour transmettre des séquences de bits d'une façon plus ou moins complexe, la parole produit des signaux pour transmettre des symboles. Ces signaux sont aléatoires, contrairement aux signaux produits par le système digital, mais sont néanmoins structurés et représentent des symboles qui ne sont *pas* aléatoires. La reconnaissance de la parole consiste essentiellement à identifier les symboles transmis. L'étude de ces symboles et des relations entre eux est donc importante pour la reconnaissance.

Ce chapitre va donc présenter quelques particularités de la parole, en particulier sa structure. Par la suite sa production, sa perception et son traitement seront présentés.

1.2 Particularités de la parole

L'objectif principal de la parole est la transmission d'information. Ainsi, elle présente une structure particulière. Tout d'abord, les règles du langage utilisé doivent être respectées. Ceci limite le nombre d'éléments exprimés, et introduit un ordre qui doit être suivi. Malheureusement, même une grammaire correcte et une structure de phrase parfaite laissent place à une grande liberté d'action.

On distingue malgré tout des éléments de base : les phonèmes (Boite & Kunt, 1987, p.10). Un phonème est le plus petit élément qui doit changer pour changer la signification d'un mot. Les phonèmes sont classés selon leurs modes de production. Tout comme le mot est distinguable par ses phonèmes, un phonème est distinguable par sa production. Le nombre de phonèmes varie selon le langage. De plus, certains phonèmes peuvent être présents dans un langage et absents dans un autre. La prononciation exacte d'un phonème est influencée par les phonèmes qui l'encadre (Junqua & Haton, 1996, p.10). La composition phonétique d'un mot peut donc varier selon le contexte (Junqua, 2000, p.8-9).

Ainsi, la parole est constituée de phonèmes. Les phonèmes sont groupés en mots. Les mots sont séparés (ou non) par des silences et peuvent être groupés en phrase. L'ordre et le choix des mots sont régis par des règles, plus ou moins strictes et plus ou moins respectées. On peut ainsi ajouter un niveau qui utilise cette structure au dessus des caractéristiques sonores.

1.2.1 Non-stationnarité de la parole

Le signal vocal est un processus aléatoire non-stationnaire (Boite & Kunt, 1987, p.57). Il faut distinguer les statistiques long-termes et les statistiques court-termes. Les statistiques long-termes doivent être calculées sur des intervalles de temps élevés, de l'ordre de plusieurs dizaines de secondes. Au contraire, les statistiques court-termes doivent être calculées sur des intervalles courts, de l'ordre de 10 à 30 ms (Boite & Kunt, 1987, p.57). Le signal vocal est quasi-stationnaire sur ces intervalles. On suppose généralement la stationnarité du signal vocal lors du traitement. Ceci n'est évidemment possible que si le signal est séparé en courts segments.

1.3 Production de la parole

La production de la parole commence par les poumons, qui expulsent de l'air et créent une onde de pression. Cette onde passe par le conduit vocal. Ce passage la module pour produire la parole. Les cordes vocales sont des membranes qui peuvent laisser l'air passer librement

ou vibrer. Si l'air passe librement, le son sera un bruit blanc. Si les cordes vocales vibrent, l'ouverture et la fermeture périodique va produire l'onde glottale. Dans les deux cas, le son va ensuite passer au travers du conduit vocal, qui va le filtrer. La distinction entre les sons produits avec vibrations des cordes vocales (ou voisé) et sans vibrations (ou non-voisé) (Boite & Kunt, 1987, p.2) est une des distinctions les plus importantes dans la parole.

Les sons voisés sont produits avec un oscillation des cordes vocales. L'air expulsé des poumons est modulé par cette oscillation. Il en résulte un signal quasi-périodique (Boite & Kunt, 1987, p.4) : l'onde glottale. Ce son est caractérisé par sa fréquence fondamentale, qui correspond aux oscillations des cordes vocales. Dans le spectre, l'enveloppe des harmoniques présente des maximums appelés formants. L'amplitude de ces formants relative à l'amplitude de la fondamentale forme le timbre du son. Les trois premiers formants sont les plus utiles pour caractériser le spectre de la parole (Boite & Kunt, 1987, p.4). La fréquence des formants découle de la position du système masticatoire et de la présence ou absence de connexion avec la cavité nasale.

Les sons non-voisés ne présentent pas de structure périodique. Le passage libre de l'air au travers de la glotte produit du bruit blanc qui est ensuite filtré par le passage dans le conduit vocal (Boite & Kunt, 1987, p.4).

Le reste de la modulation est le résultat des mouvements et de la position de la langue, des lèvres et du voile du palais. La position de la langue change les fréquences de résonance du conduit vocal, qui sont appelés les formants. Ils sont particulièrement utiles pour la classification des voyelles. La langue peut également limiter le passage de l'air ou obstruer complètement le conduit vocal. Les lèvres ont un effet similaires : elles peuvent obstruer le conduit vocal et elles peuvent être arrondies. Finalement, le voile du palais peut obstruer ou ouvrir la cavité nasale, ce qui ajoute une chambre de résonance au conduit vocal.

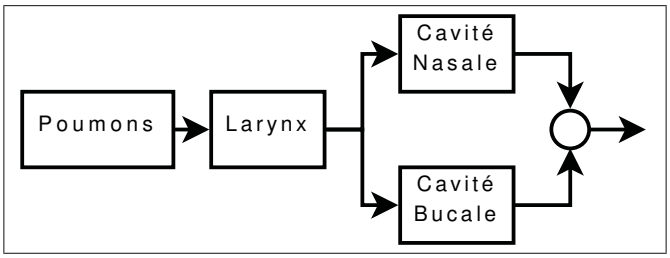


Figure 1.1 Schéma du conduit vocal.

1.3.1 Phonème

Les caractéristiques d'un phonème découlent directement de sa production. Deux phonèmes produits presque de la même façon, mais avec l'un voisé et l'autre non seront différents. Il faut noter que la production d'un phonème est influencée par les phonèmes adjacents. Le même principe s'applique pour toutes les caractéristiques des phonèmes : un seul changement de caractéristique change le phonème. On peut ensuite grouper les phonèmes selon ces caractéristiques. La table 1.1 liste les phonèmes du français, selon leur production.

Le lieu d'articulation indique où se trouve la principale obstruction du conduit vocal. Cette obstruction est provoquée par la langue ou les lèvres. Le mode d'articulation indique le comportement général des arcticulateurs durant la production d'un phonème. Une voyelle est un phonème voisé et la configuration du conduit vocal change peu durant sa production. Au contraire, une consonne peut être voisée ou non-voisée, et la configuration du conduit vocal change durant sa production.

1. Les voyelles, des sons voisé. Durant la production d'une voyelle, la configuration du conduit vocal change peu (Junqua & Haton, 1996, p.11).
2. Les consonnes, qui sont les sons qui ne sont pas des voyelles.
3. Les fricatives sont produites par une fermeture partielle du conduit vocal. Elles peuvent être voisées on non-voisées. L'obstruction du conduit vocal produit un son apériodique, similaire à du bruit (Junqua & Haton, 1996, p.17).

4. Les occlusives sont produites par une fermeture complète du conduit vocal, suivi d'un relâchement soudain de la pression. Elles peuvent être voisées ou non-voisées (Junqua & Haton, 1996, p.17).
5. Les sons nasaux sont produits par la cavité nasale, possiblement sans utiliser la cavité buccale (Junqua & Haton, 1996, p.19 ;Boite & Kunt, 1987, p.10).
6. Les liquides qui des sons difficiles à décrire, similaires aux voyelles (Junqua & Haton, 1996, p.19).
7. Cette liste n'est pas exhaustive et certaines catégories n'y sont pas représentées.

La production des voyelles est plus directe que celle des consonnes. Le degré d'ouverture indique à quel point le conduit vocal est ouvert, le lieu d'articulation indique la position de la principale obstruction et les lèvres peuvent être arrondies ou non. De cette façon, le conduit vocal peux être vu comme une chambre de résonance durant la production d'une voyelle. D'ailleurs, les voyelles peuvent être classifiées en utilisant les deux premiers formants. On peut définir des régions dans le plan construit à partir de ces formants et classifier la voyelle produite selon la position de ses formants dans ce plan.

Ces différentes catégories ont des caractéristiques spectrales et temporelles différentes, découlant directement du mode de leur production. Par exemple, les occlusives présentent une augmentation soudaine de l'amplitude, tandis que les nasales présentent des fréquences de résonances et d'antirésonances.

1.4 Perception de la parole

La perception de la parole est la contrepartie de sa production. Ce sont les caractéristiques perceptibles par le système auditif humain qui sont importantes. On peut faire une analogie avec un système de communication : la production de la parole correspond à l'émetteur. Le système auditif correspond à l'antenne et aux filtres du récepteur. L'étude de la perception de la parole par l'humain peut donc permettre d'identifier les éléments importants dans le signal pour la reconnaissance automatique de la parole

Tableau 1.1 Liste des phonèmes du français
Adapté de Calliope (1989)

Consonnes			
	Lieu d'articulation		
Mode d'articulation	Labiales	Dentales	Vélo-Palatales
Occlusives non voisées	[p]	[t]	[k]
Occlusives voisées	[b]	[d]	[g]
Fricative non voisées	[f]	[s]	[ʃ]
Fricative voisées	[v]	[z]	[ʒ]
Nasales	[m]	[n]	[ɲ]
Glissantes	[w]	[ɥ]	[j]
Liquides		[l]	[ʁ]
Voyelles			
	Lieu d'articulation		
Orales	Antérieures		Postérieures
Degrés d'ouverture	Non arrondies	Arrondies	
Fermées	[i]	[y]	[u]
	[e]	[ø]	[o]
	[ɛ]	[œ]	[ɔ]
Ouvertes	[a]		
Nasales	Antérieures		Postérieures
Fermées	[ɛ̃]		[ɔ̃]
Ouvertes		[ã]	

L'oreille est l'organe qui perçoit le son. Elle peut être séparée en trois parties : l'oreille interne, moyenne et externe. L'oreille externe est constituée du pavillon et du conduit auditif. Cette section dirige le son vers l'oreille moyenne. L'oreille moyenne commence par le tympan, une membrane qui oscille en suivant les différences de pressions dans l'air. Cette oscillation est transmisse aux osselets, trois os, qui transmettent ensuite cette oscillation a l'oreille interne.

La cochlée est la partie de l'oreille interne qui va transformer les vibrations en impulsions nerveuses. Il s'agit d'un tube, en spirale, contenant deux chambres remplies de fluide, séparées par une membrane. Des cellules placées dans ces chambres perçoivent les vibrations. Elles répondent à certaines fréquences. Le système auditif humain agit donc comme une batterie de filtres : les bandes critiques. Ces filtres sont partiellement superposés, et ont une bande passante qui augmente selon la fréquence centrale. Les fréquences centrales de ces filtres ne

sont pas placées linéairement. Elles sont placée linéairement dans les basses fréquences, et logarithmement pour les hautes fréquences. L'échelle de Mel modélise ce phénomène. Une augmentation de « fréquence »qui est perçue linéairement sera linéaire dans cette échelle et logarithmique en hertz. L'échelle de Mel est définie par :

$$M(f) = 1125 \ln(1 + f/700) \tag{1.1}$$

Des filtres placés linéairement sur cette échelle auront donc des fréquences centrales et de coupures similaires à celles des bandes critiques. L'espacement entre deux filtres sera presque constant pour les basses fréquences et augmentera dans les hautes fréquences. La bande passante aura le même comportement. Par exemple, la différence entre 1kHz et 2kHz est approximativement 520 Mel. C'est également la différence entre 0Hz et 410Hz.

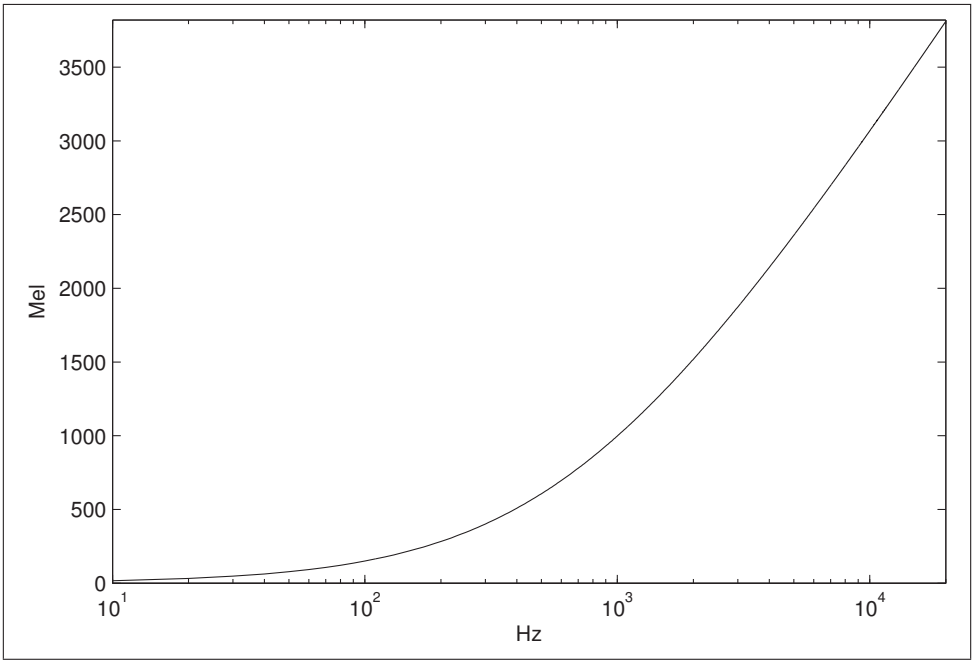


Figure 1.2 Mel selon la fréquence

On mesure l'intensité d'un son en dB, avec 0dB équivalent au seuil de perception, c'est à dire l'intensité minimale qu'un son doit avoir pour être perçu. La perception de l'intensité varie selon la fréquence. On peut modeler ce phénomène par des courbes d'égalisation de intensité

variant selon la fréquence. Des sinusoïdes pures placées sur ces courbes seront perçues comme ayant le même niveau même si elles sont d'intensités différentes.

1.5 Traitement de la parole

Les techniques de traitements de la parole sont affectées par sa production et sa perception. Même si les objectifs de ces traitements sont variés, ils utilisent des outils mathématiques similaires. Ainsi il est souvent possible d'adapter une technique à un autre problème qui est, au premier abord, complètement différent.

On peut distinguer :

1. Le rehaussement de la parole, qui consiste a rendre la parole plus claire, soit pour le bénéfice direct d'un auditeur, ou pour un traitement subséquent. La réduction du bruit additif est une exemple de rehaussement, tout comme la réduction des échos.

2. L'encodage, qui permet de réduire le nombre de bits requis pour représenter la parole. Ce type de traitement peut être très simple ou très complexe. À un extrême, on peut utiliser directement la sortie d'un convertisseur analogique-digital. À l'autre, on peut tenir compte de des limites de la perception sonore humaine pour éliminer les composants qui ne seront pas perçus.

3. La synthèse permet de produire de la parole à partir d'un texte. C'est le complément de la reconnaissance automatique : l'objectif est dans le sens opposé (texte vers parole, plutôt que parole vers texte) et les sources de difficultés sont souvent similaires. Par exemple, les variations dans la prononciation d'un phonème doivent être reproduites pour que la parole synthétisée semble naturelle. Au contraire, un système de reconnaissance doit absorber ces variations. Dans les deux cas, les variations sont un problème, mais la solution est complémentaire.

4. La reconnaissance de la parole, qui est décrite plus en détail au chapitre 3. Ce traitement présente des similitudes avec les autres. Tout comme pour l'encodage, il est important de conserver les éléments significatifs dans le signal sonore. Tout comme pour la synthèse, il est important d'identifier les variations dans la parole. Finalement, le rehaussement peut être appliqué avant la reconnaissance, pour réduire le bruit ou retirer des éléments nuisibles du signal.

1.6 Conclusion

Le traitement de la parole est délimité par la production et la perception de la parole. D'un coté, si un humain ne peut percevoir un élément du signal, ce n'est pas un élément significatif. Par exemple, les ultrasons ne sont pas important dans le traitement de la parole : ils ne sont pas perceptibles par l'oreille humaine, et peuvent être présents ou absents sans que la parole soit affectée. De l'autre coté, la production de la parole est ce qui la caractérise. Les différences entre la prononciation des différents mots sont causées, très évidemment, par la méthode de production de ces mots. Si la perception permet de définir ce dont on ne doit *pas* tenir compte, la production permet de définir ce dont on *doit* tenir compte. Par exemple, l'étude de la perception ne permet pas de savoir que les variations brusques d'amplitude sont significatives, alors que l'étude de la production nous indique que c'est une caractéristique clé des consonnes occlusives.

Si la production et la perception délimitent les caractéristiques pertinentes du signal dans lequel on reconnait la parole, c'est la structure du langage qui délimite ce que l'on tente de reconnaitre. On peut donc utiliser cette structure pour corriger les erreurs de reconnaissance, en éliminant ou pénalisant les combinaisons de symboles impossibles.

Comme les diverses techniques de traitement de la parole agissent sur le même type de signal, elles utilisent plusieurs outils communs. Les outils développés pour une technique peuvent souvent être appliqués à une autre.

CHAPITRE 2

OUTILS MATHÉMATIQUES

2.1 Introduction

La reconnaissance de la parole dépend de plusieurs disciplines : le traitement de signal, la linguistique, et les statistiques et probabilités. Ce n'est pas une caractéristique unique à la reconnaissance de la parole. Le traitement de signal est utilisé dans des domaines allant du contrôle de machineries aux systèmes de communications, la linguistique est utilisée dans tout ce qui demande l'étude des langages et les statistiques et probabilités sont omniprésentes : il suffit de lire un journal quotidien durant une campagne électorale pour le constater. Cette dépendance n'est donc qu'une instance particulière des liens communs et parfois surprenants entre les domaines du savoir.

En appliquant des transformées au signal, on peut en extraire les éléments qui décrivent la parole. La nature des transformées va changer les éléments extraits. Le choix d'une transformée appropriée est donc important : il faut que les éléments pertinents soient présents et, idéalement, que les éléments non-pertinents soient exclus.

Un modèle du langage adéquat permet de limiter les erreurs en guidant la reconnaissance. Ce modèle permet de déterminer si une phrase est possible ou impossible. Il peut également pénaliser les séquences improbables au profit des séquences plus probables. Ce modèle doit être adapté au domaine d'application du système de reconnaissance, qui va souvent être plus restreint que le langage lui-même. Il n'est pas utile, par exemple, qu'un système de composition de numéro de téléphone reconnaisse autant de mots et de phrases qu'un système de traitement de texte.

Finalement, les statistiques et les probabilités permettent de définir des modèles. La parole reste, malgré sa structure, un signal aléatoire. On va donc utiliser des statistiques découlant de

ce signal pour produire des modèles qui correspondent à un élément de la parole. Ces statistiques vont permettre d'absorber les variations naturelles dans la production de la parole. De plus, les langages eux-mêmes sont suffisamment complexes pour être difficiles à décrire. On peut utiliser des statistiques découlant d'exemples de textes pour établir des modèles probabilistes du langage. Ces modèles ne tiendront pas compte de la signification d'une séquence de mots, mais simplement de la fréquence de cette séquence dans les exemples utilisés durant la construction du modèle.

2.2 Transformée de Fourier discrète

On définit la transformée de Fourier discrète en N points $F_N\{x(n)\}$ du signal $x(n)$ par :

$$F_N\{x(n)\} = X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi k}{N} n} \tag{2.1}$$

et la transformée inverse par :

$$F_N^{-1}\{X(k)\} = x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{\frac{j2\pi n}{N} k} \tag{2.2}$$

On peut également définir la transformée de Fourier continue d'un signal discret :

$$F\{x(n)\} = X(\omega) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\omega n} \tag{2.3}$$

où $\omega = 2\pi f / f_s$ est la fréquence d'échantillonnage. Si le signal $x(n)$ est égal a zéro pour $n < 0$ et $n > N-1$ et en remplaçant ω par $2\pi f / f_s$, l'équation 2.3 devient :

$$X(f) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi \frac{f}{f_s} n} \tag{2.4}$$

Les équations 2.1 et 2.4 sont égales si

$$f = k \frac{f_s}{N} \tag{2.5}$$

Ainsi, on voit que la transformée de Fourier discrète F_N échantillonne la transformée de Fourier continue d'un signal discret. Plus N est élevé, plus l'échantillonnage est fin. Augmenter N permet donc de mieux distinguer les composants fréquentiels du signal. De plus, l'exponentielle $e^{-j2\pi}$ ajoute une période N à la transformée, et $X(k) = X(N-k)^*$, pour $0 <= k < N$. Ainsi, seuls $N/2$ points de la transformée sont significatifs.

2.3 Transformée de Fourier court-terme

La transformée de Fourier discrète ne varie pas selon le temps. Pour la reconnaissance de la parole, on doit pouvoir distinguer des changements dans le temps. On utilise donc la transformée de Fourier court-terme, définie par :

$$F_{N;ct}\{x(n)\} = X(m,k) = \sum_{n=0}^{N-1} x(n+m)w(n)e^{-\frac{j2\pi k}{N}n} \tag{2.6}$$

où $w(n)$ est une fonction *fenêtre*, égale a zéro pour $n < 0$ et $n > N-1$. La fenêtre permet de découper le signal original $x(n)$ en tranches de N points. Sans cette fonction, il y a des transitions abruptes au début et à la fin de chaque tranche. Cette fonction est multipliée avec le signal complet. On sait que la multiplication dans le temps est une convolution dans le domaine de Fourier. Donc, la transformée de Fourier de la fonction fenêtre est convoluée avec celle du signal. Comme on souhaite obtenir la transformée du signal, il faut une fenêtre avec une transformée de Fourier s'approchant d'une impulsion. Ceci n'est évidement pas possible. On utilise en pratique des fonctions avec un lobe central le plus étroit possible et des lobes latéraux avec une amplitude la plus basse possible.

On remarque que N limite la précision dans le temps : $X(m,k)$ dépend du signal $x(n)$ pour $n = m$ et $n = m+N-1$. Ainsi, il est difficile de distinguer un événement à $n = m+1$ d'un événement à $n = m+N-2$. La résolution temporelle de la transformée de Fourier court-terme est donc N/f_s. Or, la résolution fréquentielle est $f_s/(2N)$. On voit donc que pour augmenter une de ces résolutions, l'autre doit être diminuée.

2.3.1 Effet de la fenêtre

On sait que la multiplication dans le temps correspond à une convolution dans le domaine fréquentiel :

$$F\{x(n)y(n)\} = \frac{1}{2\pi}\int_{-\pi}^{+\pi} X(\lambda)Y(\omega-\lambda)d\lambda \tag{2.7}$$

Ainsi, la transformée de Fourier de la fenêtre va être convoluée avec celle du signal autour de m. Si la transformée de la fenêtre est une impulsion d'amplitude 2π, cette convolution n'affectera

pas la transformée du signal.

$$\frac{1}{2\pi}\int_{-\pi}^{+\pi} X(\tau)\cdot 2\pi\delta(\omega-\tau)d\tau = X(\omega)\int_{-\pi}^{+\pi}\delta(\omega-\tau)d\tau = X(\omega) \tag{2.8}$$

car la fonction $\delta(t)$ est définie :

$$\delta(t) = \begin{cases} 0 & ,\text{si } t \neq 0 \\ \int_A^B \delta(t)dt = 1 & ,\text{si } A < 0 \text{ et } B > 0 \end{cases} \tag{2.9}$$

Malheureusement, en appliquant la transformée de Fourier inverse à la fonction $2\pi\delta(t)$, on obtient :

$$w(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} 2\pi\delta(\omega)e^{j\omega n}d\omega = \int_{-\pi}^{\pi}\delta(\omega)e^{j0n}d\omega = 1 \tag{2.10}$$

Une telle fenêtre est évidemment inutile : le signal sera inchangé. Il faut donc utiliser une fonction dont la transformée de Fourier s'approche d'une impulsion mais qui reste localisée dans le temps. La fonction rectangle $r(n)$ de longueur L définit par :

$$r(n) = \begin{cases} 1 & ,0 \leq n < L \\ 0 & ,\text{sinon} \end{cases} \tag{2.11}$$

semble intéressante au premier abord. Cette fonction découpe simplement le signal. Elle peut donc être implémentée sans calculs. Cependant, la transformée de Fourier de cette fonction est :

$$R(\omega) = \begin{cases} e^{-j\frac{\omega}{2}(L-1)}\frac{sin\frac{\omega}{2}L}{sin\frac{w}{2}} & ,\omega \neq 0 \\ L & ,\omega = 0 \end{cases} \tag{2.12}$$

On peut donc observer que cette fonction introduit un changement de phase (le terme $e^{-j\frac{\omega}{2}(L-1)}$) et que le module va introduire des oscillations selon la fréquence.

Plutôt que d'utiliser une fenêtre rectangulaire, on peut utiliser une fenêtre de Hamming :

$$h(n) = \begin{cases} 0.54 - 0.46\cos\left(2\pi\frac{n}{L-1}\right) & ,0 \leq n < L \\ 0 & ,\text{sinon} \end{cases} \tag{2.13}$$

Cette fenêtre inclut une transition plus douce aux extrémités. Sa transformée de Fourier est :

$$W(\omega) = e^{-j\frac{\omega}{2}(L-1)}\left(\begin{array}{l} 0.54\frac{\sin\left(\frac{\omega}{2}L\right)}{\sin\left(\frac{\omega}{2}\right)} \\ +0.23\frac{\sin\left(\pi\frac{L}{L-1}\right)\cos\left(\frac{\omega}{2}L\right)-\cos\left(\pi\frac{L}{L-1}\right)\sin\left(\frac{\omega}{2}L\right)}{\sin\left(\frac{\pi}{L-1}\right)\cos\left(\frac{\omega}{2}\right)-\cos\left(\frac{\pi}{L-1}\right)\sin\left(\frac{\omega}{2}\right)} \\ +0.23\frac{\cos\left(\pi\frac{L}{L-1}\right)\sin\left(\frac{\omega}{2}L\right)+\sin\left(\pi\frac{L}{L-1}\right)\cos\left(\frac{\omega}{2}L\right)}{\cos\left(\frac{\pi}{L-1}\right)\sin\left(\frac{\omega}{2}\right)+\sin\left(\frac{\pi}{L-1}\right)\cos\left(\frac{\omega}{2}\right)} \end{array}\right) \tag{2.14}$$

On constate la présence d'un terme de phase. On peut aussi observer la présence de la transformée d'une fenêtre rectangulaire, multipliée par 0.54. Ceci est provoqué par la présence d'une fenêtre rectangulaire multipliée par 0.54 dans la fenêtre de Hamming.

La fenêtre rectangulaire présente un pic central, qui s'approche plus d'une impulsion que celui de la fenêtre de Hamming. Par contre, l'amplitude des oscillations de la fenêtre de Hamming est moins élevée. La figure 2.1 illustre les différences entre ces deux fenêtres. Dans les deux cas, la transformée de Fourier du signal va être distordue par la fenêtre. Comme la transformée du signal et la transformée de la fenêtre sont convoluées, les composants fréquentiels adjacents seront additionnés.

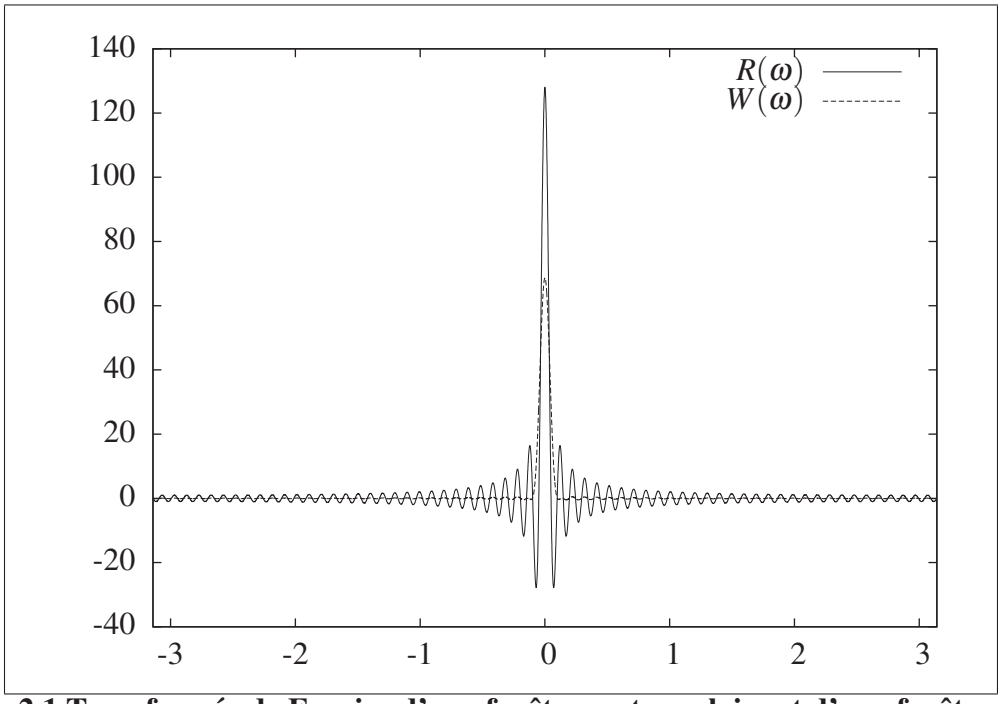


Figure 2.1 Transformée de Fourier d'une fenêtre rectangulaire et d'une fenêtre de Hamming ($\mathbf{L = 128}$).

2.4 Transformée en ondelettes discrète

La plupart des transformées d'un signal discret ont une définition de la forme :

$$T\{f(n)\}_\theta = \sum_{n=-\infty}^{\infty} f(n)\Theta_\theta(n) \tag{2.15}$$

Où Θ_θ est une fonction qui varie selon le paramètre θ. On compare donc la fonction à analyser à un ensemble de fonction pour obtenir des coefficients qui varient selon θ. Par exemple, pour la transformée en Z, $\theta = z$ et $\Theta_\theta(n) = z$. Pour une transformée de Fourier discrète court-terme, $\theta = k$ et $\Theta_\theta(n) = w(n)e^{-\frac{j2\pi k}{N}n}$. Dans ces deux cas, θ est simplement un paramètre d'une fonction, mais ce n'est pas obligatoirement le cas. Par exemple, ont pourrait définir $\theta = \{w(n), k\}$. Dans ce cas, l'ensemble de fonction dépendrait de plusieurs fonctions fenêtres $w(n)$ et de k.

La transformée en ondelettes discrète est une transformée temps-échelle. On obtient le *détail* et *l'approximation* du signal à différentes échelles. L'approximation correspond au signal, vu avec plus ou moins de détails. Le détail correspond à la différence entre une approximation et l'approximation de niveau supérieur. On utilise deux fonctions : $\psi(n)$ et $\phi(n)$, qui correspondent respectivement au détail et à l'approximation (Mallat, 1989). Le lien entre ces fonctions sera présenté ultérieurement.

La fonction $\psi(n)$ est l'ondelette. On l'utilise pour définir un ensemble de fonctions Θ_θ tel que :

$$\Theta_\theta(n) = \psi_{\{s,k\}}(n) = 2^{-s}\psi(2^{-s}n - k) \tag{2.16}$$

Ainsi, θ contient l'échelle s et la translation k. La fonction $\phi(n)$ est la fonction de mise à l'échelle. On l'utilise de la même façon que $\psi(n)$:

$$\phi_{\{s,k\}}(n) = 2^{-s}\phi(2^{-s}n - k) \tag{2.17}$$

Ces fonctions permettent d'obtenir l'approximation et le détail selon l'échelle et le temps. Comme il y a plusieurs fonctions possibles, la transformée en ondelettes discrète est en fait un ensemble de transformées. Les figures 2.2 et 2.3 illustrent un choix de fonctions possible, qui correspond à l'ondelette de Daubechies à 5 moments nuls.

La transformée en ondelettes discrète est similaire à la transformée de Fourier court-terme présentée précédemment. En particulier, on remarque que l'ondelette utilisée sert de fenêtre.

Comme le paramètre d'échelle change la longueur de l'ondelette, la résolution dans le temps varie selon l'échelle. Ainsi, plus l'échelle est réduite, plus la résolution temporelle est élevée. Contrairement à la transformée de Fourier court-terme, on peut donc dire que la transformée en ondelettes discrète est multi-résolutions.

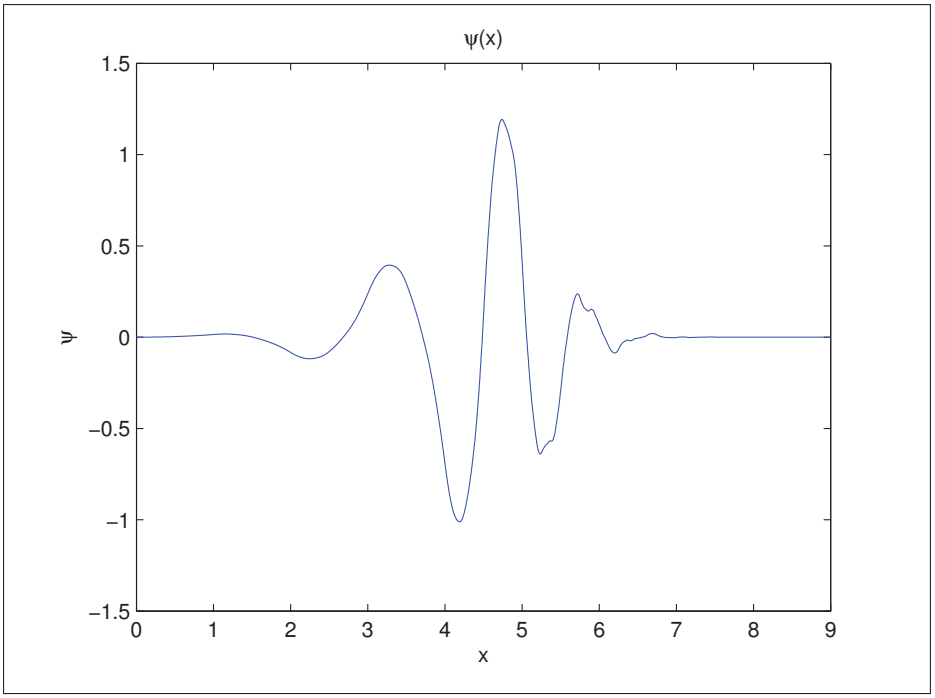


Figure 2.2 Exemple de fonction ψ.

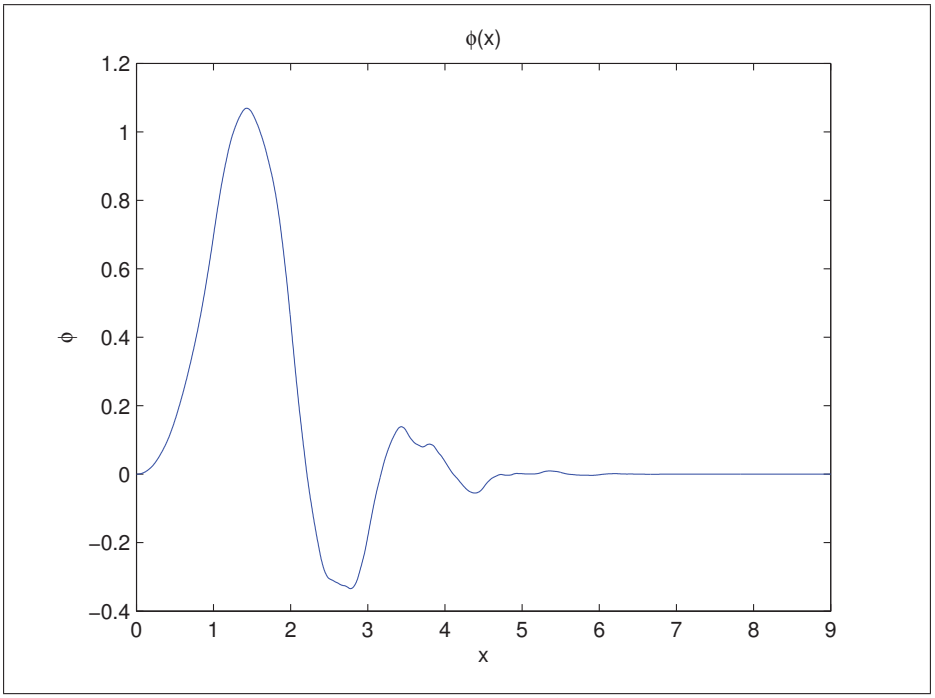


Figure 2.3 Exemple de fonction ϕ.

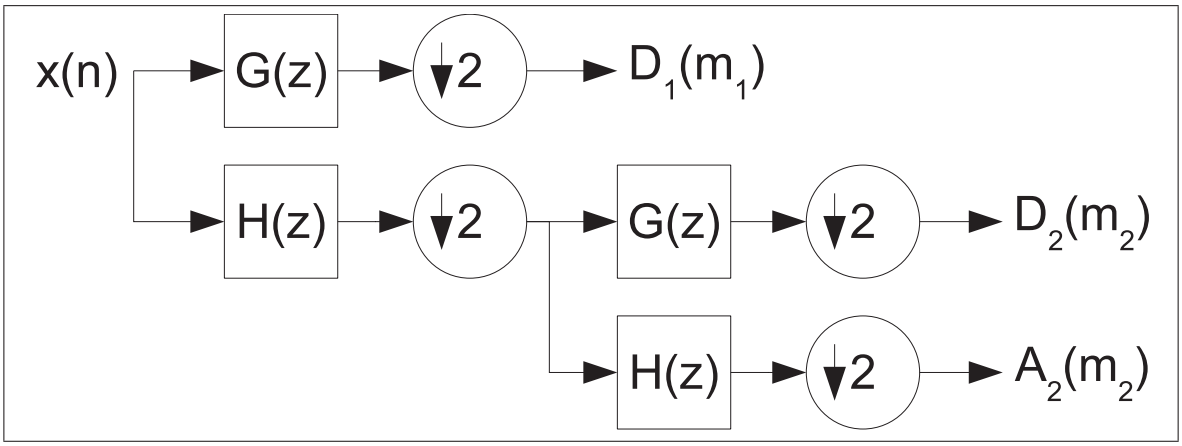


Figure 2.4 Pyramide de décomposition.

2.4.1 Algorithme de Mallat

Mallat (1989) a démontré que l'on peut définir un filtre à réponse impulsionelle finie $h(n)_{n\in\mathbb{N}}$ correspondant à une fonction de mise à l'échelle :

$$h(n) =< \phi_{2^{-1}}(u), \phi(u-n) >= \int_{-\infty}^{+\infty} 2^{-1}\phi(2^{-1}u)\phi(u-n)du \tag{2.18}$$

On définit également le filtre miroir $\bar{h}(n) = h(-n)$. Le filtre $\bar{h}(n)$ est un filtre passe-bas. On peut obtenir l'approximation $A_{2^{-s}}$ de $f(n)$ en filtrant $A_{2^{-s+1}}$ de $f(n)$ avec $\bar{h}(n)$ et en décimant la sortie par 2. La transformée de Fourier $\hat{\psi}(\omega)$ de l'ondelette $\psi(x)$ est :

$$\hat{\psi}(\omega) = e^{-j\omega/2}\bar{H}\left(\frac{\omega+\pi}{2}\right)\hat{\psi}\left(\frac{\omega}{2}\right) \tag{2.19}$$

On peut obtenir un filtre $g(n)$ similaire à $h(n)$:

$$g(n) =< \psi_{2^{-1}}(u), \phi(u-n) > \tag{2.20}$$

Le filtre $\bar{g}(n)$ est un filtre passe-haut. Ce filtre permet d'obtenir le détail $D_{2^{-s}}$ de $f(n)$ en filtrant $A_{2^{-s+1}}$ de $f(n)$ avec $\bar{g}(n)$ et en décimant la sortie par 2. On peut donc obtenir l'approximation et le détail à une échelle s en filtrant avec $\bar{h}(n)$ ou $\bar{g}(n)$ et en décimant par 2. La transformée en ondelettes discrète peut ainsi être obtenue à partir d'une pyramide de filtres et décimateurs. La figure 2.4 illustre cette pyramide. Les figures 2.5 et 2.7 illustrent les réponses impulsionelles des filtres correspondants aux fonctions ψ et ϕ des figures 2.2 et 2.3. Leur transformées de

Fourier sont illustrées par les figures 2.6 et 2.8. Dans ces deux figures, les lignes pointillées indiquent les pulsations de coupures.

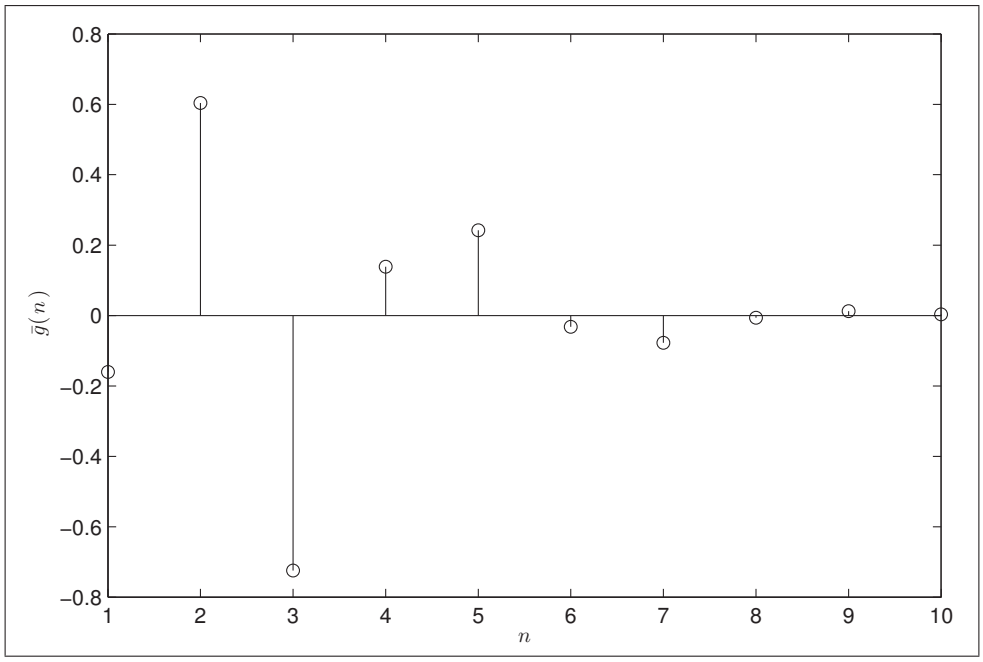


Figure 2.5 Exemple de filtre $\bar{g}(n)$.

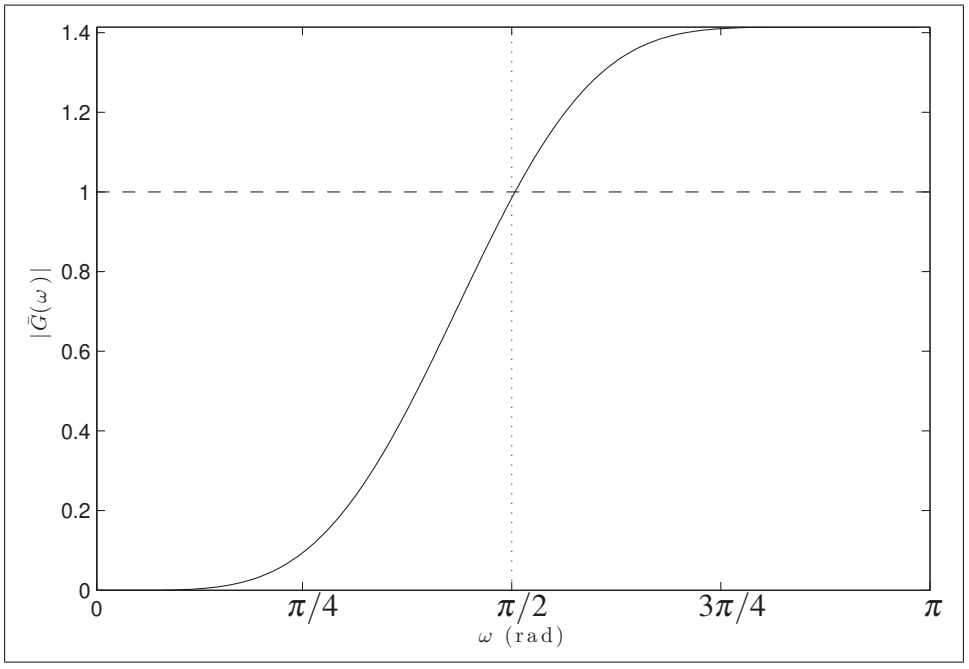


Figure 2.6 Transformée de Fourier continue du filtre $\bar{g}(n)$.

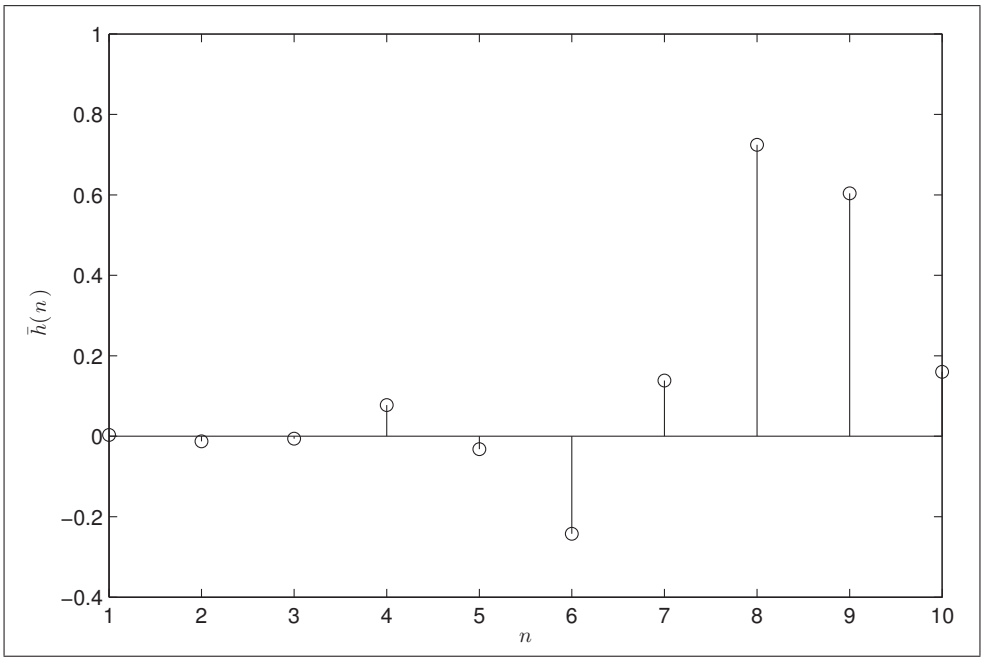


Figure 2.7 Exemple de filtre $\bar{h}(n)$.

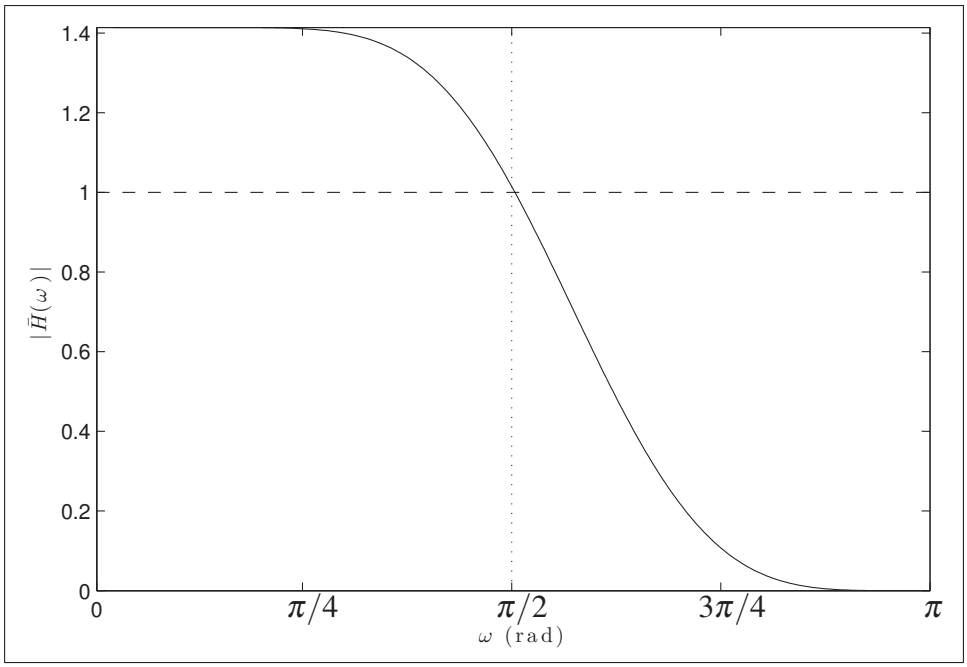


Figure 2.8 Transformée de Fourier continue du filtre $\bar{h}(n)$.

2.4.2 Décimation d'un signal

La décimation par N d'un signal consiste à diviser la fréquence d'échantillonnage par N en conservant un échantillon sur N. La transformée en Z d'un signal décimé par 2 $X_2(z)$ peut être

obtenue à partir de la transformée en Z du signal original :

$$X_2(z) = \frac{1}{2}\left[X\left(z^{1/2}\right) + X\left(-z^{1/2}\right)\right] \tag{2.21}$$

On peut obtenir la transformée de Fourier d'un signal discret à partir de sa transformée en Z, si le cercle unitaire fait partie de la région de convergence :

$$X(2\pi f/f_s) = X_z\left(e^{j2\pi f/fs}\right) \tag{2.22}$$

Où f_s est la fréquence d'échantillonnage après la décimation. Ceci permet de réécrire X_2 en fonction de f/f_s :

$$X_2(2\pi f/f_s) = \frac{1}{2}\left[X\left(\frac{2\pi f/f_s}{2}\right) + X\left(\frac{2\pi f/f_s}{2} + \frac{2\pi f_s}{2f_s}\right)\right] \tag{2.23}$$

On peut également obtenir X_2 en fonction de la fréquence d'échantillonnage originale $f_{sorg} = 2f_s$:

$$X_2\left(2\pi\frac{f}{f_{sorg}/2}\right) = \frac{1}{2}\left[X\left(2\pi f/f_{sorg}\right) + X\left(2\pi f/f_{sorg} + \frac{2\pi f_{sorg}/4}{f_{sorg}}\right)\right] \tag{2.24}$$

Pour éviter le recouvrement et respecter le théorème de Nyquist, $X\left(2\pi f/f_{sorg}\right)$ doit être 0 pour $f_{sorg}/4 \leq f < f_{sorg}$. Dans le cas de la transformée en ondelettes, on remarque que les détails sont obtenus à partir d'un filtre passe-haut suivi d'un décimateur. Dans ce cas, la moitié inférieure du spectre du signal original doit être filtrée. Il va y avoir recouvrement, mais comme les basses fréquences ont été filtrées, ce n'est pas un problème.

La première identité de Noble spécifie qu'un décimateur d'ordre 2 suivi d'un filtre $F(z)$ peuvent être remplacés par un filtre $F(z^2)$ suivit d'un décimateur. En appliquant successivement cette identité à une série de N filtres $F_i(z)$ suivi de décimateurs d'ordre 2, on obtient un filtre $F_{total}(z)$ suivi d'un décimateur d'ordre 2^N, avec

$$F_{total}(z) = \prod_{i=0}^{N-1} F_i\left(z^{2^i}\right) \tag{2.25}$$

On peut donc voir chaque détail et approximation comme un filtre passe-bande suivi d'un décimateur. Les détails sont le résultat d'une série de filtres passe-bas, qui coupe les fréquences

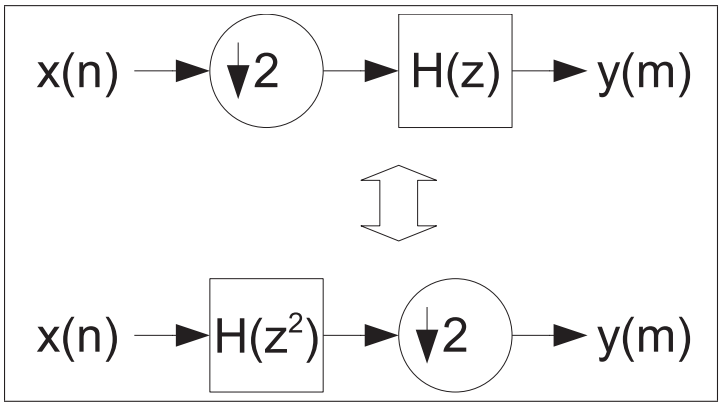


Figure 2.9 Identité de Noble.

supérieures, suivi d'un filtre passe-haut, qui coupe une partie des fréquences inférieures. L'approximation finale est un filtre passe-bas. Les caractéristiques exactes des ces filtres dépendent de l'ondelette utilisée. Comme les filtres coupent à $\pi/2$, on peut avoir une idée approximative du comportement en fréquence de la transformée pour toutes les ondelettes. Plus l'échelle est élevée, plus le découpage des basses fréquences est fin : le premier filtre passe-bas coupe à $f_{sorg}/4$, le second à $f_{sorg}/8$ et ainsi de suite.

Comme les filtres utilisés pour les détails sont des filtres passe-hauts, le décimateur va avoir un effet perceptible dans la transformée de Fourier continue. Les approximations dépendent uniquement de filtre passe-bas. Chaque filtre coupe la moitié supérieur du signal original. Si on suppose que les filtres sont idéaux et donc que $X(2\pi f/f_{sorg}) = 0$, pour $f_{sorg}/4 \leq f < f_{sorg}$, l'équation 2.24 devient :

$$X_2\left(2\pi\frac{f}{f_{sorg}/2}\right) = \frac{1}{2}\left[X(2\pi f/f_{sorg}) + X\left(2\pi f/f_{sorg} + \frac{2\pi f_{sorg}/4}{f_{sorg}}\right)\right] = \frac{1}{2}\left[X(2\pi f/f_{sorg})\right] \tag{2.26}$$

Ceci éviterait le recouvrement. Comme les filtres ne sont pas idéaux, il va y avoir un peu de recouvrement, qui devrait être réduit par le filtre. Pour les filtres passe-haut, cette condition n'est *pas* respectée. Si on suppose un filtre passe-haut idéal, avec $X(2\pi f/f_{sorg}) = 0$, pour $0 \leq f < f_{sorg}/4$, l'équation 2.24 devient :

$$X_2\left(2\pi\frac{f}{f_{sorg}/2}\right) = \frac{1}{2}\left[X\left(2\pi f/f_{sorg} + \frac{2\pi f_{sorg}/4}{f_{sorg}}\right)\right] \tag{2.27}$$

On voit que la décimation va déplacer les fréquences supérieures vers 0. Il faut se rappeler que cette transformée correspond à une transformée en Z, évaluée sur le cercle unitaire $e^{j\omega}$. Cette fonction a donc une période de 2π. Si on réécrit l'équation 2.21 en fonction de $z = e^{j\omega}$, on obtient :

$$\begin{aligned} & = \tfrac{1}{2}\left[X\left(e^{j\omega/2}\right) + X\left(-e^{j\omega/2}\right)\right] \\ X_2\left(e^{j\omega}\right) & = \tfrac{1}{2}\left[X\left(e^{j\omega/2}\right) + X\left(e^{j\pi}e^{j\omega/2}\right)\right] \\ & = \tfrac{1}{2}\left[X\left(e^{j\omega/2}\right) + X\left(e^{j\omega/2+j\pi}\right)\right] \end{aligned} \tag{2.28}$$

Comme $X\left(e^{j\pi+j\theta}\right) = X^*\left(e^{j\pi-j\theta}\right)$, où $X^*(z)$ est la conjuguée complexe de $X(z)$, on voit qu'en plus d'être périodique, $X(e^{j\omega})$ est symétrique autour de $\omega = \pi$. Si on combine ces propriétés, on voit que les hautes fréquences sont déplacées prés de $\omega = 0$ et que, pour $0 \leq \omega < \pi/2$, l'ordre des fréquences réelles est inversé : $\omega = 0$ correspond à $f_{sorg}/2$ et $\omega = \pi/2$ correspond à $f_{sorg}/4$. La figure 2.10 illustre les sections du cercle unitaire de la transformée en Z originale et leurs positions dans la transformée en Z du signal décimé. On y voit clairement les sections qui seront superposées après la décimation.

2.5 Paquets d'ondelettes

On peut ajouter des branches à la pyramide de décomposition, ce qui permet de découper les hautes fréquences en plus des basses. Ainsi, on peut aller chercher une résolution en fréquence plus fine, au prix d'une résolution temporelle plus réduite.

Ce découpage du plan temps-fréquence permet d'aller chercher des résolutions variables en fréquence. On sait que le système auditif humain peut être modélisé par une batterie de filtres passes-bande dont la bande passante augmente avec la fréquence centrale. La transformée en ondelettes discrète a une propriété similaire : la résolution fréquentielle peut être réduite afin d'augmenter la résolution temporelle. Ainsi, on peut obtenir une meilleur résolution temporelle dans les hautes fréquences, sans réduire la résolution fréquentielle dans les basses fréquences. Comme on filtre et décime après les filtres passes-hauts, il faut tenir compte de l'inversion des fréquences ainsi causée par les décimateurs. Ce n'est pas important pour la transformation en paquets d'ondelettes, mais c'est essentiel pour relier les fréquences aux paquets. Chaque filtre

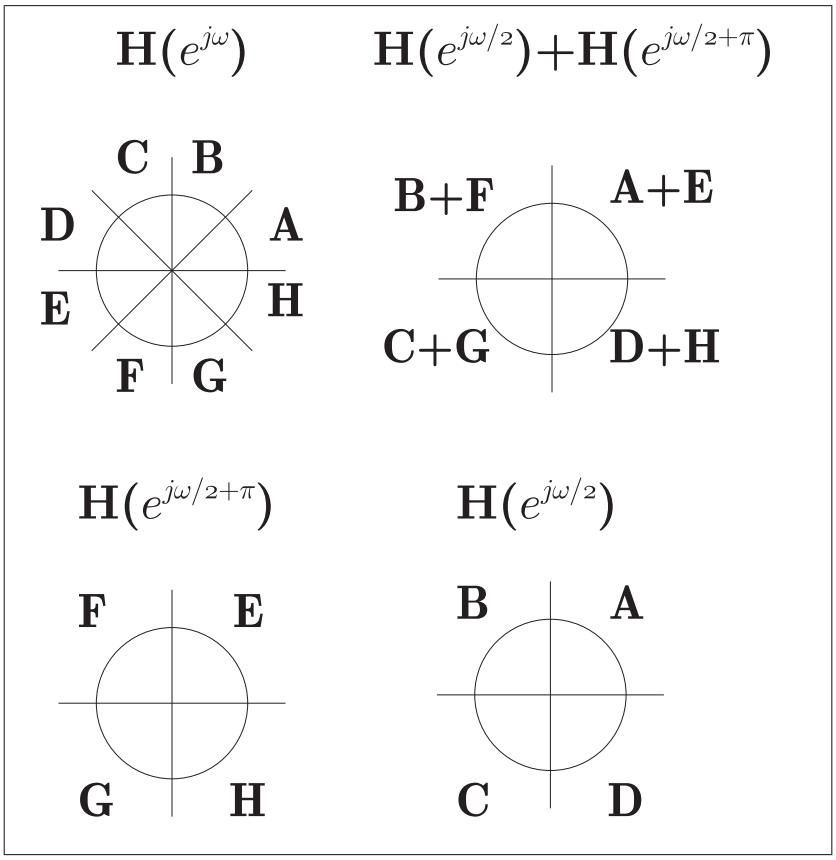


Figure 2.10 Effet d'une décimation par 2 sur la transformée en Z.

passe-haut inverse les fréquences.

2.6 Modèles du langage

Le phonème est le plus petit élément qui doit changer pour changer la signification d'un mot (Boite & Kunt, 1987). Un mot est donc composé de phonèmes. Contrairement aux mots, les phonèmes n'ont pas de sens par eux mêmes. On peut donc observer une structure pyramidale allant de la phrase vers le phonème. De la même manière, un système de reconnaissance automatique de la parole peut utiliser comme symboles fondamentaux des phrases, mots ou phonèmes. Le nombre élevé de phrases possibles rend habituellement la reconnaissance au niveau de la phrase impossible. Selon l'application, le nombre de mots dans le vocabulaire peut permettre la reconnaissance au niveau du mot. Dans les autres cas, il faut reconnaître des phonèmes et ensuite les regrouper en mots et en phrases. Les caractéristiques d'un phonème sont

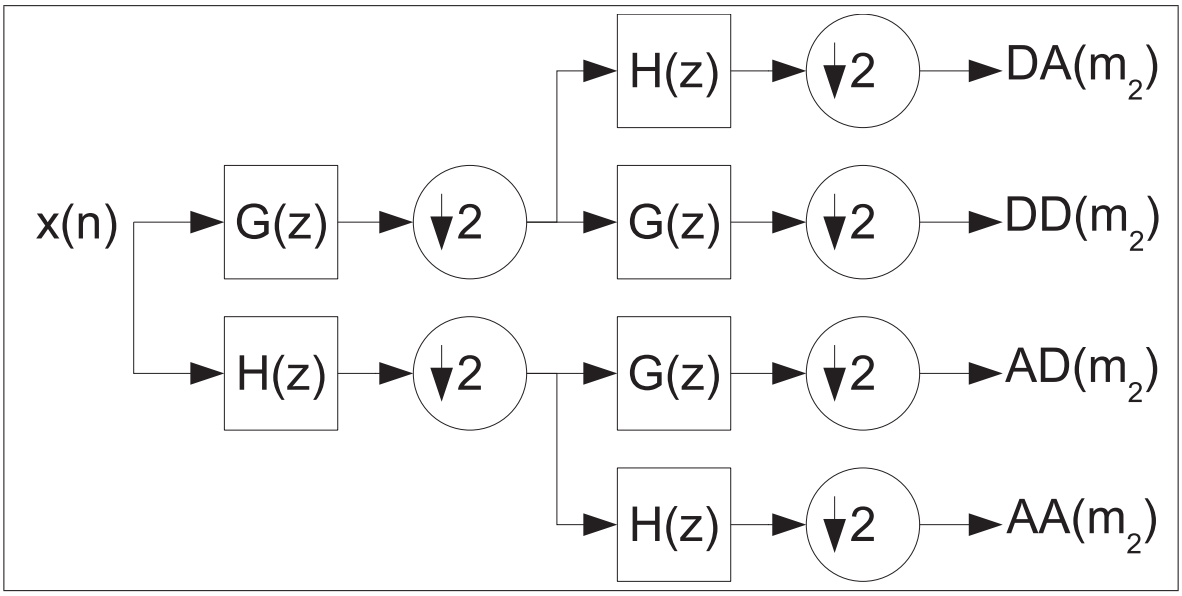


Figure 2.11 Pyramide de décomposition en paquets d'ondelettes.

le résultat direct de sa production. Elles ont donc été présentées dans la section 1.3.1, page 7.

Un simple dictionnaire permet de passer du phonème vers le mot. Ce dictionnaire limite les séquences de phonèmes possibles : seules les séquences qui correspondent aux mots qu'il contient sont acceptables. Il faut également regrouper les mots en phrases. Pour ce faire, on utilise un modèle de langage.

Selon l'application, ce modèle peut être très simple ou très complexe. Il peut s'agir d'une simple boucle de mots dans laquelle tout les mots du vocabulaire se suivent sans ordre. À l'autre extrême, on pourrait utiliser un modèle qui tient compte de la signification des mots et du sens de la phrase. Un tel modèle relève cependant plus du domaine de l'intelligence artificielle que de la reconnaissance de la parole. En pratique, le choix du modèle utilisé va dépendre de l'application. Le modèle permet de limiter l'espace de recherche lors de la reconnaissance. Lorsque un mot semble être terminé, c'est le modèle du langage qui fournit les mots suivants.

2.6.1 Grammaire artificielle

Une grammaire artificielle sera généralement utilisée quand le domaine d'application permet de limiter les phrases possibles. Il s'agit d'une séries de règles, qui limite les phrases possibles. Ces règles sont exprimées sous la forme d'une machine à états finis (Huang *et al.*, 2001). On peut ainsi produire un nombre, potentiellement infini, de phrase. Par exemple, un système utilisé pour le contrôle d'un téléphone cellulaire pourrait utiliser une grammaire artificielle. Un tel système utiliserait un nombre limité de commandes, suivi d'un nombre limité de paramètres. Ce type de système décrit seulement les transitions mots à mots possibles et ne tient pas compte de la probabilité d'une phrase ou d'une transition. Il peut inclure des boucles infinies de mots, par exemple une série de chiffres pour composer un numéro de téléphone quelconque, ou des chemins parallèles, par exemple le nom de la personne plutôt que son numéro de téléphone.

2.6.2 n-Grammes

La grammaire et la structure du langage naturel sont très complexes, et peuvent difficilement être représentées par une grammaire artificielle. Par contre, on peut voir le langage comme une chaîne de Markov : la probabilité du prochain mot dépend des mots précédents. On utilise alors des n-grammes (Huang *et al.*, 2001). Le « n »de n-Grammes indique le nombre de mots utilisés pour définir une transition. Ainsi, un monogramme dépend uniquement du mot en question et ne tient pas compte de ce qui précède : la probabilité d'un mot est donc uniquement fonction du mot lui même. Un bigramme tient compte du mot précédent : la probabilité d'un mot est donc fonction du mot et du mot qui le précède. Un modèle basé sur des n-grammes est donc défini par :

$$P\left(w_i|w_{i-1}, w_{i-2}, \cdots, w_{i-(n-1)}\right) \tag{2.29}$$

c'est-à-dire la probabilité que le i-éme mot soit w_i selon les $n-1$ mots précédents. On dit que l'ordre d'un n-gramme est n. Pour un monogramme $n = 1$, on a :

$$P\left(w_i\right) \tag{2.30}$$

Pour un bigramme $n = 2$, on a :

$$P(w_i|w_{i-1}) \tag{2.31}$$

Pour un trigramme $n = 3$, on a :

$$P(w_i|w_{i-2}, w_{i-1}) \tag{2.32}$$

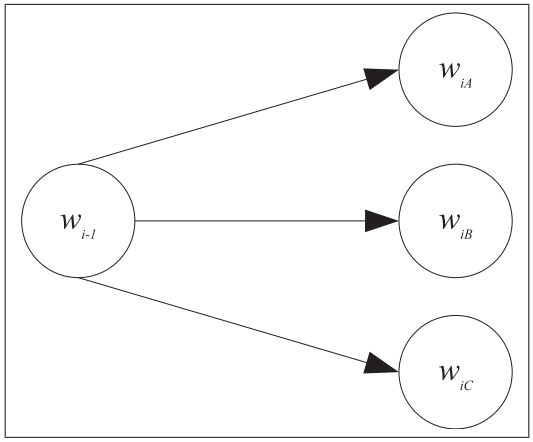


Figure 2.12 Exemple de bigramme.

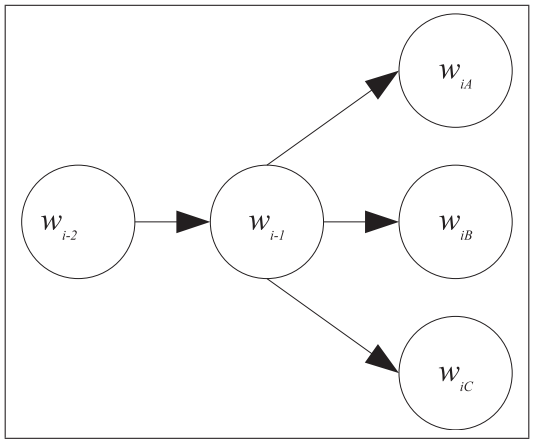


Figure 2.13 Exemple de trigramme.

Il faut évidement évaluer ces probabilités. Pour ce faire, on doit utiliser un texte d'entraînement, qui contient des phrases typiques en grand nombre. Pour un monogramme, il s'agit simplement

de calculer la fréquence du mot dans le texte. Quand l'ordre est plus élevé, on doit utiliser la fréquence du mot et de ce qui le précède :

$$P\left(w_i|w_{i-1},w_{i-2},\cdots,w_{i-(n-1)}\right)=\frac{C\left(w_{i-n+1},w_{i-n+2},\cdots,w_{i-1},w_{i)}\right)}{C\left(w_{i-n+1},w_{i-n+2},\cdots,w_{i-2},w_{i-1}\right)} \tag{2.33}$$

où $C(\mathbf{W})$ est le nombre d'occurrences de la séquence de mots $\mathbf{W}$ dans le texte. Pour un trigramme $n = 3$, on a donc :

$$P\left(w_i|w_{i-2},w_{i-1}\right)=\frac{C\left(w_{i-2},w_{i-1},w_i\right)}{C\left(w_{i-2},w_{i-1}\right)} \tag{2.34}$$

Ceci illustre un problème dans l'évaluation : quand le vocabulaire est large, le nombre de paires de mots possibles est très élevé. Plusieurs séquences ne seront pas présentes, et auront donc une probabilité nulle. Ces séquences absentes vont donc rendre la reconnaissance des phrases qui les contiennent impossible. Or, plusieurs de ces phrases seront valides et doivent être reconnaissables. Des techniques d'aplatissement permettent de corriger ce problème en augmentant la probabilité des séquences rares ou absentes et en réduisant la probabilité d'autres séquences.

Une méthode d'aplatissement couramment utilisée est celle de Katz (1987). Pour un bigramme, elle est définie par :

$$P_{Katz}(w_i|w_{i-1})=\begin{cases} C(w_{i-1},w_i)/C(w_{i-1}) & r>k \\ d_r C(w_{i-1},w_i)/C(w_{i-1}) & k\geq r>0 \\ \alpha(w_{i-1})P(w_i) & r=0 \end{cases} \tag{2.35}$$

où $r = C(w_{i-1},w_i)$, d_r est un facteur de réduction dépendant de r et $\alpha(w)$ est le facteur d'augmentation. Cette méthode tient compte de la fréquence des mots qui constituent le bigramme pour déterminer la probabilité corrigée des bigrammes absents. Ce type d'approche réduit donc temporairement l'ordre du modèle pour les n-grammes absents. Le paramètre k est fixé. On suppose que les bigrammes présents plus de k fois sont représentatifs de leurs fréquences réelles. Il faut donc trouver des valeurs de d_r et $\alpha(w)$ adéquates.

Les valeurs de d_r sont basées sur l'estimation de Good-Turing (Huang *et al.*, 2001). Pour cette estimation, ont ajuste le compte d'un n-gramme spécifique selon le nombres de n-gramme qui

ont ce compte. Si $r = C(\mathbf{W})$, le compte estimé est :

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \tag{2.36}$$

où n_r est le nombre de n-grammes présents exactement n fois dans le texte ou la fréquence de la fréquence r. Cette estimation ajuste donc les comptes selon leurs fréquences. En particulier, pour les n-grammes absents ($r = 0$) l'estimation sera $r^* = n_1/n_0$. Pour chaque fréquence, le nombres d'instances $N = n_r \cdot r$. Ainsi, après ajustement, on estimera $n_0 \cdot n_1/n_0 = n_1$ instances des n-grammes absents.

Comme le nombre de n-grammes contenus dans le texte doit être constant, il faut faire en sorte que les comptes ajoutés aux n-grammes absents soient également retirés d'autre n-grammes. Comme les n-grammes présents plus de k fois ne sont pas ajustés et l'estimation pour les n-grammes absents est n_1, il faut redistribuer n_1 comptes des n-grammes présents entre 1 et k fois. Le facteur de réduction d_r est donc supposé proportionnel à r^*/r tel que :

$$1 - d_r = \mu\left(1 - \frac{r^*}{r}\right) \tag{2.37}$$

Comme le nombre de n-grammes ne doit pas varier, on peut également ajouter la condition :

$$\sum_{r=1}^{k} n_r(1 - d_r) = n_1 \tag{2.38}$$

Ces deux conditions permettent de trouver la valeur de d_r :

$$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}} \tag{2.39}$$

En réduisant la probabilité d'un bigramme, le ration d_r en réduit également le compte $C(w_{i-1}, w_i)$. Il faut ensuite redistribuer ces comptes. Le facteur $\alpha(w)$ permet cette redistribution :

$$\alpha(w) = \frac{1 - \sum_{v:C(w,v)>0} P_{Katz}(v|w)}{1 - \sum_{v:C(w,v)>0} P(v)} \tag{2.40}$$

$$P(v) = \frac{C(v)}{N} \tag{2.41}$$

On constate donc que $P_{Katz}(\mathbf{W})$ est récursif. Cependant, comme les d_r et α ne varient pas, ils peuvent être précalculés et simplement être inclus dans le modèle du langage.

Ainsi, la technique de Katz utilise les comptes réels quand le nombre d'instances est suffisamment élevé, réduit les comptes en utilisant la technique de Good-Turing quand le compte est petit mais non-zéro, et augmente artificiellement le compte quand il est de zéro. De plus, l'augmentation artificielle tient compte du n-gramme d'ordre inférieur, c'est-à-dire du monogramme pour un modèle de langage basé sur des bigrammes.

2.7 Chaînes de Markov cachées

Les chaînes des Markov ont été brièvement présentées dans la section 2.6.2. On peut également les utiliser pour modéliser un phonème. Dans ce cas, on utilise habituellement une chaîne de Markov d'ordre 1. Les états permettent de représenter les variations des caractéristiques du phonème dans le temps et d'inclure les variations de durée d'un phonème dans le modèle.

Pour le modèle du langage, on connaît l'état dans lequel se trouve le modèle : il s'agit du mot et de ceux qui le précèdent. Ce n'est pas le cas pour un modèle de phonème : on ne sait pas quel est l'état qui a produit l'observation. On utilise alors une chaîne de Markov cachée. Dans une chaîne de Markov, seule la transition est incertaine : la sortie d'un état est fixe. Pour une chaîne de Markov cachée, la sortie d'un état varie. On ajoute donc une probabilité à chaque sortie possible pour un état. Il y a donc deux sources d'incertitude dans la chaîne : la sortie et la transition. Les paramètres sont :

A : La matrice de probabilités de transition, où a_{ij} est la probabilité de passer de l'état i à l'état j.

B : La matrice de probabilités de sortie, où $b_i(\overrightarrow{O})$ est la probabilité de que le vecteur-observation $\overrightarrow{O}$ soit produit par l'état i.

π **:** La matrice de probabilités initiale, où π_i est la probabilité que l'état initial soit l'état i.

On compare le vecteur-observation $\overrightarrow{O}$ à un modèle établi à partir des données d'entrainement

pour trouver la probabilité $b_i(\overrightarrow{O})$ d'une sortie selon l'état. Le vecteur-observation a été produit à partir du signal. C'est le résultat d'une série de transformations et d'autres traitements, mais il n'est pas important de connaitre sa méthode de production pour trouver **B**. La production du vecteur-observation est importante, et plusieurs techniques existent. Le chapitre 3 contient plus de détail sur ce sujet.

Durant la reconnaissance de la parole, on veut trouver le modèle le plus probable. On doit donc trouver la séquence d'états la plus probable. L'algorithme de Viterbi permet de trouver cette séquence (Huang *et al.*, 2001). On utilise une matrice **V**, qui contient la probabilité maximale $V_t(i)$ de passer par l'état i au temps t, et une matrice **C**. La matrice **C** va contenir le chemin emprunté. On initialise $V_1(i) = \pi_i \cdot b_i(\overrightarrow{O}_1)$ pour chaque état i. Ensuite, on itère pour chaque observation :

$$V_t(i) = b_i(\overrightarrow{O}_t) \cdot \max_j \left[V_{t-1}(i)a_{ij}\right]$$

Ceci va continuer les chemins précédents pour chaque états. On conserve uniquement la transition la plus probable qui conduit à un état i au temps t. Pour la matrice **C**, on ajoute l'état source de cette transition, tel que

$$C_t(i) = \arg\max_j \left[V_{t-1}(i)a_{ij}\right]$$

Lorsque toutes les observations ont été utilisées (au temps T,) on identifie l'état final le plus probable selon V_T. On utilise ensuite **C** pour trouver le chemin qui a conduit à cet état. Ainsi, on obtient la séquence d'état la plus probable selon une série d'observation et la probabilité en question. On peut représenter l'algorithme de Viterbi par un treillis, constitué d'une colonne par observation et d'une ligne par état.

La figure 2.14 est un exemple de chaîne de Markov, avec 3 états émetteurs, et un état de fin de séquence. Cet état peut servir à connecter deux chaînes. Si l'on ajoute une matrice **B** de probabilité d'émission, on obtient une chaîne de Markov cachée. Par exemple, avec la matrice

$$\mathbf{B} = \begin{bmatrix} 0.25 & 0.75 & 0.90 \\ 0.50 & 0.10 & 0.05 \\ 0.25 & 0.15 & 0.05 \end{bmatrix} = \begin{bmatrix} b_1(A) & b_2(A) & b_3(A) \\ b_1(B) & b_2(B) & b_3(B) \\ b_1(C) & b_3(C) & b_3(C) \end{bmatrix}$$

et la séquence d'observation $\{B,C,A,B\}$ ont obtient le treillis illustré par la figure 2.15. La probabilité de cette séquence est environ 0.5%. La séquence d'états correspondante est $\{1,1,2,2\}$ si touts les états peuvent être l'état final. Si l'état fin doit être l'état final, la séquence d'état est $\{1,1,2,3,fin\}$, et la probabilité est environ 0.02%.

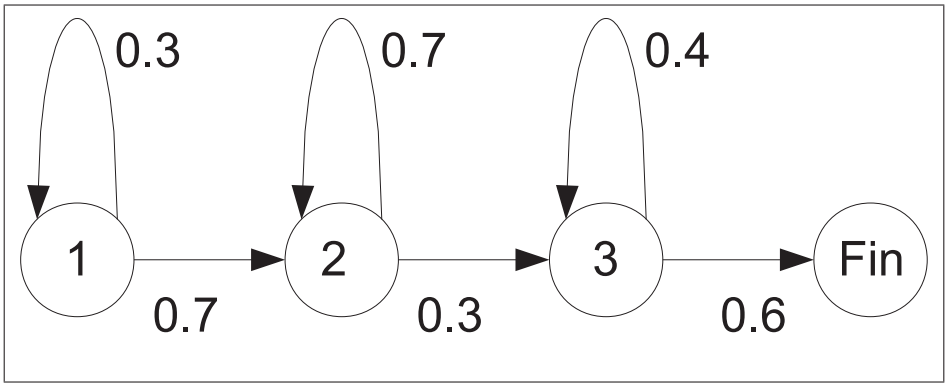


Figure 2.14 Exemple de chaîne de Markov.

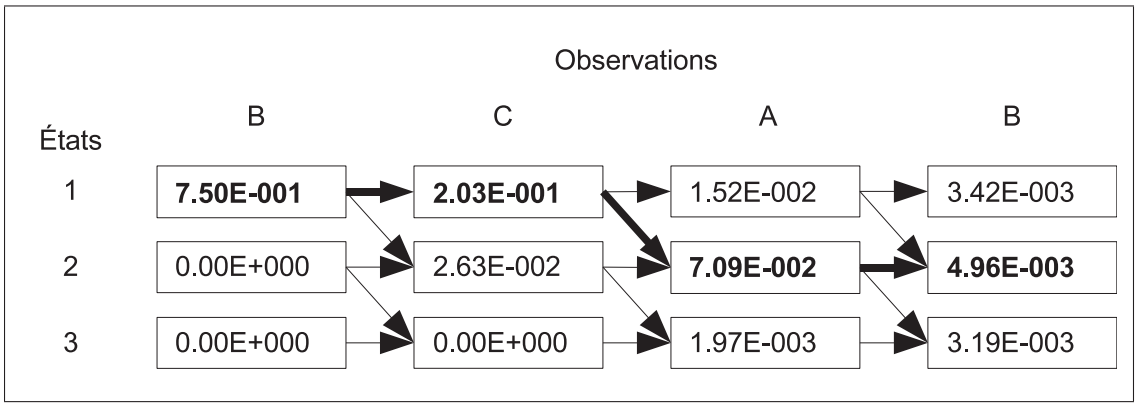


Figure 2.15 Exemple de treillis.

2.8 Analyse discriminante

L'analyse discriminante permet de retirer les lignes non-significatives du vecteur-observation. Cette analyse produit une matrice de transformation linéaire qui est multipliée par le vecteur-observation. On produit ainsi un nouveau vecteur-observation, composé d'une combinaison linéaire du vecteur original. La matrice de transformation est choisie de façon à minimiser

la variance à l'intérieur d'une classe (ici un état du modèle de Markov) tout en maximisant la variance entre les classes. Intuitivement, l'effet de ce processus est de concentrer la distribution des observations qui correspondent à une classe, tout en séparant les classes.

Kumar & Andreou (1998) ont présenté une approche qui permet de trouver la matrice de transformation optimale. Cette approche maximise le logarithme de la vraisemblance selon le modèle. Ce processus peut être facilement inclus dans la procédure de ré-estimation des paramètres des modèles. Avec $V_t(j)$, la probabilité d'être à l'état j au temps t et $\overrightarrow{o}_t$ le vecteur-observation au temps t, on définit :

$$N_j = \sum_{t=1}^{T} V_t(j) \tag{2.42}$$

$$N = \sum N_j \tag{2.43}$$

$$\bar{N}_j = N_j / N \tag{2.44}$$

$$\bar{X} = \frac{\sum_{t=1}^{T} \overrightarrow{o}_t}{N} \tag{2.45}$$

$$\bar{X}_j = \frac{\sum_{t=1}^{T} V_t(j) \overrightarrow{o}_t}{N_j} \tag{2.46}$$

$$\bar{W}_j = \sum_{t=1}^{T} \frac{V_t(j)}{N_j} (\overrightarrow{o}_t - \bar{X}_j)(\overrightarrow{o}_t - \bar{X}_j)^{\mathrm{T}} \tag{2.47}$$

$$\bar{W} = \sum \bar{N}_j \bar{W}_j \tag{2.48}$$

$$\bar{T} = \sum_{t=1}^{T} \frac{1}{N} (\overrightarrow{o}_t - \bar{X})(\overrightarrow{o}_t - \bar{X})^{\mathrm{T}} \tag{2.49}$$

$\bar{T}$ correspond à la variance inter-classe, alors que $\bar{W}_j$ correspond à la variance de la classe j. $\bar{X}$ est la moyenne globale, alors que $\bar{X}_j$ est la moyenne de la classe j. Si la matrice de covariance est diagonale, ce qui est le cas dans les tests qui seront présentés ultérieurement, Kumar & Andreou (1998) ont démontré qu'il faut trouver la matrice de transformation $\hat{\theta}_D$ qui maximise le logarithme de la vraisemblance :

$$\hat{\theta}_D = \underset{\hat{\theta}_D}{\operatorname{argmax}} \left\{ -\frac{N}{2} \log \left| \operatorname{Diag} \left(\theta_{n-p}^{\mathrm{T}} \hat{T} \theta_{n-p} \right) \right| - \sum_{j=1}^{J} \frac{N_j}{2} \log \left| \operatorname{Diag} \left(\theta_p^{\mathrm{T}} \hat{W}_j \theta_p \right) \right| + \log |\theta| \right\} \tag{2.50}$$

La matrice θ est une matrice carré, de dimension $n \times n$, où n est le nombre de lignes du vecteur-observation original. Elle est composée de deux sections : θ_p et θ_{n-p}. θ_p est la partie de la matrice qui produit la transformation : les p colonnes correspondent aux p lignes du vecteur-observation final alors que θ_{n-p} est le reste de la matrice. La transformation se fait avec $\theta_p^{\mathrm{T}} \overrightarrow{o}_t$.

2.9 Conclusion

La transformée de Fourier permet d'extraire une représentation du signal en fonction de la fréquence. On peut donc savoir ce qui se produit, mais on ne peut pas savoir *quand* cela se produit. On peut utiliser une fonction fenêtre pour découper le signal en plus petites tranches. Ce découpage va limiter la résolution fréquentielle de la transformée, mais va permette d'augmenter la résolution temporelle. La fonction fenêtre va également distordre le résultat de la transformée.

La transformée en ondelettes permet de varier les résolutions temporelles et fréquentielles. Cette transformée ne demande pas l'utilisation d'une fonction fenêtre, puisque le fonction d'analyse est localisée dans le temps. Plusieurs fonctions d'analyses peuvent être utilisées. La transformée en paquets d'ondelettes permet de découper le plan temps-fréquence pour obtenir la résolution temporelle voulue dans une bande de fréquence, en réduisant la résolution fréquentielle, ou d'augmenter la résolution fréquentielle en réduisant la résolution temporelle.

Ces deux transformées permettent d'extraire l'information pertinente du signal et de produire des vecteurs-observations. On peut ensuite utiliser des chaînes de Markov cachées pour modéliser les symboles à reconnaitre. Les vecteurs-observations vont permettre d'établir la probabilité qu'un état de la chaîne produise l'observation correspondante. L'analyse discriminante permet de raffiner les vecteurs-observations en supprimant les lignes du vecteur peu significatives au profit des lignes significatives. Ainsi, les modèles contiennent l'information qui permet de reconnaitre un symbole et excluent le plus d'information non-significative possible. Le modèle du langage utilisé permet ensuite de relier les symboles qui ont été reconnus entre eux.

CHAPITRE 3

RECONNAISSANCE DE LA PAROLE

3.1 Introduction

La reconnaissance de la parole consiste à produire une représentation textuelle de la parole à partir d'un signal sonore correspondant. C'est donc un processus qui extrait des symboles discrets d'un signal, un peu comme un système de communication digital transforme une forme d'onde en 1 et 0. Une fois ces symboles obtenus, on peut les utiliser de plusieurs façons. Par exemple, pour manipuler des fenêtres dans une interface usager graphique, pour écrire un texte, pour produire les sous-titres d'une émission de télévision ou pour composer un numéro de téléphone. Le choix de l'application va informer la conception et les limites du système de reconnaissance, mais les techniques utilisées seront essentiellement les mêmes. Selon les besoins, ont peut utiliser de la reconnaissance de forme, qui compare le signal, habituellement transformé en vecteur-observation, à un modèle et utilise une mesure de distance pour déterminer le symbole, ou encore on peut utiliser un système basé sur la probabilité que l'observation ai été produite par un modèle défini statistiquement à partir d'exemple de parole. Dans les deux cas, le signal est comparé à un modèle. Le type de modèle et la nature de la comparaison changent, mais la comparaison elle-même reste.

Dans ce chapitre, on présentera d'abord un système de reconnaissance probabiliste. Nous verrons ensuite la notion de robustesse aux variations, les types d'approche qui permettent d'obtenir cette robustesse et les variations en question. Finalement, on a mentionné précédemment et à plusieurs reprises le vecteur-observation. Ce vecteur caractérise la parole pour une tranche de temps spécifique. Il est le résultat de transformations et de traitement du signal. Plusieurs méthodes existent pour produire ces vecteurs. Deux seront présentées dans ce chapitre :

1. Les coefficients cepstraux avec fréquence de Mel (Davis & Mermelstein, 1980), qui ont été choisis pour leur ubiquité ;

2. Les coefficients de prédiction linéaire cepstraux perceptuels (Hermansky *et al.*, 1992), qui ont été choisis pour leurs robustesse aux variations de canal.

3.2 Présentation d'un système de reconnaissance de la parole

On peut produire des systèmes de reconnaissance automatiques de la parole en utilisant des chaînes de Markov cachées. Dans ces systèmes, on produit un treillis à partir du modèle de langage, de l'expansion en phonèmes des mots et finalement à partir de la chaîne de Markov correspondant à un phonème. Si le vocabulaire est réduit, il est possible de représenter les mots par des chaînes de Markov, sans avoir de modèles pour les phonèmes. On peut ensuite appliquer l'algorithme de Viterbi à ce treillis.

Dans tous les cas, le signal qui contient la parole devra être converti en vecteur-observation. Ce traitement se fera à intervalles réguliers (le pas de traitement), pour une section du signal. On trouvera la probabilité que ce vecteur soit produit par un état donné. Le modèle de langage va permettre va relier les chaînes de Markov, ce qui va produire une phrase. On obtiendra ainsi, finalement, la phrase qui a le plus de chance d'être produite par les modèles. Il faut noter que cette méthode ne peut pas conclure qu'une phrase n'a pas été produite par les modèles. Par exemple, un système qui doit uniquement détecter des chiffres ne peut pas conclure que la phrase ne contient pas de chiffre. On doit donc inclure des chaînes de Markov pour les silences et, en particulier si le vocabulaire est réduit, pour les mots qui ne sont pas dans le vocabulaire. Par exemple, si ont veux détecter les mots « OUI » et « NON », il faudra au minimum quatre modèles : « OUI », « NON », « SILENCE » et « AUTRE ». On pourrait également utiliser des modèles qui décrivent les phonèmes des mots « OUI » et « NON », en plus des deux autres modèles. Il faudra également un modèle de langage approprié au problème. Ce modèle va limiter les séquences possibles. Ainsi, si l'on désire seulement un mot, il devrait uniquement permettre des phrases composées d'un seul symbole, en plus des silences. Pour des modèles plus complexes, on va tenir compte de la fréquence d'un mot ou d'une séquence de mots dans les textes d'entrainement.

Les modèles de langages et les chaînes de Markov cachées ont été présentés dans le chapitre 2. Plusieurs méthodes existent pour produire le vecteur-observation. Deux seront décrites ultérieurement et une nouvelle sera proposée au chapitre 4.

3.3 Reconnaissance robuste

Les systèmes de reconnaissance de la parole sont basés sur une comparaison entre le signal de parole à reconnaître et des données d'entraînement. La reconnaissance robuste de la parole doit fonctionner dans un environnement différent de l'environnement d'entraînement. Même dans les conditions les plus idéales, le signal à reconnaître sera différent des signaux qui ont été utilisés pour l'entraînement. Il s'agit donc d'une question de degré : un système est plus ou moins robuste à certaines différences.

La taille et la nature du vocabulaire affectent les performances du système. Un système peut avoir un taux d'erreur réduit avec un vocabulaire réduit, même en présence de fortes différences, et un taux d'erreur large avec un vocabulaire large, même en présence de faibles différences.

Ainsi, un système idéal serait robuste à toutes différences, peu importe son vocabulaire. Un tel système n'est pas présentement possible. Il faut donc choisir le type de système de reconnaissance de la parole en fonction des conditions d'utilisations prévues.

3.3.1 Utilité de la reconnaissance robuste

Pour être vraiment utile, la reconnaissance de la parole doit avoir un faible taux d'erreur. En effet, la correction des erreurs du système n'est pas toujours facile, ou même possible. Chaque erreur ralentit et stresse l'utilisateur en le forçant à corriger ou même recommencer l'opération.

Dans des conditions idéales, le taux d'erreur est suffisamment réduit pour de nombreuses applications. Malheureusement, la plupart des environnements où la reconnaissance de la parole

est désirée ne sont pas idéaux. Ils présentent une ou plusieurs difficultés, auxquelles le système doit être robuste.

3.3.2 Difficultés et variations

On peut distinguer deux grandes catégories de variations : celles reliées à l'environnement et celles reliées au locuteur.

Les variations environnementales sont celles qui modifient la parole après sa production. Il s'agit du bruit ambiant, des distorsions introduites par le canal et du bruit ajouté par le canal.

Le variation reliées au locuteur sont celles qui modifient la production de la parole. Elles incluent les variations intra-locuteur et inter-locuteurs. Il faut noter que le locuteur est conscient de l'environnement et modifie sa parole en conséquence. On considère que ces variations sont reliées au locuteur, même si l'environnement en est partiellement responsable.

L'environnement est rarement parfaitement silencieux. Le niveau et la nature du bruit ambiant rendent la parole plus difficile à reconnaître.

Si le rapport signal-à-bruit (RSB) est faible, il est difficile de distinguer les périodes d'activités et d'inactivités vocales. De plus, les caractéristiques du bruit vont être importantes par rapport aux caractéristiques du signal.

Il est possible d'entraîner un système de reconnaissance pour un RSB donné. Ceci améliore les performances dans les conditions d'entraînement. Une augmentation ou *diminution* du RSB va réduire les performances. Das *et al.* (1993) présentent un bon exemple de ce phénomène.

Avant d'être traitée, la parole passe par un canal qui a une réponse fréquentielle. Ce canal introduit des distorsions. Si la réponse fréquentielle utilisée pour l'entraînement est différente

de celle utilisée durant la reconnaissance, ces distorsions vont réduire les performances.

Le canal inclut les chemins entre la source et le microphone ainsi que le microphone lui-même. On peut également ajouter le système de communication qui transmet le signal capturé au système de reconnaissance de la parole. Chaque microphone a une réponse fréquentielle qui varie selon le type de microphone. De plus, la position de la source par rapport au microphone influence la réponse fréquentielle. Cette variation est plus ou moins prononcée selon le type de microphone. Das *et al.* (1993) présentent un bon exemple de ce phénomène. On voit bien que la réponse fréquentielle du canal varie facilement :

1. On ne peut pas toujours forcer le choix du microphone ;
2. On ne peut pas toujours prévenir les changements de microphone ;
3. On ne peut pas toujours contrôler la position du locuteur par rapport au microphone.

Les système de reconnaissance de la parole doivent souvent reconnaître la parole de plusieurs locuteurs. Chaque locuteur ne s'exprime pas de la même façon. Un système entraîné pour un locuteur ne sera pas aussi performant avec un autre locuteur. Benzeghiba *et al.* (2007) passent en revue quelques unes des causes de ces variations.

La parole spontanée est différente de la lecture à voie haute. Il y a des différences dans la prononciation et des différences dans la structure. Certains phonèmes sont sous-prononcés, tandis que d'autres sont sur-prononcés. Adda-Decker *et al.* (2005) présentent quelques unes de ces altérations pour le français. On remarque également des pauses, hésitations, répétitions et d'autres changements dans la structure de la phrase. Byrne *et al.* (2004) présentent, entre autres, ce genre de difficultés.

Le rythme de la parole est une autre variation importante. Il ne s'agit pas simplement de réduire ou augmenter la durée des phonèmes. La variation de la durée des phonèmes n'est pas uniforme. La prononciation est aussi modifiée. Janse (2004) montre bien ce phénomène.

Les caractéristiques spectrales de la parole changent selon l'état émotionnel du locuteur. Comme ces caractéristiques sont la base de la reconnaissance de la parole, tout les systèmes de reconnaissances sont affectés. En plus des émotions, le réflexe Lombard est une adaptation de la parole au bruit ambiant. Bou-Ghazale & Hansen (2000) présentent l'effet de deux émotions et du réflexe Lombard sur la reconnaissance automatique de la parole. De plus, l'effet est différent si le locuteur tente de communiquer avec le système, ou si il récite une liste (Junqua *et al.*, 1999).

Un taux d'erreurs élevé peut stresser, frustrer ou même enrager le locuteur. Ce changement émotionnel risque d'augmenter le taux d'erreur, résultant en une boucle de mauvaise performance. Le locuteur risque de devenir hostile aux systèmes de reconnaissances. L'effet de l'état émotionnel sur les systèmes de reconnaissances est, indirectement, une autre raison de l'importance de la robustesse.

3.3.3 Types d'approche pour améliorer la robustesse

On peut améliorer la robustesse d'un système de plusieurs façons. On peut distinguer deux grandes catégories : celles reliées à l'analyse du signal sonore et celles reliées au modèle de la parole.

Les approches reliées à l'analyse du signal sonore tentent d'en extraire l'information requise pour la reconnaissance et d'exclure les éléments superflus. On veut idéalement une représentation du son qui ne contient que les éléments reliés à la parole et qui représente fidèlement les éléments importants.

Les approches reliées au modèle de la parole tentent de mieux décrire les unités linguistiques qui la constituent. Le modèle peut agir à plusieurs niveaux, de la phonème à la phrase. On pourrait même inclure la compréhension de la parole.

3.4 Coefficients à spectre relatif pour la reconnaissance robuste

On sait que le conduit vocal peut être vu comme un filtre. Les poumons et cordes vocales fournissent l'excitation sous forme de l'onde glottale ou d'un bruit blanc. On peut modéliser ce filtre afin d'extraire des caractéristiques utiles pour la reconnaissance de la parole. Le cepstre transforme la convolution de l'excitation et de la réponse à l'impulsion du filtre en addition. Si on tient compte des bandes critiques lors de la production du cepstre, on obtient des coefficients cepstraux avec fréquence de Mel (Davis & Mermelstein, 1980). Un filtre auto-régressif est également bien adapter à la description du conduit vocal : on peut facilement en obtenir les coefficients, et il décrit bien le spectre de puissance du signal. On peut également tenir compte des limites du système auditif humain dans cette technique. On obtient alors des coefficients de prédiction linéaire perceptuels (PLP), qui ont été présentés par Hermansky *et al.* (1986). On peut calculer des coefficients cepstraux à partir de ces coefficients. De plus, on peut ajouter d'autres étapes dans la production de ces coefficients afin d'augmenter la robustesse aux changements de canal. On obtient alors des coefficients à spectre relatif (RASTA-PLP ou RPLP) qui ont été présentés par Hermansky *et al.* (1992).

Nous verrons d'abord comment calculer directement le cepstre d'un signal, en tenant compte des bandes critiques. Le modèle auto-régressif sera également présenté. Ensuite, nous verrons comment ajouter le modèle perceptuel, incluant l'élément RASTA. Les coefficients de prédiction linéaire cepstraux perceptuels (Hermansky *et al.*, 1992) qui en résulte sont particulièrement robustes aux variations de canal.

3.4.1 Coefficients cepstraux avec fréquence de Mel

Davis & Mermelstein (1980) ont présenté les MFCCs (Mel-Frequency Cepstral Coefficients) comme représentation de la parole pour la reconnaissance automatique. Ces coefficients sont courament utilisés pour produire des vecteurs-observations. Les MFCCs sont basés sur une batterie de filtres passe-bandes triangulaires, dont la fréquence centrale est placée linéairement

dans les fréquences de Mel. Les bandes passantes correspondent aux bandes critiques du système auditif humain. On utilise habituellement une transformée de Fourier court-terme, suivi d'une série de convolution pour implémenter cette étage. Ensuite, on applique une transformation cepstrale pour obtenir les coefficients.

La transformation cepstrale est la transformée en cosinus discrets du logarithme de $|X(i)|$, l'énergie de la transformée de Fourier discrète de $x(n)$:

$$C_k = \sum_{i=0}^{N-1} \log |X(i)| \cos\left(\frac{\pi k}{N} i\right) \tag{3.1}$$

On sait que la convolution dans le temps devient une multiplication dans le domaine fréquentiel. On sait également que $\log AB = \log A + \log B$. Ainsi, la convolution dans le temps devient une addition dans le cepstre. Comme la parole est produite par la convolution de la réponse à l'impulsion du conduit vocal et d'un bruit blanc (pour les sons non-voisés) ou de l'onde glottique (pour les sons voisés), le cepstre correspond à l'addition de ces deux signaux. De plus, les composants périodiques du spectre produiront un pic dans le cepstre. Ceci permet de bien identifier les formants d'un phonème. Ainsi, le cepstre met en évidence les caractéristiques du conduit vocal. Comme ces caractéristiques différencient les phonèmes, il s'agit évidement d'une propriété utile dans la reconnaissance vocale.

Pour les MFCCs, on n'utilise pas directement le résultat de la transformée de Fourier discrète. On va plutôt utiliser les sorties d'une batterie de filtres passe-bandes triangulaires qui se chevauchent. Ces filtres sont placés à intervalles réguliers sur l'échelle de Mel. Ainsi, les filtres placés dans les basses fréquences sont à intervalles presque linéaires en Hz, tandis que les filtres placés dans les hautes fréquences sont à intervalles logarithmiques. Ainsi, les MFCCs d'un signal sont :

$$MFCC_k = \sum_{i=0}^{N-1} \log F_i \cos\left(\frac{\pi k}{N} i\right) \tag{3.2}$$

où F_i est l'énergie du filtre i.

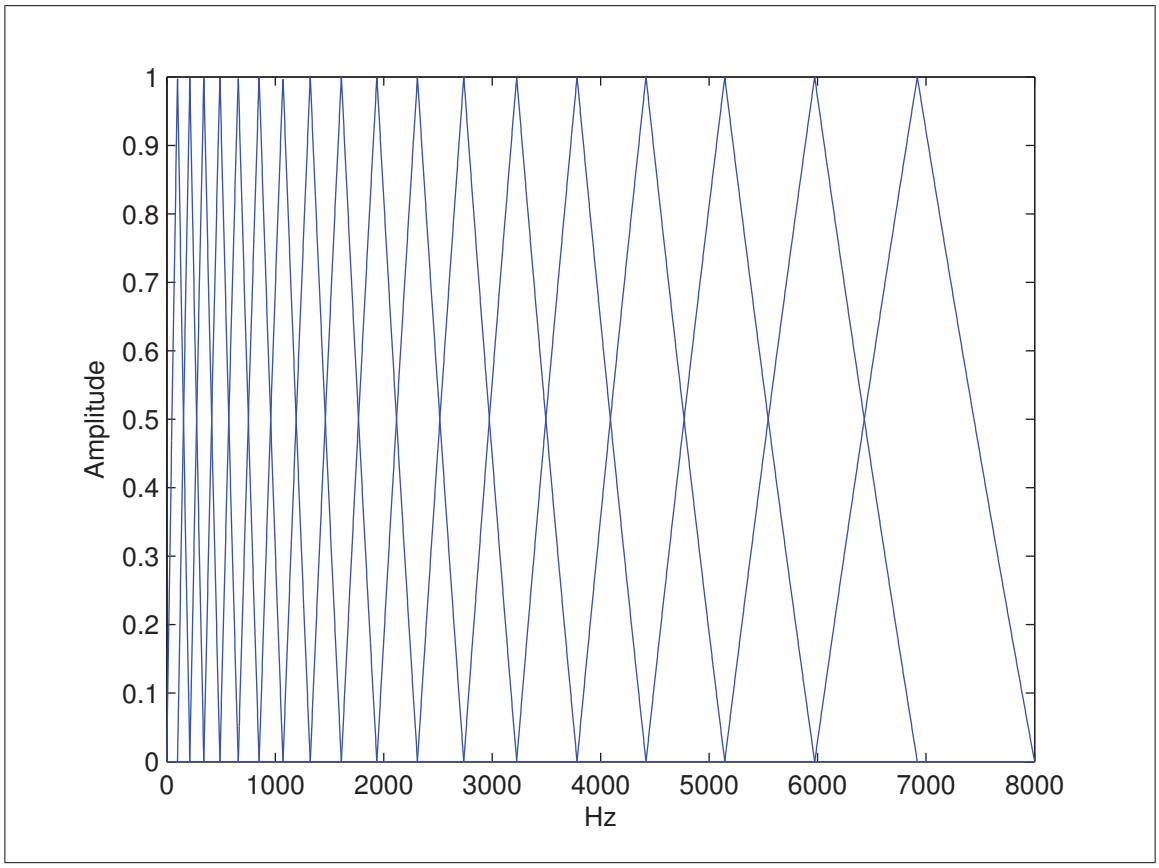


Figure 3.1 Batterie de filtres (échelle en Hertz).

3.4.2 Modèle auto-régressif

On suppose que le conduit vocal a une fonction de transfert de la forme :

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{3.3}$$

pour un modèle d'ordre p (Huang *et al.*, 2001). On voit que cette fonction de transfert dépend de l'entrée et des p valeurs précédentes de la sortie. La transformée en Z inverse de la sortie $Y(z)$ est :

$$y(n) = x(n) + \sum_{i=1}^{p} a_i y(n-i) \tag{3.4}$$

On dit que ce modèle est à prédiction linéaire car l'échantillon $y(n)$ peut être prédit en fonction des p échantillons de la sortie qui le précédent. La valeur prédite $\hat{y}(n)$ est donc :

$$\hat{y}(n) = \sum_{i=1}^{p} a_i y(n-i) \tag{3.5}$$

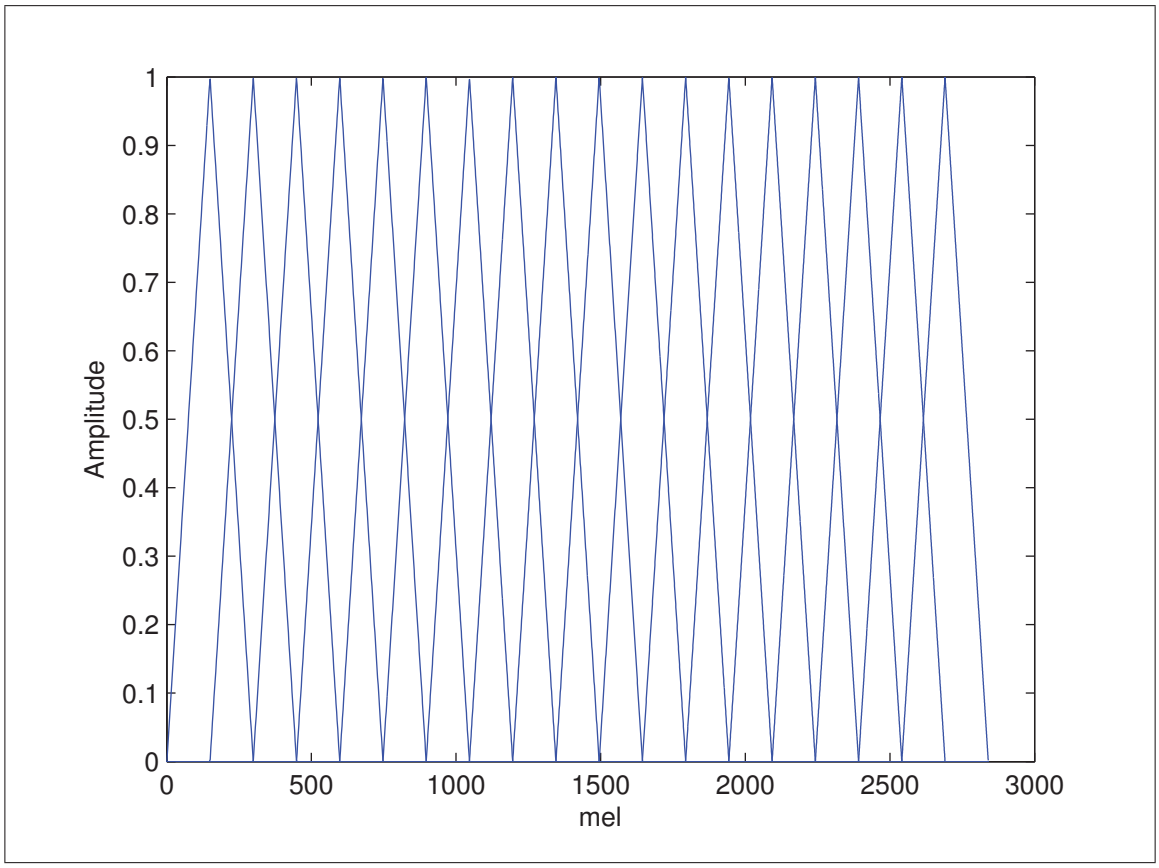


Figure 3.2 Batterie de filtres (échelle en mel).

L'erreur de prédiction $e(n)$ est :

$$e(n) = y(n) - \hat{y}(n) = y(n) - \sum_{i=1}^{p} a_i y(n-i) \tag{3.6}$$

Comme les caractéristiques de la parole changent dans le temps, on doit découper le signal avec une fenêtre, tout comme pour la transformée de Fourier. Ensuite, les coefficients sont évalués pour cette période. On suppose que le signal $y(n)$ est nul à l'extérieur de cette période. On définit l'erreur totale pour une période d'analyse :

$$E = \sum_{n=0}^{N+p-1} e^2(n) \tag{3.7}$$

On doit ajouter les p échantillons qui suivent la période car la sortie prédite durant ces échantillons n'est pas nulle : elle découle des p dernières sorties de la période d'analyse. On trouve

les coefficients a_i qui minimisent cette l'erreur totale. Pour un coefficient a_i, l'erreur est minimisée quand :

$$\frac{\partial E}{\partial a_i} = 0 \tag{3.8}$$

On va donc obtenir un système de p équations et p variables. La dérivation nous donne :

$$0 = \sum_n y(n-i)\left(y(n) - \sum_{j=1}^{p} a_j y(n-j)\right) = \sum_n y(n-i)e(n) \tag{3.9}$$

On peut manipuler cette équation et obtenir :

$$\sum_n y(n)y(n-i) = \sum_{j=1}^{p} a_j \sum_n y(n-i)y(n-j) \tag{3.10}$$

On définit l'auto-corrélation $R(k)$:

$$R(k) = \sum_{n=0}^{N-1-k} x(n)x(n+k) \tag{3.11}$$

Ainsi, on peut obtenir un système d'équation basé sur l'auto-corrélation en combinant les équations 3.10 et 3.11 :

$$R(i) = \sum_{j=1}^{p} a_j R(|i-j|) \tag{3.12}$$

Ce système d'équations peut être exprimé sous forme de matrice :

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \tag{3.13}$$

Les coefficients a_i forment un filtre $H(z) = 1/A(z)$. On peut obtenir la transformée de Fourier à partir d'une fonction en Z stable par la substitution $z = e^{j\omega}$. De cette façon, on peut comparer la transformée de Fourier court-terme du signal et le spectre obtenu à partir des coefficients de prédiction linéaire. La figure 3.3 illustre cette comparaison. On voit la correspondance entre les deux spectres. Celui obtenu à partir des LPCs est plus graduel. On voit qu'il suit l'enveloppe de la transformée de Fourier. On voit bien les fréquences de résonances, mais les harmoniques de l'onde glottale ne sont pas visibles.

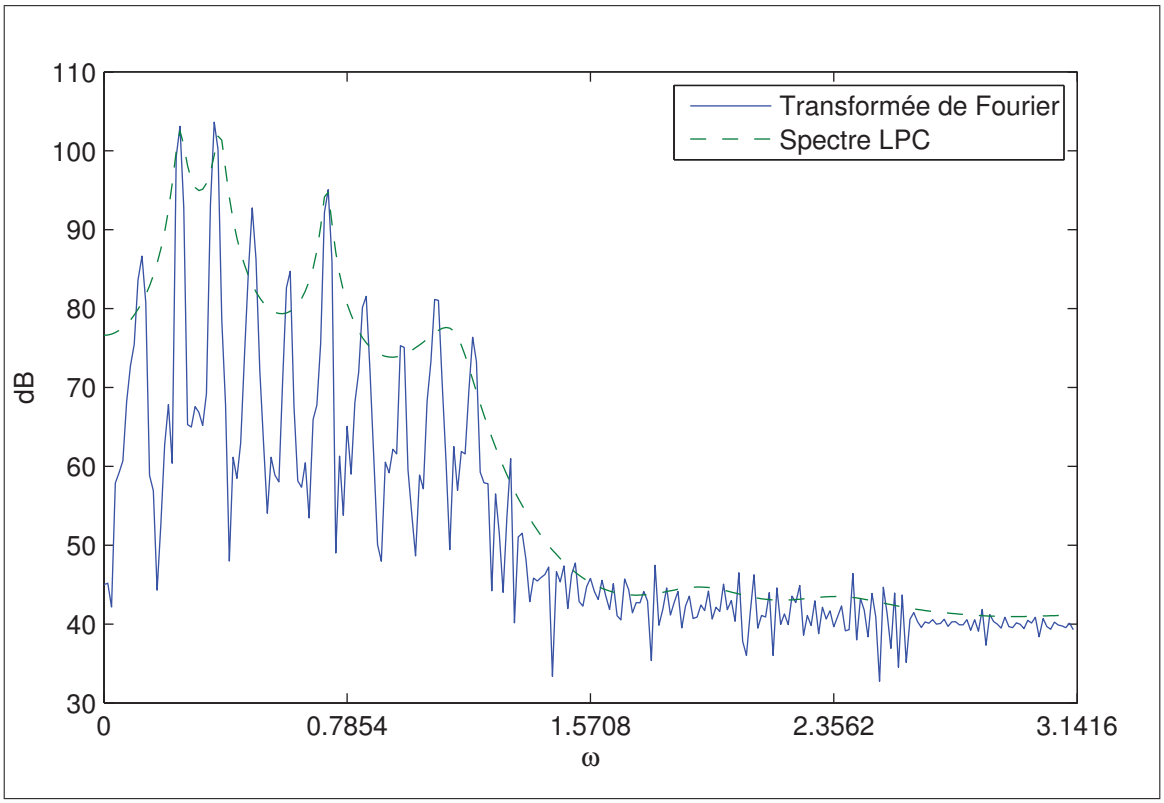


Figure 3.3 Comparaison entre la transformée de Fourier court-terme et le spectre LPC.

On peut obtenir le cepstre à partir de $H(e^{j\omega})$:

$$C_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|H(e^{j\omega})|e^{j\omega n}d\omega \tag{3.14}$$

Il est possible d'utiliser une récursion à partir des coefficients de prédiction linéaire, ce qui demande beaucoup moins de calcul. Le nombre théorique de coefficients cepstrales est infini. Cependant, on peut habituellement tronquer la séquence. La figure 3.4 illustre la différence entre le cepstre obtenu directement à partir de la transformée de Fourier court-terme du signal et celui obtenu à partir des LPCs.

3.4.3 Prédiction linéaire perceptuel et à spectre relatif

On doit ajuster le signal pour tenir compte des caractéristiques du système auditif (Hermansky *et al.*, 1986). On utilise une batterie de filtres passe-bandes, qui correspondent aux bandes critiques. On égalise ensuite la force sonore, pour tenir compte du fait que la sensibilité du

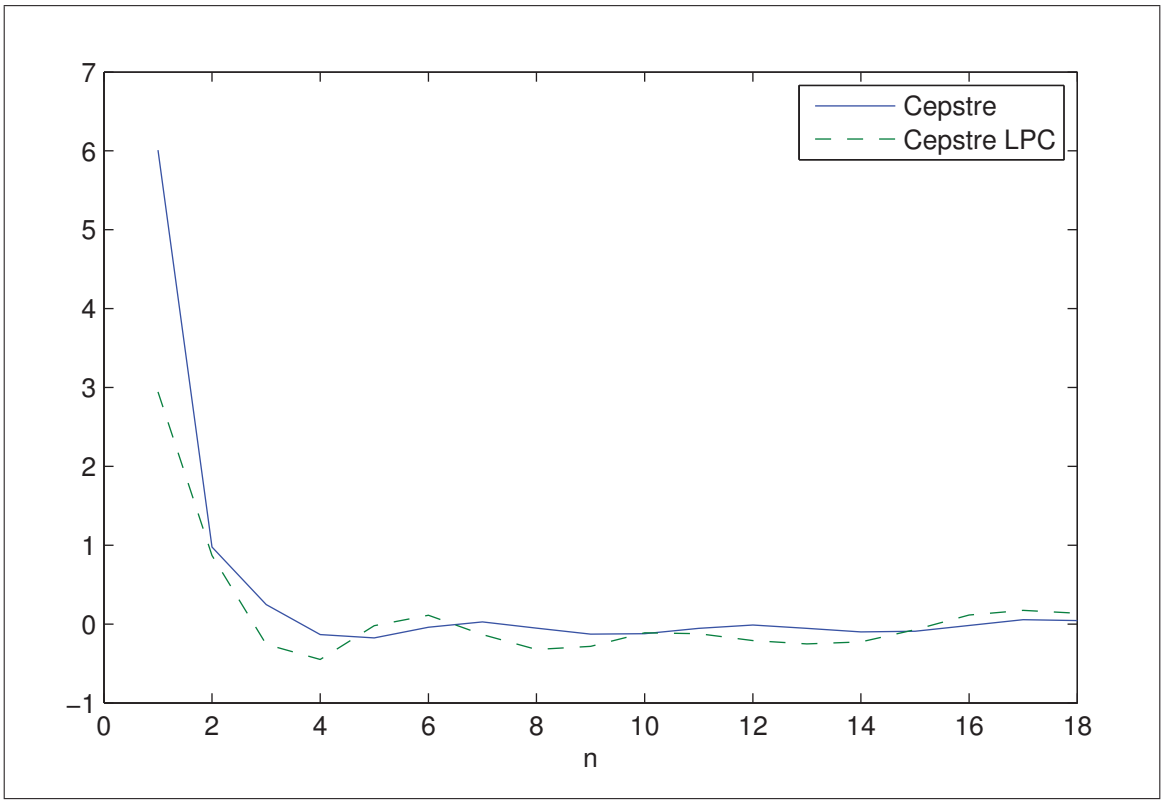


Figure 3.4 Comparaison entre le cepstre et le cepstre LPC.

système auditif varie selon la fréquence. Finalement, on applique une conversion intensité-force sonore, habituellement une racine cubique. On applique la transformée de Fourier inverse au résultat, ce qui nous donne notre signal ajusté. Il suffit ensuite de calculer les coefficients de prédiction linéaire et leurs cepstres.

La technique RASTA-PLP est similaire (Hermansky *et al.*, 1992). Cependant, on tient compte de certaine limites supplémentaires : le système auditif est sensible aux variations du signal. Si les caractéristiques changent trop lentement, le changement ne sera pas perçu. De plus, les articulateurs ne peuvent pas bouger trop rapidement. Si les caractéristiques changent trop rapidement, elles ne sont peut-être pas produites par de la parole.

On va donc calculer le logarithme des amplitudes des bandes critiques. Ainsi, le canal passe

d'une multiplication dans le domaine de fréquentielle à une addition. On passe ensuite le résultat dans un filtre passe-bande. Ce filtre retire les éléments qui varient rapidement et les éléments qui varient lentement. On applique ensuite le logarithme inverse, avant de continuer avec le traitement PLP.

3.4.4 Compression lin-log pour la robustesse au bruit additif

Hermansky & Morgan (1994) ont présenté une variante de RASTA-PLP qui améliore la robustesse au bruit additif. Cette variante ajoute un paramètre J, qui varie la linéarité de la fonction de compression/expansion. Plutôt que d'utiliser $\log x$, ont utilise

$$\log 1 + Jx \tag{3.15}$$

L'expansion se fait avec

$$e^y / J \tag{3.16}$$

et non $(e^y - 1)/J$, afin d'éviter les valeurs négatives. Ce paramètre J doit être adapté au ratio signal-bruit du signal. Quand ce ratio est élevé, J doit être élevé : on veut une compression non-linéaire. Quand ce ratio est bas, J doit être réduit : on veut une compression partiellement linéaire. La fonction de compression peut être réécrite :

$$\log J + \log \frac{1}{J} + x \tag{3.17}$$

Le filtre passe-bande va retirer la constante $\log J$. La puissance x est translatée par le terme $1/J$. Si la puissance du bruit est élevée, ce terme sera également élevé et la translation va placer x à l'extérieur de la section où le logarithme varie rapidement. De cette façon, les variations causées par le bruit additif auront moins d'impact. Évidemment, les variations rapides sont normalement le comportement voulu. Par conséquent, quand la puissance du bruit est faible, il est plus approprié d'utiliser un J haut, ce qui va limiter la translation et laisser x dans la section où le logarithme varie rapidement.

3.5 Conclusion

On a vu que les outils présentés précédemment peuvent être forgés en un système de reconnaissance de la parole. On a également vu des causes de variations problématiques pour ces systèmes, et les types d'approche qui peuvent améliorer les performances d'un système face à ces difficultés. Un de ces types d'approche est relié à l'analyse du signal. Ce type d'approche se trouve au niveau de la construction du vecteur-observation, qui décrit le signal à reconnaitre durant un intervalle de temps.

Le développement de nouvelles techniques a permit de produire des systèmes plus robustes, entre autres face aux variations de canal. Comme la robustesse est une question de degré, on peut toujours l'améliorer et la recherche de nouvelles techniques continue.

CHAPITRE 4

MÉTHODE PROPOSÉE

4.1 Introduction

L'augmentation de la robustesse par analyse du signal est une approche courante. Comme ces approches n'affectent que les étages initiaux d'un système de reconnaissance de la parole, on peut les implémenter et tester rapidement. De plus, ces approches sont du domaine du traitement de signal. On peut donc utiliser des techniques développées pour le traitement de signal en général à la reconnaissance de la parole, ce qui ouvre plusieurs avenues de recherche.

L'objet de ce mémoire est l'amélioration de la robustesse face aux changements de canal. On a choisi de se limiter aux approches basées sur le traitement de signal. Les changements de canal vont, par définition, altérer la transformée de Fourier d'un signal. On sait également que l'information contenue dans la parole reste présente malgré ces changements. Par exemple, la voix d'un interlocuteur va être altérée par le système téléphonique, mais reste néanmoins compréhensible, et ce même si l'identification de la personne qui parle peut être problématique. Comme l'être humain a moins de difficultés face aux changements de canal que les systèmes de reconnaissance automatiques, on peut supposer qu'une partie de l'information présente dans le signal original est rejetée, ou rendue moins accessible par les méthodes de reconnaissance actuelles. On peut également remarquer que ces méthodes ont la transformée de Fourier en commun.

Plutôt que d'utiliser la transformée de Fourier court-terme, suivie d'une convolution avec des filtres triangulaires pour produire une batterie de filtres passe-bandes, la transformée en paquets d'ondelettes pourrait être utilisée dans ces applications. L'information du signal original sera ainsi présentée d'une façon similaire, mais légèrement différente. En particulier, on obtiendra une résolution temporelle et fréquentielle différente selon la bande. De plus, il est possible d'utiliser une fonction d'analyse autre qu'une sinusoïde. Finalement, la fenêtre fait directement

partie de la fonction d'analyse.

On propose donc une extension de la technique RASTA-PLP développée par Hermansky *et al.* (1992). Le terme *« WRPLP »* sera utilisé pour désigner cette méthode. Divers choix et modifications découlent de cette extension, et seront décrits dans ce chapitre.

4.2 Détails de la méthode proposée

On a vu, au chapitre 3, les coefficients de prédiction linéaire cepstraux perceptuels (Hermansky *et al.*, 1992). La méthode proposée est une extension de cette méthode, où une analyse en paquets d'ondelettes est substituée à la transformée de Fourier court-terme.

1. On applique l'analyse en paquets d'ondelettes. On obtient ainsi plusieurs signaux, correspondant à des bandes de fréquences du signal original. La fréquence d'échantillonnage de ces signaux varie selon leur résolutions fréquentielles. Pour chacun de ces signaux,
   (a) On applique une fonction de compression ;
   (b) On applique un filtre passe-bande ;
   (c) On applique une fonction d'expansion ;
   (d) On applique une courbe d'égalisation de la force sonore qui modélise la perception variable de l'intensité selon la fréquence ;
   (e) On applique une racine cubique.
2. On augmente la fréquence d'échantillonnage des signaux, pour qu'ils aient tous la fréquence maximale ;
3. On applique la transformée de Fourier inverse aux signaux ;
4. On utilise le signal ajusté ainsi obtenu pour calculer le cepstre des coefficients de prédiction linéaire, tel que décrit dans la section 3.4 ;
5. On utilise les coefficients cepstraux ainsi obtenus pour construire un vecteur-observation.

4.3 Analyse en paquets d'ondelettes

Il a été mentionné précédemment que la transformée en ondelettes est une batterie de filtres passe-bandes. Si on décompose les détails en plus des approximations, on obtient la transformée en paquets d'ondelettes. Il est possible de construire une pyramide d'analyse qui corresponde approximativement aux batteries de filtres basés sur l'échelle de Mel. Comme le découpage se fait en divisant la fréquence d'échantillonnage par 2, cette pyramide va changer selon cette fréquence. Cette contrainte n'est pas particulièrement problématique : tout système digital qui doit correspondre à une fréquence précise doit tenir compte de la fréquence d'échantillonnage.

La batterie de filtres triangulaires utilisée habituellement est basée sur un modèle du système auditif humain. Ce modèle nous indique que le système auditif se comporte comme une batterie de filtres. La bande passante de ces filtres, ainsi que la distance entre les fréquences centrales, augmente avec la fréquence. C'est donc ce comportement qui doit être reproduit par notre pyramide de décomposition. Il n'est pas nécessaire de diviser le spectre exactement comme l'aurait fait une batterie de filtres. Cependant, la position et la bande passante de ces filtres peuvent servir de guides dans la conception de la pyramide. Carnero & Drygajlo (1999) ont effectué un travail similaire, dans le but de coder et débruiter la parole. La pyramide qu'ils ont utilisée peut être utilisée si la fréquence d'échantillonnage est la même. Comme c'est le cas pour les bases de données utilisées pour les tests, cette pyramide a pu être utilisée directement. Il est néanmoins pertinent de la comparer à la batterie de filtres, puisque le modèle utilisé pour la concevoir n'est pas exactement celui utilisé pour concevoir la batterie de filtres.

Les filtres triangulaires sont centrés dans l'échelle de Mel. La base du triangle s'étend du centre du filtre précédent au centre du filtre suivant. Il y a donc un recouvrement partiel entre les filtres adjacents. Les filtres sont espacés uniformément dans l'échelle de Mel. L'intervalle s'obtient en convertissant simplement la fréquence maximale qui peut être dans le signal ($f_s/2$) en Mel, et en divisant par le nombre de filtres souhaités plus 1. Le filtre dont la fréquence centrale est la

plus élevée s'étendra ainsi jusqu'à la fréquence maximale. On peut ainsi placer les fréquences centrales des filtres, ce qui permet également de placer la base des triangles. Il faut remarquer que le nombre de filtres influence les fréquences centrales. Comme la pyramide va produire 21 bandes, 21 filtres seront utilisés.

Tableau 4.1 Positions des filtres passe-bandes

	Estimation des fréquences de coupures de la transformée en ondelettes		Fréquences de coupures des filtres triangulaires		Fréquences de début et fin des filtres triangulaires	
Filtre #	Basse (Hz)	Haute (Hz)	Basse (Hz)	Haute (Hz)	Début (Hz)	Fin (Hz)
0	0	125	59	111	0	180
1	125	250	152	209	85	287
2	250	375	256	320	180	407
3	375	500	371	443	287	541
4	500	625	502	582	407	692
5	625	750	647	738	541	861
6	750	875	811	914	692	1050
7	875	1000	992	1110	861	1263
8	1000	1250	1200	1328	1050	1501
9	1250	1500	1430	1575	1263	1768
10	1500	1750	1688	1852	1501	2067
11	1750	2000	1977	2161	1768	2403
12	2000	2250	2301	2508	2067	2780
13	2250	2500	2667	2899	2403	3202
14	2500	3000	3075	3335	2780	3676
15	3000	3500	3535	3825	3202	4207
16	3500	4000	4048	4374	3676	4802
17	4000	5000	4622	4989	4207	5470
18	5000	6000	5269	5681	4802	6219
19	6000	7000	5994	6455	5470	7058
20	7000	8000	6806	7324	6219	8000

On a vu précédemment que la transformée en ondelettes pouvait être calculée à partir de filtres et de décimateurs. On sait que la fréquence de coupure de ces filtres est $f_s/4$, ou encore $\omega_c = \pi/2$. En supposant que les filtres sont idéaux, on peut ainsi facilement estimer la bande passante d'une branche. Ainsi, on sait que la bande passante d'une branche de 6 filtres

sera $f_s/2^{6+1}$, et celle d'une branche de 3 filtres $f_s/2^{3+1}$. La table 4.1 indique les fréquences de coupure de ces filtres, les fréquences de coupure de la batterie de filtres, ainsi que les fréquences de début et fin des filtres triangulaires. Une fréquence d'échantillonnage $f_s = 16000$Hz a été utilisée pour produire cette table. On constate que les bandes passantes ne sont pas exactement superposées. Ce résultat n'est pas surprenant : même en supposant que la bande passante des filtres 0 soit identique, l'échelle de Mel va produire une augmentation graduelle de la bande passante, alors que celle de la transformée en ondelettes sera constante. Ainsi, pour la transformée en ondelettes, la bande passante des filtres 0 à 7 est de 125Hz, alors que pour les filtres triangulaires, elle passe de 52Hz à 118Hz. Comme l'objectif est d'approximer le comportement des bandes critiques de l'être humain, reproduire exactement le comportement des filtres triangulaires n'est pas essentiel. On obtient bien le comportement désiré : la bande passante et la distance entre les fréquences centrales augmentent avec la fréquence.

On a vu précédemment qu'à chaque ondelette correspond un filtre passe-bas et un filtre passe-haut. Ces filtres ne sont évidemment pas idéaux. Bien que leurs fréquences de coupures soient $\omega_c = \pi/2$, le gain du filtre ne sera pas nul dans la bande de coupure et ne sera pas constant dans la bande passante. Tel que vu précédemment, l'identité de Noble nous permet de trouver la transformée en Z d'une branche. On sait donc que chaque branche est le résultat de la mise en cascade des filtres de l'ondelette dilatés. Comme la bande passante est définie comme étant la bande où l'atténuation par rapport au gain maximal est inférieur à 3dB, le fait que les gains maximaux des filtres ne sont pas superposés va changer la bande passante, ainsi que la fréquence du gain maximal de la branche. Comme la forme exacte va varier selon l'ondelette, la bande passante exacte va également varier. Par contre, la fréquence de coupure fixe des filtres nous indique le comportement général, peu importe l'ondelette choisie. Les figures 4.2 et 4.3 illustrent l'effet de la dilatation sur les filtres qui correspondent à l'ondelette de Daubechies d'ordre 5 (db5 dans le populaire logiciel Matlab).

La figure 4.4 illustre les bandes passantes ainsi que les fréquences où le gain est maximal de la transformée en ondelettes et des filtres triangulaires. Les figures 4.5, 4.6, 4.7 et 4.8 comparent

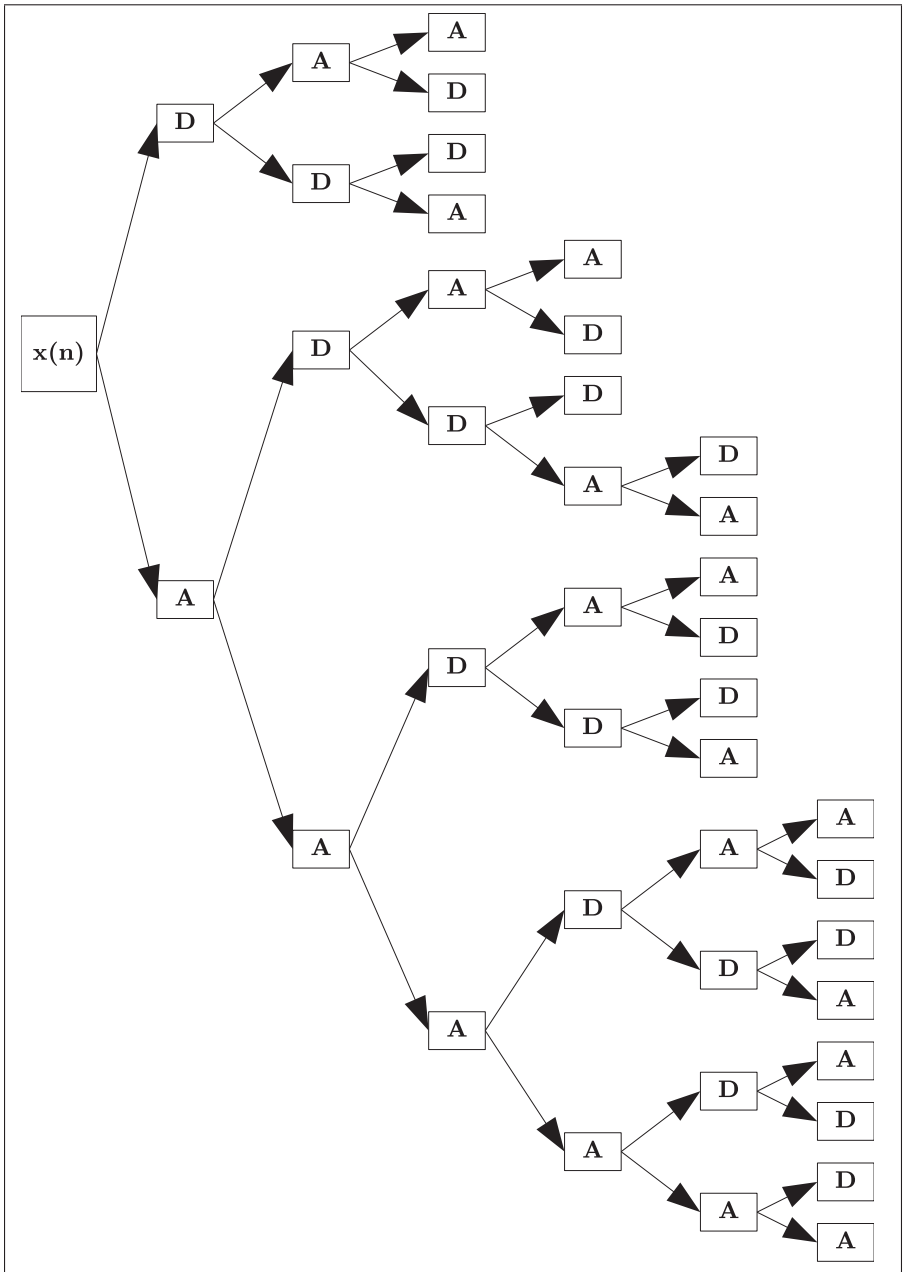


Figure 4.1 Pyramide de décomposition en paquets d'ondelettes.

la forme des filtres basés sur la transformée en ondelettes à celle des filtres triangulaires.

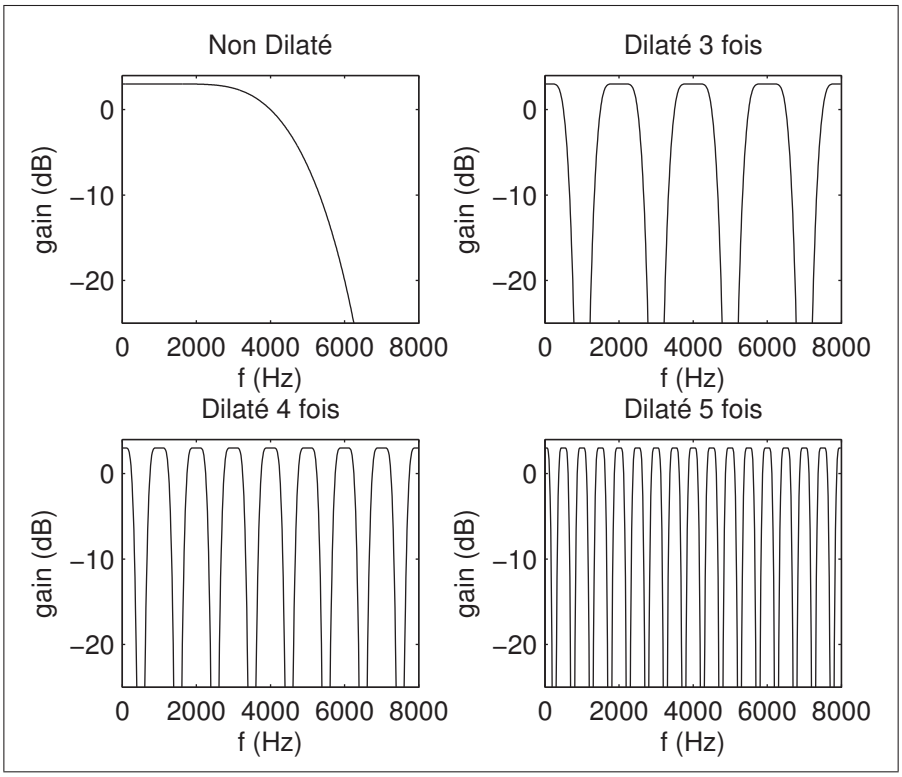


Figure 4.2 Exemples de filtres passe-bas avec dilatation.

Comme les formes des filtres obtenus à partir de la transformée en ondelettes, leurs bandes passantes et leurs fréquences maximales sont similaires à celles des filtres triangulaires, ils peuvent leurs être substitués. Comme ils ne sont pas identiques, on peut en conclure que cette substitution va avoir un effet sur la reconnaissance de la parole.

Les filtres utilisés pour la transformée en ondelettes sont des filtres à réponse impulsionnelle finie. La mise en cascade de ces filtres est donc également à réponse impulsionnelle finie. C'est cette propriété qui remplace la fenêtre de la transformée de Fourier court-terme. La longueur va varier selon la branche et la longueur des filtres :

$$L_w = 1 + (L_f - 1) \cdot (2^{L_b} - 1) \tag{4.1}$$

Où L_w est la longueur de la fenêtre, L_f est la longueur des filtres et L_b est la longueur de la branche. De plus, les décimateurs de la branche rejettent des échantillons. Ainsi, un échantillon

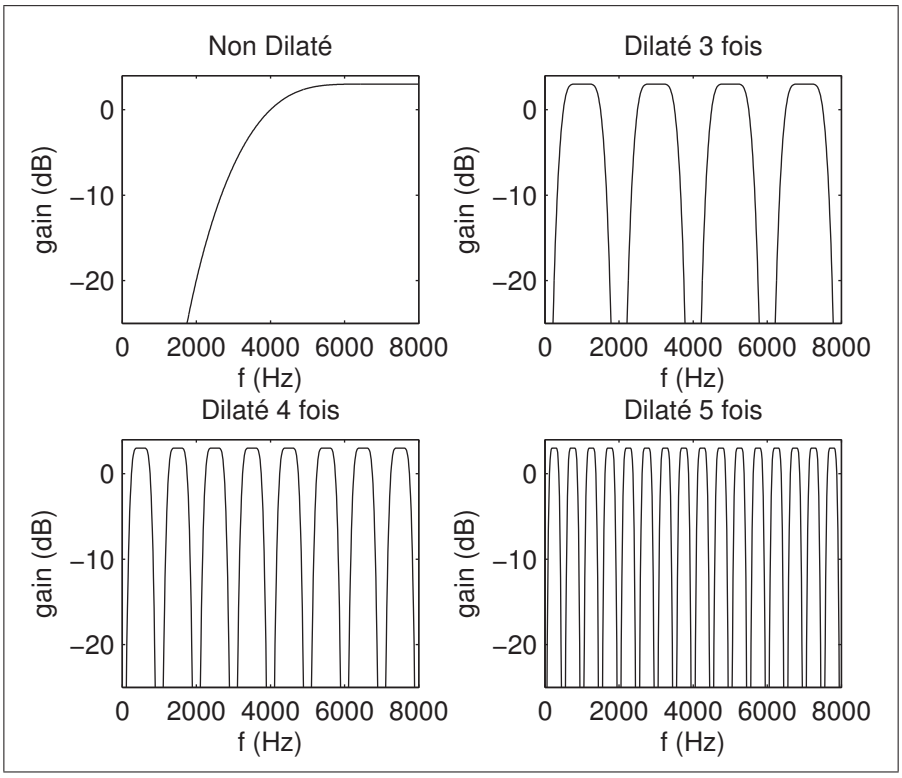


Figure 4.3 Exemples de filtres passe-hauts avec dilatation.

de la sortie représente plusieurs échantillons du signal. Le nombre d'échantillons représentés est le pas de la branche P_b. Ce pas varie seulement selon la longueur de la branche :

$$P_b = 2^{L_b} \tag{4.2}$$

Chaque échantillon de la sortie dépend donc de L_w échantillons du signal et représente P_b échantillons. Si $L_f > 2$, L_w sera également plus grand que P_b, ce qui veut dire que des échantillons du signal vont contribuer à plusieurs échantillons de la sortie.

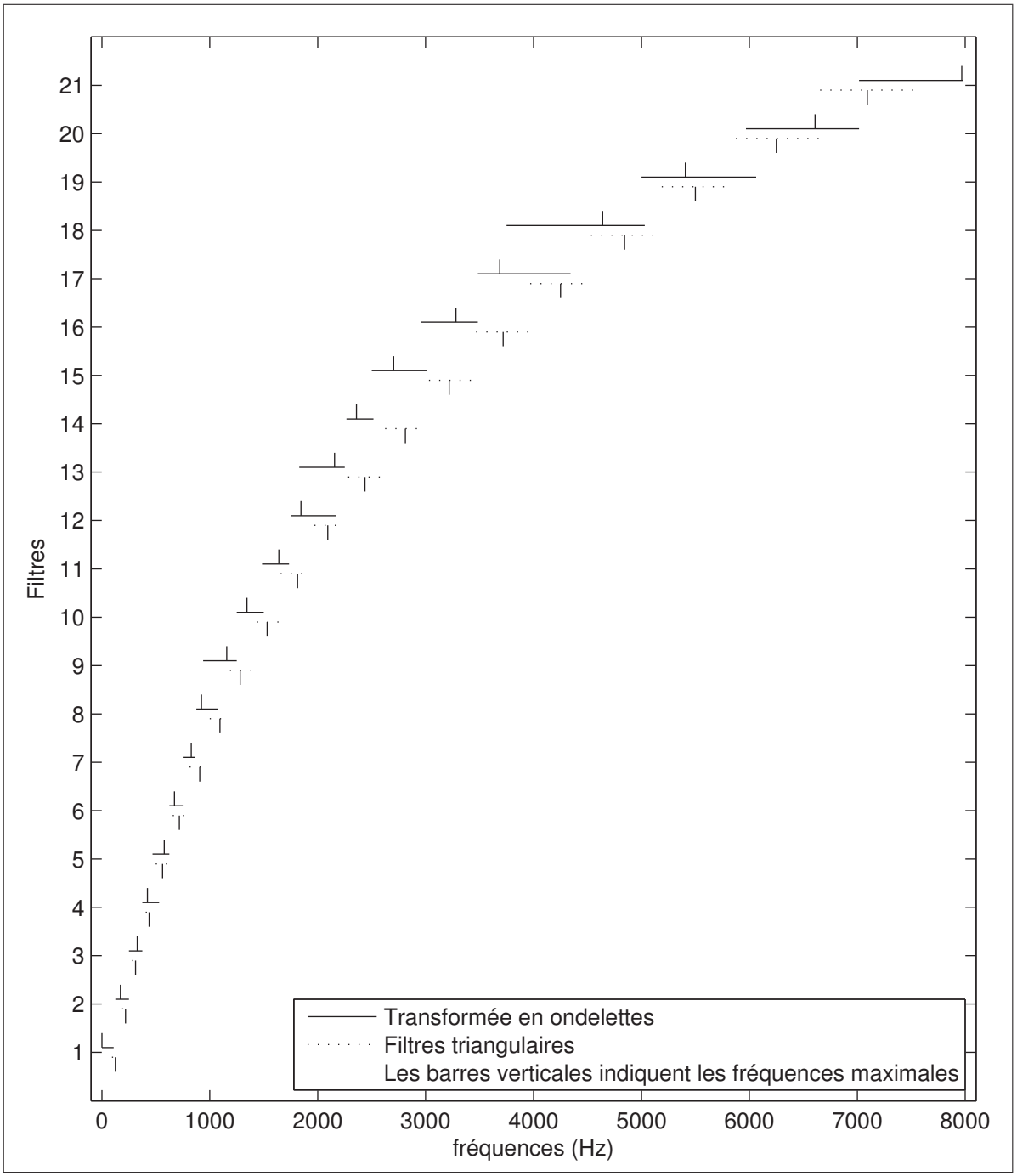


Figure 4.4 Position de la bande passante de la batterie de filtre.

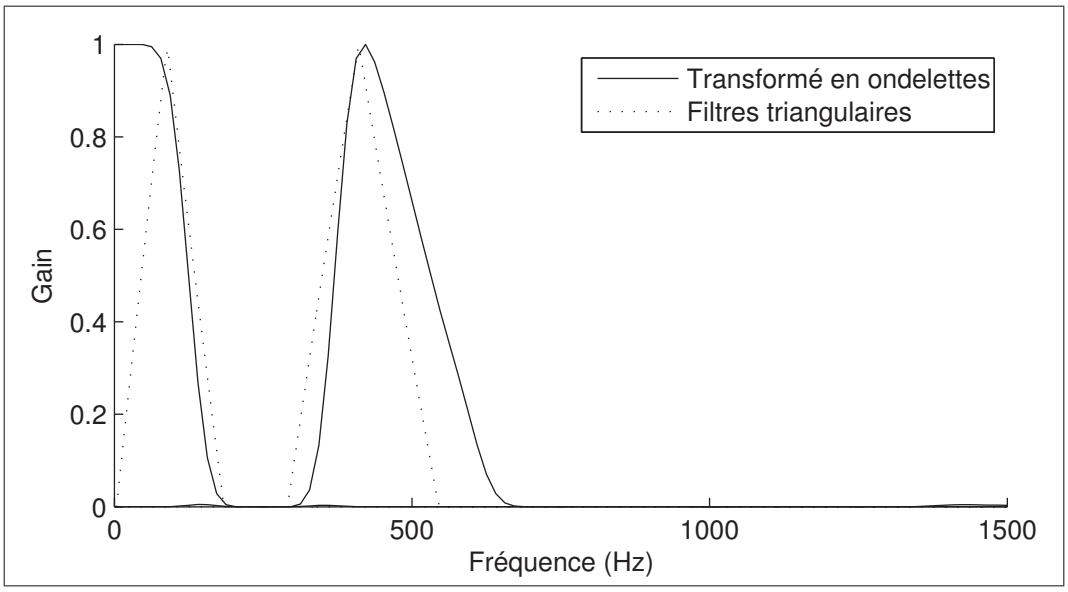


Figure 4.5 Forme des filtres 1 et 4.

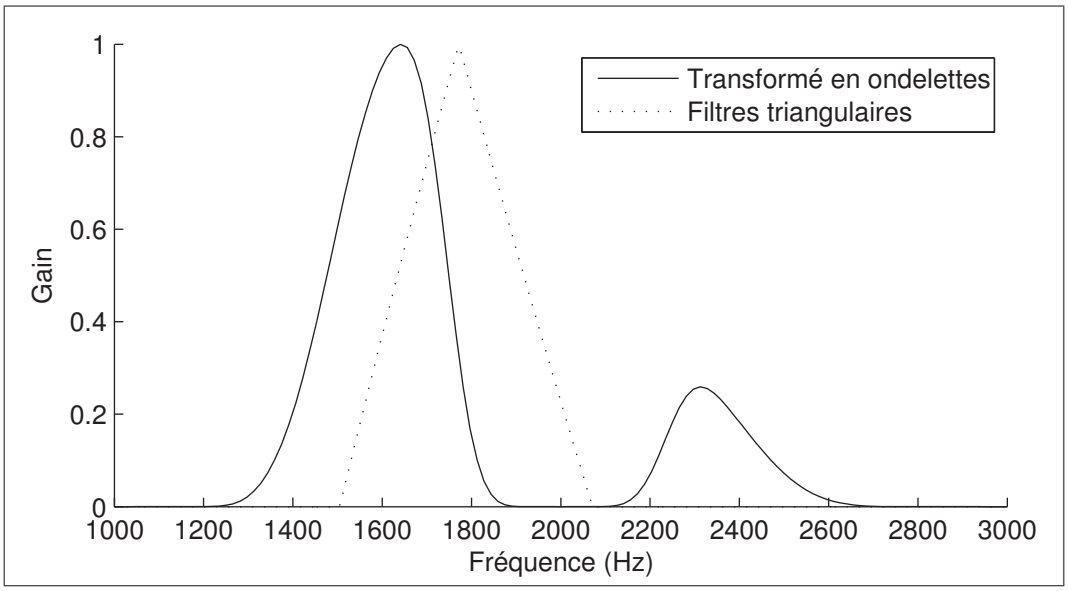


Figure 4.6 Forme du filtre 11.

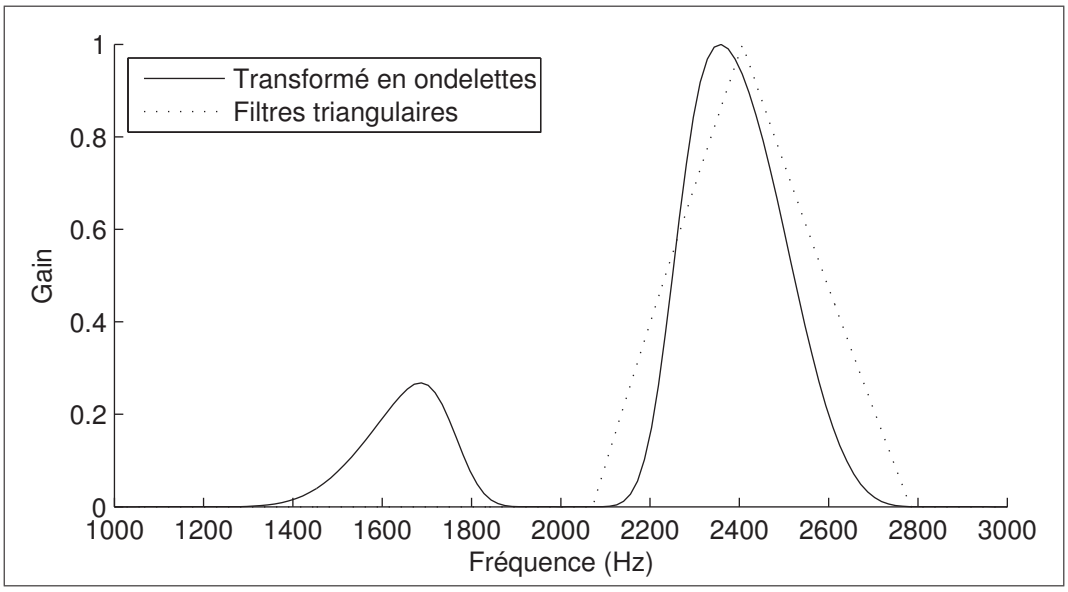


Figure 4.7 Forme des filtres 14 (ondelette) et 13 (triangulaire).

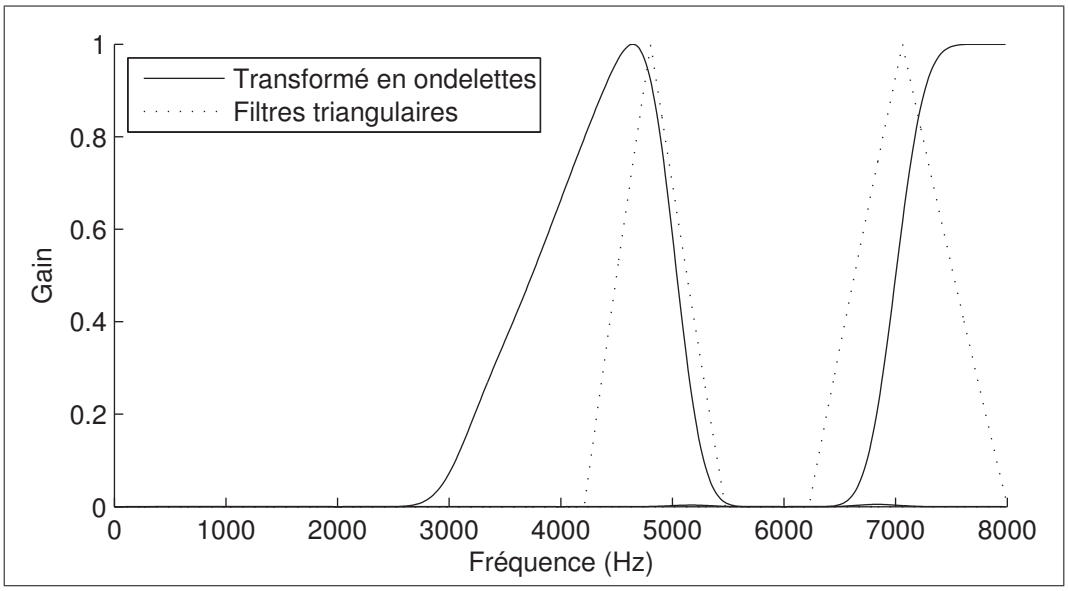


Figure 4.8 Forme des filtres 18 et 21.

La puissance de la réponse impulsionnelle varie d'une branche à l'autre. On applique donc une normalisation aux branches après la décimation. Le facteur de normalisation est simplement $1/\sum_{n=0}^{L_h} h(n)^2$, où $h(n)$ est la réponse impulsionnelle de la branche et L_h est la longueur cette réponse. Ce facteur est appliqué à la sortie d'une branche après en avoir pris le carré. On considère en effet que la puissance de chaque échantillon est la puissance dans la bande passante durant le pas de traitement. C'est l'équivalent de l'application du théorème de Parceval qui est utilisé dans la batterie de filtres triangulaires.

4.4 Filtres passe-bande

L'étape clé du traitement RASTA est l'application de filtres passe-bandes avant le calcul des coefficients de prédictions linéaires. Comme la fréquence d'échantillonnage varie selon le niveau de décomposition, il faut concevoir un filtre par niveau. Ces filtres peuvent avoir les mêmes bandes passantes (en Hz). Hermansky & Morgan (1994) ont proposé un filtre dont la fréquence de coupure inférieure était 0.26Hz et qui s'atténuait aux environs de 12.8Hz. Ces valeurs ont donc été utilisées comme guides dans la conception des filtres.

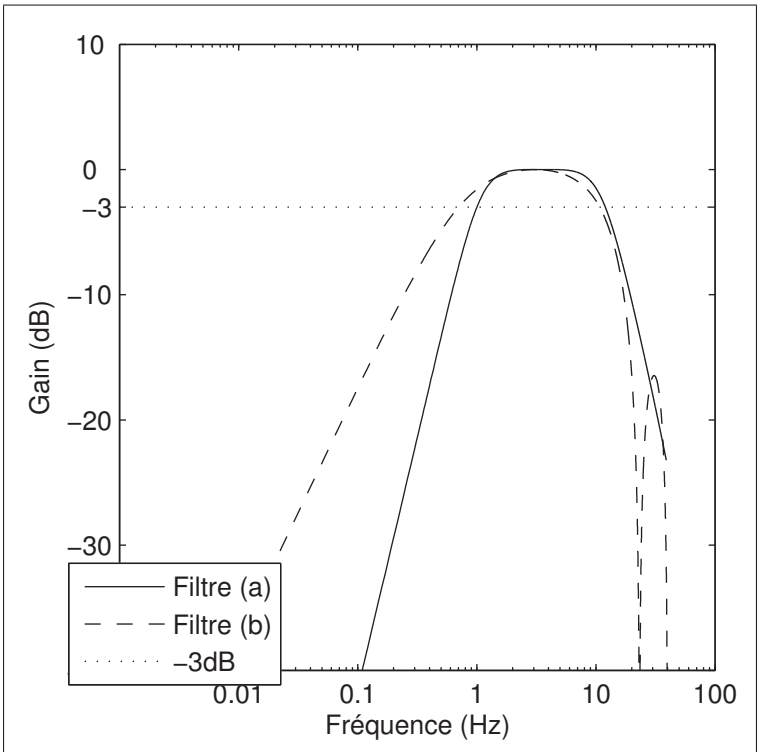
Pour simplifier la conception, l'outil « fdatool » du logiciel Matlab a été utilisé pour produire des filtres de Butterworth d'ordre 4. Les fréquences de coupures ont été fixées à 1Hz et 12Hz. La figure 4.9 compare un de ces filtres au filtre optimal trouvé par Hermansky & Morgan (1994). Les gains ont été normalisés pour faciliter la comparaison. Les filtres des autres niveaux de décomposition sont similaires et n'ont pas été inclus afin d'alléger la figure. La table 4.2 liste les coefficients des filtres.

4.5 Ajustement de la fréquence d'échantillonnage

Comme la fréquence d'échantillonnage varie d'une branche à l'autre, il faut les ajuster : la conversion en LPC suppose que la sortie de la batterie de filtres décrit une période d'échantillonnage commune. Comme plusieurs échantillons des hautes fréquences correspondent à un seul échantillon des basses, cette restriction n'est pas respectée. Il était possible de réduire

Tableau 4.2 Coefficients des filtres passe-bandes

Filtres	**F_s(Hz)**	**Transformée en Z**
Hermansky & Morgan (1994)	80	$\frac{0.2+0.1z^{-1}+0z^{-2}-0.1z^{-3}-0.2z^{-4}}{1-0.94z^{-1}}$
Niveau 6	250	$\frac{0.0159+0z^{-1}-0.0318z^{-2}+0z^{-3}+0.0159z^{-4}}{1-3.5990z^{-1}+4.8770z^{-2}-2.9543z^{-3}+0.6764z^{-4}}$
Niveau 5	500	$\frac{0.0043+0z^{-1}-0.0087z^{-2}+0z^{-3}+0.0043z^{-4}}{1-3.8014z^{-1}+5.4257z^{-2}-3.4467z^{-3}+0.8224z^{-4}}$
Niveau 4	1000	$\frac{0.0011+0z^{-1}-0.0023z^{-2}+0z^{-3}+0.0011z^{-4}}{1-3.9014z^{-1}+5.7097z^{-2}-3.7152z^{-3}+0.9069z^{-4}}$
Niveau 3	2000	$\frac{0.0003+0z^{-1}-0.0006z^{-2}+0z^{-3}+0.0003z^{-4}}{1-3.9509z^{-1}+5.8541z^{-2}-3.8555z^{-3}+0.9523z^{-4}}$



(a) Filtre pour $f_s = 250$Hz
(b) Filtre proposé par Hermansky & Morgan (1994)

Figure 4.9 Comparaison des filtres passe-bandes.

la fréquence d'échantillonnage des bandes supérieures ou d'augmenter celle des bandes inférieures. Puisque la méthode proposée dépend de la résolution temporelle supérieure dans les hautes fréquences, il faut augmenter la fréquence d'échantillonnage des basses. Pour ce faire, on suppose simplement que la valeur de la sortie est constante durant la période d'échantillonnage. Comme la période double par niveau, un échantillon est reproduit 2 fois par niveau de décomposition. Le niveau 3 a la fréquence d'échantillonnage maximale et n'est donc pas ajusté. On obtient 2 échantillons pour le niveau 4, 4 pour le niveau 5 et 8 pour le niveau 6.

4.6 Construction du vecteur-observation

Après la conversion en LPC, on obtient un pas de traitement de $8/f_s$: le pas de traitement le plus court de la transformée en paquets d'ondelettes utilisée. Comme $f_s = 16000$Hz, ce pas de traitement est de 0.5ms. On utilise habituellement des pas de traitement de l'ordre de 10 à 20ms. Intuitivement, on voit bien que le pas de traitement est beaucoup trop court : la quantité de calcul requis pour la reconnaissance va augmenter énormément. Les changements d'un échantillon à l'autre seront également très réduit : la voix est quasi-stationnaire sur une période plus élevée, qui correspond au pas de traitement habituel. Cependant, la voix n'est que quasi-stationnaire : on suppose qu'il y a des variations significatives durant la période de quasi-stationnarité. Ce sont ces variations que l'on veut utiliser pour améliorer l'indépendance envers le canal. On ne peut donc pas simplement décimer la sortie. On va donc construire un nouveau vecteur-observation, qui va décrire plusieurs pas de traitement. Le nombre de lignes dans ce vecteur sera plus élevé que dans les vecteurs originaux.

La solution la plus simple serait de concaténer des vecteurs-observations. Cette solution n'est pas appropriée : un délai dans le temps deviendrait un changement de ligne dans le vecteur. Le nombre de lignes serait également très élevé, ce qui demanderait beaucoup plus de calcul. Il convient donc de trouver une autre solution. Deux méthodes de construction ont été testées. La conception de ces méthodes est arbitraire. On a tenté de fournir de l'information significative au système de reconnaissance, mais il est difficile d'identifier les éléments importants des vecteurs

d'observation initiaux. Les méthodes de construction ont donc été choisies pour assister dans l'identification des éléments importants plus que pour la reconnaissance.

Le nombre de lignes dans le vecteur-observation résultant de ces méthodes est grand. De plus, comme les méthodes utilisées pour réduire la fréquence d'échantillonnage et organiser le vecteur sont arbitraires, on peut supposer qu'il contient de l'information redondante et non-significative pour la reconnaissance de la parole. L'analyse discriminante est une technique qui pourrait régler ces problèmes.

4.6.1 Construction par moyenne

La méthode de construction par moyenne est la plus directe. On prend la moyenne de chaque coefficient pour 8 pas de traitement. On sépare ensuite en groupe de 4 valeurs moyennes. Finalement, le vecteur est construit à partir de la moyenne de chaque groupe, ainsi que la différence entre les 3 dernières valeurs du groupe et la moyenne. Le pas de traitement est augmenté par un facteur de 32, et le nombre de lignes dans le vecteur par un facteur de 4. Au final, on a la moyenne de 32 pas de traitement pour chaque coefficient et des différences entre ces moyennes et des moyennes plus localisées dans le temps. Comme on sait que la moyenne cepstrale représente les éléments constants du canal et, par conséquent, la configuration du conduit vocal, les moyennes à grande échelle vont fournir l'information sur les variations lentes du conduit vocal, alors que les différences vont fournir de l'information sur les variations rapides. On peut remarquer qu'un délai suffisant va se traduire en changement de ligne de différence. On espère que le pas de traitement des différences de $8 \cdot 8/f_s = 4$ms va réduire ce problème. Dans tout les cas, cette construction initiale a pour but de vérifier si l'ajout des différences va réduire le taux d'erreur et de valider le processus. Comme l'étape suivante est l'application d'une analyse discriminante, on suppose que les lignes du vecteur qui ne sont pas significatives vont être supprimées par l'analyse. Au contraire, les lignes significatives vont être conservées.

4.6.2 Construction par différence

Cette méthode de construction est la plus complexe. On utilise la moyenne cepstrale pour 32 pas de traitement ainsi que des dérivées :

1. La moyenne pour 32 pas de traitement de la dérivée du vecteur original ;
2. La dérivée du vecteur original décimé par un facteur de 32 ;
3. La dérivée de la moyenne pour 32 pas de traitement.

Une fois de plus, l'analyse discriminante va retirer les lignes non-significatives. Le but de cette construction est de vérifier si les variations sont significatives pour la reconnaissance.

4.6.3 Construction par dérivées d'ordre supérieurs

Cette méthode a pour but de vérifier l'utilité des variations à l'intérieur des groupes de 32. On va donc les retirer. Pour maintenir le nombre de ligne dans le vecteur-observation, on y ajoute des dérivées :

1. La moyenne pour 32 pas de traitement ;
2. La dérivée de la moyenne pour 32 pas de traitement ;
3. La dérivée seconde de la moyenne pour 32 pas de traitement ;
4. La dérivée troisième de la moyenne pour 32 pas de traitement.

4.7 Conclusion

Une méthode de construction du vecteur-observation a été proposée pour augmenter la robustesse aux variations de canal. Cette méthode utilise la transformée en paquets d'ondelettes pour simuler le système auditif. Les caractéristiques de la pyramide de décomposition utilisée ont été comparées à celle de la batterie de filtres triangulaires habituellement utilisée. On a vu que les bandes passantes sont similaires, bien que différentes. On peut donc supposer que toute l'information présentée par la batterie de filtre sera présentée par l'analyse en paquets d'ondelettes.

L'application de filtres passe-bandes aux sorties de cette analyse va permettre de tenir compte des limites du système auditif humain. Ce type d'approche a déjà été utilisé avec succès par Hermansky *et al.* (1992). Il est possible que la nature multi-résolutionnelle de la transformée en ondelettes conserve de l'information qui est rejetée ou cachée par la transformée de Fourier. Cette même nature fait cependant en sorte que les fréquences d'échantillonnage varient selon la bande de fréquences représentées et demande la conception de plusieurs filtres passe-bandes. Il faut également faire en sorte que cette fréquence soit la même pour toutes les bandes durant les dernières étapes du traitement.

Finalement, trois méthodes de construction du vecteur-observation ont été décrites. Ces méthodes réduisent la fréquence d'échantillonnage finale et ont été conçues pour mettre en valeur différentes caractéristiques des coefficients produits par les étages antérieurs.

CHAPITRE 5

RÉSULTATS EXPÉRIMENTAUX

5.1 Introduction

La description de la méthode proposée émet implicitement l'hypothèse que cette méthode va améliorer les performances. Il reste à définir les performances en question. Comme l'environnement de test va avoir un impact direct sur les performances, il faut également le décrire.

Dans ce chapitre, les deux sources d'échantillons de parole utilisées pour entrainer un système de reconnaissance de la parole basé sur la méthode proposée seront présentées. Ces deux sources contiennent également des échantillons qui seront utilisés pour tester l'hypothèse. L'outil logiciel qui permet de construire, entrainer et tester le système sera également présenté. Les mesures de performance utilisées seront définies et les tests eux-mêmes seront décrits. Finalement, les résultats obtenus et une discussion de ces résultats seront présentés.

5.2 Bases de données

Les bases de données Timit (Fisher *et al.*, 1986) et NTimit (Jankowski *et al.*, 1990) ont été utilisées pour l'entrainement des modèles et les tests. Pour ces deux bases de données, la fréquence d'échantillonnage est de $f_s = 16$kHz et les échantillons sont des entiers signés de 16bits. La base de données NTimit est basée sur Timit. La base de données Timit est constituée de 6300 phrases et inclut 630 locuteurs, des hommes et des femmes. Ces locuteurs proviennent de plusieurs régions des États-Unis. Il s'agit donc d'une base de données de l'anglais américain.

Chaque locuteur a produit 10 phrases. Deux phrases (« sa1 » et « sa2 ») ont été produites par touts les locuteurs. Le groupe « sx » contient 450 phrases, choisies pour offrir une bonne couvertures des paires de phonèmes et des contextes de prononciations. Chaque phrase de ce groupe a été assignée à sept locuteurs. Finalement, le groupe « si » est constitué de 1890 phrases

diverses. Chacune de ces phrases a été assignée à un seul locuteur. La base de données inclut des transcriptions phonétiques des phrases.

La base de données est divisée en deux section : une section de test et une section d'entrainement. La section de test est constituée des phrases produites par 168 locuteurs, des deux sexes et de toutes les régions. La section d'entrainement est constituées des phrases produites par les autres locuteurs. On remarque que cette division fait en sorte que les phrases « sa1 » et « sa2 » font partie de l'ensemble d'entrainement et de l'ensemble de test. Ces phrases sont également les phrases les plus fréquentes dans les deux ensembles. Pour éviter de distordre les résultats, ces deux phrases ne sont pas utilisées durant l'entrainement et durant les tests.

NTimit contient les mêmes phrases que Timit. Cependant, ces phrases ont été transmises par le réseau téléphonique. Cette transmission va introduire des changements de canaux. En particulier, la fréquence de coupure sera de 4kHz plutôt que 8kHz. Il y a également un peu de bruit additif. Plusieurs canaux téléphoniques ont été utilisés. Ainsi, non seulement les canaux sont différents de celui de Timit, mais ils sont également différents entre eux. Ces changements de canaux sont la *seule* différence entre Timit et NTimit. Toute différence dans la performance d'un système de reconnaissance sera donc causée par ces changements. Cette paire de bases de données est donc particulièrement bien adaptée aux tests de la robustesse aux changements de canaux.

5.3 Outils d'entrainement et de test

L'outil HTK (Hidden Markov models ToolKit, Woodland *et al.* (1994)) permet d'initialiser, entraîner et raffiner des modèles de Markov cachés. Ce logiciel est également capable d'utiliser les modèles ainsi produits dans plusieurs scénarios de reconnaissance de la parole. Il inclut, entre autres, un module de traitement de signal pour produire les vecteurs-observations et un module qui implémente l'algorithme de Viterbi adapté à la reconnaissance de la parole continue.

HTK est constitué de plusieurs modules et programmes spécialisés. Les programmes sont principalement commandés par lignes de commande et par des fichiers de configuration, ce qui permet de les contrôler facilement avec des scripts. On peut donc automatiser l'entrainement et les tests du système.

En particulier, les outils HRest et HERest permettent de mettre à jour des modèles. HRest utilise des transcriptions pour découper les échantillons de parole en segment correspondant aux chaînes de Markov cachées à entrainer, alors que HERest ré-estime en utilisant des phrases complètes. Les programmes HInit et HCompV permettent d'initialiser les modèles. Finalement, le programme HVite utilise l'algorithme de Viterbi pour reconnaitre la parole et le programme HResults permet de calculer le taux d'erreurs. D'autres programmes permettent, entre autre, d'adapter les modèles, et de gérer et construire des modèles de langage.

Le module de traitement de signal est capable de produire des vecteurs-observations PLP, qui sont la base de la techniques RASTA-PLP présentée précédemment. Cette technique n'est cependant pas supportée par HTK. Il a donc fallu modifier cette librairie afin d'ajouter l'étage de compression-filtrage-expansion requise, ainsi que le support des variables de configuration correspondantes.

L'implémentation de la transformée en paquets d'ondelettes aurait demandé des modifications beaucoup plus complexes. Ces paramètres ont donc été calculés en utilisant l'outil Matlab. Pour minimiser les différences, des sections de HTK ont été converties du langage de programmation C au langage Matlab. Les fichiers d'observations produits de cette façon peuvent être directement lus et utilisés par HTK.

Ainsi, HTK est utilisé pour l'entrainement du système, les tests du système et pour le calcul des paramètres RASTA-PLP. Matlab est utilisé pour le calcul des paramètres basés sur la transformée en paquets d'ondelettes.

5.4 Descriptions des tests

On utilise le taux erreurs-mots pour comparer les résultats :

$$TEM = \frac{I+S+D}{N} \tag{5.1}$$

Où I est le nombre de mots insérés dans la phrase, S est le nombre de mots substitués et D est le nombre de mots retirés. N est le nombre de mots total dans la phrase correcte. Il est possible que ce taux d'erreurs dépasse 100%, si le nombre d'erreurs est supérieur au nombre de mots dans la phrase à reconnaitre. On calcul ce taux pour chaque phrase de test et on calcul la moyenne. C'est alors le TEM_m. On peut également le calculer directement pour l'ensemble des phrases. C'est alors le TEM_t. Ces deux mesures de performances sont similaires. Le TEM_t est en fait le taux erreurs-mots moyen pondéré par la longueur de la phrase :

$$TEM_m = \frac{1}{P}\sum_{n=1}^{P}\frac{I_n+S_n+D_n}{N_n} \tag{5.2}$$

$$TEM_{mp} = \frac{1}{P}\sum_{n=1}^{P}\left(\frac{N_n}{\sum_{m=1}^{P}N_m}\right)\frac{I_n+S_n+D_n}{N_n} = \frac{\sum_{n=1}^{P}I_n+S_n+D_n}{\sum_{n=1}^{P}N_n} = TEM_t \tag{5.3}$$

Où P est le nombre de phrases utilisées pour le calcul du taux. Afin d'alléger le texte, seul le taux erreurs-mots moyen sera utilisé dans les résultats détaillés. Les deux seront utilisés pour les résultats globaux. Comme les phrases longues ont plus d'impact que les phrases courtes sur le TEM_t, on va ainsi pouvoir confirmer que la longueur de la phrase n'influence pas les résultats.

La technique RASTA-PLP sera utilisée comme référence. Tout les tests utilises des modèles de Markov cachés à 3 états émetteurs, un pas de traitement de 16ms et des modèles de triphones. On utilise les dérivées d'ordre un à trois dans les vecteurs-observations RASTA-PLP, afin d'obtenir 36 lignes par observation. Les vecteurs basés sur la transformée en paquets d'ondelettes ont tous 36 lignes. Pour mieux représenter la fonction de densité de probabilité, des mixtures de gaussiennes sont utilisées dans le calcul de la probabilité qu'une observation soit produite par un état donné.

Les variables suivantes ont été identifiées :

1. La valeur de J, qui doit être adaptée au rapport signal-bruit, tel que décrit dans la section 3.4 ;
2. Le nombre de lignes conservées par l'analyse discriminante, tel que décrit dans la section 2.8 ;
3. La méthode de construction, tel que décrit dans le chapitre 4.

Les tests sont de deux types : environnements correspondants et divergents. Les tests à environnements correspondants permettent d'obtenir le taux d'erreur d'une technique quand le canal n'a pas changé. Les tests à environnements divergents permettent d'obtenir l'effet des changements de canal selon la technique. De cette façon, on peut vérifier l'influence des variables et de la méthode proposée sur la robustesse aux variations de canaux. On a également effectué des tests utilisant une autre ondelette et d'autres filtres passe-bandes, afin de vérifier l'impact de ces éléments sur les résultats.

5.4.1 Procédure d'entrainement

La procédure d'entrainement des modèles est la même pour toute les caractérisations. Les outils utilisés sont ceux fournis par HTK. Les modèles sont initialiser par l'outil HCompV, qui calcul les moyennes et variances globales. L'outil HERest est ensuite utiliser afin de différencier les modèles de phonèmes. Ces deux étapes permettent d'obtenir des modèles initiaux. Les outils HInit et HRest auraient également pût être utilisés, mais dépendent des transcriptions en phonèmes incluses dans Timit. Ces transcriptions ne sont pas parfaites et les deux méthodes produisaient des résultats comparables.

Les étapes d'initialisation ne placent pas de silence entre les mots. Elles utilisent seulement des silences au début et à la fin des phrases. On utilise l'état central de ce modèle de silence long pour produire un modèle de silence court, composé d'un seul état émetteur et qui peut être de durée 0. L'outil HVite est utilisé pour produire de nouvelles transcriptions phonétiques qui incluent ces silences courts. On utilise ensuite HERest pour raffiner les modèles.

On transforme ensuite les modèles de monophones en modèles de triphone, en utilisant les outils de manipulation de modèles HLed et HHEd. Comme certains triphones sont peu présents, ou même absents, dans les données de test, leurs modèles sont reliés à celui de triphones similaires mais plus communs. Ceci permet d'avoir plus d'exemples pour chaque modèle. Par exemple, si le triphone « a-b+c » est rare, il pourrait être relié à « e-b+d », ou encore ces deux triphones pourrait être reliés à « u-b+c ». Le choix des liens est automatiquement calculé par HTK, mais les liens possibles doivent être définies par l'utilisateur.

Finalement, on augmente le nombre de gaussiennes utilisées pour modéliser un état en utilisant l'outil HHed. On applique ensuite une analyse discriminante et on utilise HERest pour terminer l'entrainement des modèles.

5.5 Résultats

Pour tout les tests, la base de données d'entrainement est Timit. Comme les bases de données utilisées ne présentent pas beaucoup de bruit additif, $J = 10^{-6}$ a été utilisé pour les tests initiaux. Comme ce paramètre dépend du ratio signal-bruit, on suppose que la même valeur sera adéquate pour RASTA-PLP et pour la technique proposée (WRPLP). Le tableau 5.1 montre que le nombre de lignes optimal testé est 9. On a ensuite vérifié l'effet du paramètre J, en utilisant 9 lignes. Les résultats sont présentés dans le tableau 5.2.

Tableau 5.1 Taux erreurs-mots selon le nombre de lignes du vecteur-observation (RPLP)

	Données de test	
Nombre de lignes	Timit	NTimit
9	16%	88%
18	21%	89%
27	25%	95%

On a d'abord testé les méthodes de construction du vecteur-observation présentées dans la section 4.6. Le paramètre J a été initialement fixé à 10^{-6}. L'effet du nombre de lignes a également

Tableau 5.2 Taux erreurs-mots selon J (RPLP)

J	Données de test	
	Timit	NTimit
10^{-6}	16%	88%
10^{-5}	15%	85%
$5 \cdot 10^{-3}$	19%	94%

été testé. Le tableau 5.3 présente les résultats pour la construction par moyenne (section 4.6.1), et le tableau 5.4 présente ceux de la construction par différence (section 4.6.2). On constate que le nombre de lignes optimal est 9 dans les deux cas. Comme c'est également le cas dans les tests de RASTA-PLP, 9 lignes ont été utilisées dans les tests subséquents. De plus, la construction par différence a le taux erreurs-mots minimal. Cette construction a donc été utilisée pour les autres tests.

Tableau 5.3 Taux erreurs-mots selon le nombre de lignes du vecteur-observation (WRPLP, construction par moyenne)

Nombre de lignes	Données de test	
	Timit	NTimit
9	26%	68%
18	27%	79%
27	35%	88%

Tableau 5.4 Taux erreurs-mots selon le nombre de lignes du vecteur-observation (WRPLP, construction par différence)

Nombre de lignes	Données de test	
	Timit	NTimit
9	26%	53%
18	27%	64%
27	29%	68%

Le tableau 5.5 présente l'effet du paramètre J. On voit que l'effet de ce paramètre est différent

selon les données de test. Quand on utilise Timit, c'est à dire quand le canal de test est similaire à celui d'entrainement, l'augmentation de J améliore les performances. Quand les canaux divergent, l'augmentation de J diminue les performances.

Tableau 5.5 Taux erreurs-mots selon J
(WRPLP, construction par différence)

J	Données de test	
	Timit	NTimit
10^{-6}	26%	53%
10^{-5}	19%	67%
$5 \cdot 10^{-3}$	16%	90%

Finalement, comme les choix de l'ondelette et des filtres passe-bandes ont été arbitraires, une autre ondelette et un autre ensemble de filtres ont été utilisés. Le paramètre J a été fixé à 10^{-5} pour ces tests, puisque cette valeur semblait être un bon compromit entre Timit et NTimit. L'ondelette utilisée pour le premier test était une Haar. Cette ondelette est parente à l'ondelette utilisée dans les autres tests. Les filtres passe-bandes du second test ont été conçus en utilisant l'outil « FDAtool », sans fixer l'ordre du filtre. Les fréquences de coupures étaient situées à 5Hz et 20Hz. L'atténuation minimale était de 20dB à 2Hz et 40Hz. Le tableau 5.6 présente l'effet de ces variations sur le taux erreurs-mots. On y constate que le changement de l'ondelette n'a pas eu d'effet significatif alors que le changement de filtre a causé une détérioration des performances.

Tableau 5.6 Taux erreurs-mots selon
l'ondelette et les filtres
(WRPLP, construction par différence)

Variation	Données de test	
	Timit	NTimit
Ondelettes	20%	66%
Filtres	31%	81%
Aucune variation	19%	67%

Finalement, l'effet de la construction par dérivées d'ordre supérieurs (section 4.6.3) a été testé. Le tableau 5.7 présente les résultats selon le paramètre J.

Tableau 5.7 Taux erreurs-mots selon J
(WRPLP, construction par dérivées d'ordre supérieurs)

J	Données de test	
	Timit	NTimit
10^{-6}	27%	57%
10^{-5}	16%	63%

Le tableau 5.8 récapitule les résultats importants obtenus en avec des vecteurs-observations de 9 lignes. On y a ajouté les taux erreurs-mots totaux (section 5.4), en plus des taux moyens. Ceci permet de confirmer que la longueur des phrases influence peu les résultats. La figure 5.1 illustre les taux moyens. Les TEMs placés à gauche des barres sont ceux avec obtenus avec Timit, ceux à droites sont obtenus avec NTimit.

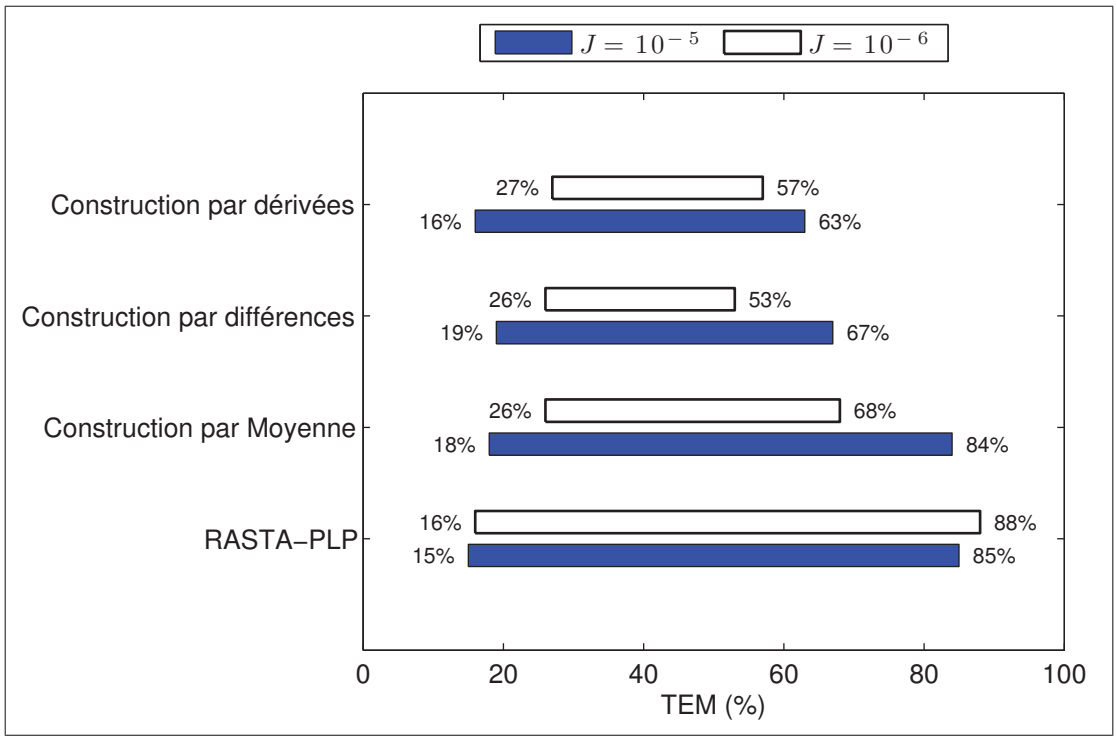


Figure 5.1 Comparaison des taux erreurs-mots

Tableau 5.8 Comparaison des taux erreurs-mots

	TEM_m		TEM_t	
	Données de test		Données de test	
Environnement	Timit	NTimit	Timit	NTimit
RASTA PLP				
$J = 10^{-5}$	15%	85%	15%	85%
$J = 10^{-6}$	16%	88%	16%	87%
Construction par moyenne				
$J = 10^{-5}$	18%	84%	19%	83%
$J = 10^{-6}$	26%	68%	26%	68%
Construction par différences				
$J = 10^{-5}$	19%	67%	19%	67%
$J = 10^{-6}$	26%	53%	27%	54%
Construction par dérivées				
$J = 10^{-5}$	16%	63%	16%	64%
$J = 10^{-6}$	27%	57%	27%	58%

5.6 Discussion

Les résultats obtenus permettent de conclure que WRPLP est plus robuste aux variations de canal que RASTA-PLP. Il est cependant plus difficile de déterminer la (ou les) cause(s) de cette amélioration. Les méthodes de construction du vecteur-observation qui ont produit les TEM_m les plus réduits sont les deux qui utilisent le plus de variations externes au pas de traitement. On conclut donc que les variations à long termes sont significatives et augmentent la robustesse aux changements de canal.

La construction par moyenne, qui utilise uniquement les variations dans le pas de traitement, a un TEM_m comparable a celui des autres constructions quand les environnements de test et d'entrainement sont similaires, et le TEM_m est plus réduit que celui de RASTA-PLP quand les environnements sont différents. On peut en conclure que les variations durant un pas de traitement ne sont pas significatifs quand il y a peu de variation de canal. Comme le TEM_m le plus réduit quand le canal change a été obtenu avec la construction par différences, qui inclut des variations à l'intérieur et à l'extérieur du pas de traitement, on peut être tenté de conclure que les deux types de variations sont significatifs. Cependant, ce résultat dépend du paramètre J, puisque la construction par dérivées d'ordres supérieurs à un TEM_m plus élevé pour $J = 10^{-6}$, mais également un TEM_m plus *réduit* pour $J = 10^{-5}$. Si on observe la différence de performance entre les environnements, ces deux constructions sont équivalentes pour $J = 10^{-5}$. Le TEM_m a augmenté de 48% pour la construction par différences et de 47% pour la construction par dérivées. Pour $J = 10^{-6}$, on constate que l'augmentation est plus réduite pour la construction par différences (27%) que pour la construction par dérivées (30%). Il semble donc que la construction par différences soit plus robuste aux variations de canal, mais plus sensible au paramètre J. Comme cette construction dépend des variations à long terme et des variations à l'intérieur d'un pas de traitement, on peut conclure que les deux types de variations sont significatives, mais que les variations internes sont plus sensibles au paramètre J.

La détérioration des performances est moindre pour les tests utilisant la transformé en ondelette. On ne peut malheureusement pas tirer de conclusion certaines sur la cause de cette amélioration. Cependant, on peut émettre l'hypothèse que la transformé de Fourier supprime de l'information temporelle. Si il n'y a pas de variations de canal, cette information est peut-être superflue, ou même nuisible. Comme la variation de canal va altérer la réponse en fréquence, cette information redondante qui varie dans le temps peut devenir significative. La transformé en ondelette pourrait, grâce à sa résolution temporelle variable, conserver cette information. Les résultats obtenus ne contredisent pas cette hypothèse : les TEMs sont moins élevés pour RASTA-PLP que pour la méthode proposée quand il n'y a pas de variations de canal et sont plus élevés quand il y a des variations.

Les résultats illustrent une faiblesse de la méthode proposée : le bruit additif. Les variations du paramètre J ont eux beaucoup plus d'effet sur les résultats de ces tests que sur ceux des tests initiaux avec RASTA-PLP. Comme la valeur optimale de ce paramètre dépend du ratio signal-bruit, ont peut conclure que WRPLP est plus sensible au bruit que RASTA-PLP.

Les tests utilisant une ondelette Haar indiquent que le choix de l'ondelette est peu important. Finalement, les tests utilisant des filtres passe-bandes différents indiquent que le choix du filtre a un impact important sur les performances. On peut donc conclure que les filtres ont un impact important.

CONCLUSION

La reconnaissance de la parole est l'extraction de symboles d'un signal sonore. Comme ces symboles sont produits pour transmettre de l'information, ils sont reliés entre eux. On reconnait les symboles par des caractéristiques qui sont présentes dans le signal. L'environnement va modifier ces caractéristiques et ainsi causer des erreurs dans la reconnaissance. Différentes techniques existent pour réduire l'influence de l'environnement. Certaines utilisent les liens entre les symboles, puisqu'ils ne dépendent pas de l'environnement. D'autres utilisent des méthodes de caractérisation qui sont moins affectées par l'environnement.

Les variations de canal vont réduire les performances des systèmes de reconnaissance de la parole. Comme ces variations sont souvent inévitables, il est important de développer des systèmes qui leurs sont robustes. C'est à cette fin qu'une nouvelle méthode de caractérisation a été présentée dans ce mémoire. Cette méthode utilise la transformée en paquets d'ondelettes comme étage initial de caractérisation. On a testé les performances de cette méthode face à des variations de canal. Les résultats de ces tests ont montré une amélioration de la robustesse quand ils ont été comparés à une méthode existante. Les taux d'erreurs-mots obtenus en utilisant la caractérisation RASTA-PLP étaient 85% et 88%, selon le paramètre J utilisé. Dans les même circonstances, la construction par dérivées de la méthode proposée produisaient des taux d'erreurs-mots de 63% et 57%.

Ces tests ont permit d'identifier que le paramètre J avait une forte influence sur les résultats. Comme ce paramètre dépend du bruit additif, il est possible que la méthode proposée y soit sensible. Une investigation de l'effet de ce bruit sera donc des plus pertinente. Si elle confirme la présence d'une faiblesse au bruit, cette investigation devrait également tenter de modifier la méthode afin d'améliorer sa robustesse au bruit.

La transformée en paquets d'ondelettes requiert plusieurs choix. En particulier on doit choisir une (ou plusieurs) ondelettes mères. On doit également déterminer quelles branches seront

décomposées. Une seule pyramide de décomposition a été utilisée dans les tests. Comme cette pyramide doit seulement approximer les bandes critiques du système auditif humain, d'autres pyramides sont possibles. On a testé les performances en utilisant une seconde ondelette mère. Cette substitution a légèrement changé les performances. Il serait donc souhaitable d'étudier l'effet de la pyramide de décomposition et de l'ondelette mère sur les performances.

Trois méthodes ont été utilisées pour produire le vecteur-observation. Ces méthodes ont été choisies dans le but de déterminer l'importance des variations durant un pas de traitement. Le choix était donc restreint. D'autre méthodes, basées sur d'autres considérations, pourraient améliorer les performances. En particulier, les coefficients cepstraux ont tous été utilisés de la même façon. Il est possible, par exemple, que seules les variations rapides de certains soient importantes. La construction exacte du vecteur-observation pourrait également être ajoutée dans la procédure d'entrainement. Ce serait donc un paramètre du modèle, plutôt qu'un élément constant.

Finalement, des filtres passe-bandes sont utilisés durant la caractérisation. Les tests ont démontrés que ces filtres ont un effet sur les performances. La même constatation s'applique à la méthode qui a inspiré la méthode proposée. Comme la conception de ces filtres était quelque peu arbitraire, une étude plus détaillé de ces filtres pourrait permettre d'améliorer les performances. En particulier, il se pourrait que les réponses fréquentielles optimales varient selon la bande de fréquence.

BIBLIOGRAPHIE

Adda-Decker, Martine, Philippe Boula de Mareuil, Gilles Adda et Lori Lamel. 2005. « Investigating syllabic structures and their variation in spontaneous French ». *Speech Communication*, vol. 46, n°2, p. 119–139.

Benzeghiba, M., R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi et C. Wellekens. 2007. « Automatic speech recognition and speech variability : A review ». *Speech Communication*, vol. 49, n°10-11, p. 763–786.

Boite, Rene et Murat Kunt. 1987. *Traitement de la parole*. « Traité d'électricité ». Lausanne (Suisse) : Presses Polytechniques romandes, 240 p.

Bou-Ghazale, S. E. et J. H. L. Hansen. 2000. « A comparative study of traditional and newly proposed features for recognition of speech under stress ». *Speech and Audio Processing, IEEE Transactions on*, vol. 8, n°4, p. 429–442.

Byrne, W., D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward et Zhu Wei-Jing. 2004. « Automatic recognition of spontaneous speech for access to multilingual oral history archives ». *Speech and Audio Processing, IEEE Transactions on*, vol. 12, n°4, p. 420–435.

Calliope. 1989. *La parole et son traitement automatique*. « Collection technique et scientifique des télécommunications ». Paris : Masson, 717 p.

Carnero, B. et A. Drygajlo. 1999. « Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms ». *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 47, n°6, p. 1622–1635.

Das, S., R. Bakis, A. Nadas, D. Nahamoo et M. Picheny. 1993. « Influence of background noise and microphone on the performance of the IBM Tangora speech recognition system ». In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*. vol. 2, p. 71–74.

Davis, S. et P. Mermelstein. 1980. « Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences ». *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, vol. 28, n°4, p. 357–366.

Fisher, W., G. Doddington et K. Goudie-Marshall. 1986. « The DARPA Speech Recognition Research Database : Specifications and Status ». In *Proc. DARPA Workshop on Speech Recognition*. p. 93–99.

Hermansky, H. et N. Morgan. 1994. « RASTA processing of speech ». *Speech and Audio Processing, IEEE Transactions on*, vol. 2, n°4, p. 578–589.

Hermansky, H., K. Tsuga, S. Makino et H. Wakita. 1986. « Perceptually based processing in automatic speech recognition ». In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86)*. vol. 11, p. 1971–1974.

Hermansky, H., N. Morgan, A. Bayya et P. Kohn. 1992. « RASTA-PLP speech analysis technique ». In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*. vol. 1, p. 121–124.

Huang, Xuedong, Alex Acero et Hsiao-Wuen Hon. 2001. *Spoken Language Processing : a guide to theory, algorithm, and system development*. Prentice-Hall, 980 p.

Jankowski, C., A. Kalyanswamy, S. Basson et J. Spitz. 1990. « NTIMIT : a phonetically balanced, continuous speech, telephone bandwidth speech database ». In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. vol. 1, p. 109–112.

Janse, Esther. 2004. « Word perception in fast speech : artificially time-compressed vs. naturally produced fast speech ». *Speech Communication*, vol. 42, n°2, p. 155–173.

Junqua, J. C., S. Fincke et K. Field. 1999. « The Lombard effect : a reflex to better communicate with others in noise ». In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*. vol. 4, p. 2083–2086.

Junqua, Jean-Claude. 2000. *Robust Speech Recognition in Embedded Systems and PC Applications*. Norwell (Massachusetts) : Kluwer, 177 p.

Junqua, Jean-Claude et Jean-Paul Haton. 1996. *Robustness in Automatic Speech Recognition : Fundamentals and Applications*. Norwell (Massachusetts) : Kluwer, 440 p.

Katz, S. 1987. « Estimation of probabilities from sparse data for the language model component of a speech recognizer ». *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, n°3, p. 400–401.

Kumar, Nagendra et Andreas G. Andreou. 1998. « Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition ». *Speech Communication*, vol. 26, n°4, p. 283–297. doi : DOI : 10.1016/S0167-6393(98)00061-2.

Mallat, S. G. 1989. « A theory for multiresolution signal decomposition : the wavelet representation ». *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, n°7, p. 674–693.

Woodland, P. C., J. J. Odell, V. Valtchev et S. J. Young. 1994. « Large vocabulary continuous speech recognition using HTK ». In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1994)*. vol. ii, p. II/125–II/128 vol.2.